AI Coding Agents Rival Human Methodological Diversity in Social Science

Meysam Alizadeh
University of Zurich
University of Oxford

Fabrizio Gilardi
University of Zurich

Mohsen Mosleh
University of Oxford

Enkelejda Kasneci
Technical University of Munich

Abstract

Large language model (LLM) systems often struggle to produce the diverse, human-like creativity expected in open-ended tasks without clear ground-truth answers, raising concerns about AI homogenization. We argue that observational social science provides a strong test case because researchers analyzing the same hypothesis and data often reach different methodological choices and conclusions, especially when core concepts and quantities are open to broad interpretation. Here, we examine whether LLM-based coding agents exhibit comparable methodological diversity or instead converge toward dominant analytic paradigms. Using a many-analysts design on a prominent immigration and social-policy hypothesis, we evaluate repeated runs of Claude Code and Codex. Codex produces methodological diversity similar to human research teams, whereas Claude Code exhibits nearly three times more diversity across model specification, robustness strategies, and analytic workflows. Despite this variation, agents mostly produce effect estimates and substantive conclusions similar to those reached by humans. Prompt-induced researcher priors do not meaningfully alter estimates or final verdicts, although they partially shift methodological decisions. By contrast, an explicit confirmatory prompt substantially changes Claude Code's final interpretations without comparably changing its coefficient distributions, suggesting that the failure emerges primarily in the interpretation layer rather than the estimation layer. The same manipulation has little effect on Codex. Together, the results challenge concerns about AI homogenization by showing that coding agents can rival or exceed human methodological diversity while remaining vulnerable to confirmatory framing at the interpretation stage.

***K*eywords** AI in Science · AI Coding Agents · Artificial Hivemind

1 Introduction

Scientific discovery depends not only on the availability of data, but also on the diversity of methods used to interpret it [1, 2]. Across disciplines, progress has historically emerged from methodological pluralism, in which competing analytical strategies generate alternative explanations tested against empirical evidence [3, 4], collectively shaping scientific understanding [5]. Such diversity is particularly important in research on human societies, where core concepts and quantities are often open to broad interpretation [6]. Recent advances in LLM-based agents now enable autonomous execution of substantial parts of the research workflow, including code generation, replication of published analyses, and machine-learning experimentation [**?**, **?**, **?**, 7]. As these systems increasingly participate in methodological decision-making, a central question emerges: can they reproduce the methodological diversity observed in scientific inquiry, or do they converge toward dominant analytic paradigms leading to epistemic homogenization?

LLMs often show reduced creative diversity in problems without definitive ground-truth answers [8, 9, 10, 11], raising concerns about AI homogenization [12, 13, 14]. We argue that observational social science provides

a stringent test case for this concern. Many core social-science constructs, such as socioeconomic status or partisanship, are inherently unobservable and admit multiple competing operationalizations [15, 16, 17]. These operationalizations reflect broader theoretical and normative assumptions about what a construct should capture [18, 19]. As a result, researchers analyzing the same hypothesis and data often arrive at different methodological choices and conclusions [6].

Here, we examine whether LLM-based agents exhibit methodological diversity comparable to that observed in observational social science research. Building on a many-analysts dataset in which 73 research teams independently tested the same prominent hypothesis [6]—that greater immigration reduces public support for social policy [20]—using identical data, we investigate three questions. First, do AI agents exhibit methodological diversity without producing substantially different empirical conclusions? Second, do frontier coding agents converge toward similar methodological decisions, or diverge from one another? Third, are these agents robust to prompt-induced researcher priors and confirmatory framing attempts? We find that Codex exhibits methodological diversity comparable to human analyst teams, whereas Claude Code produces substantially greater diversity while remaining vulnerable to confirmatory framing primarily at the interpretation layer.

Methodological diversity, however, is also a source of fragility. Even when researchers act in good faith, small analytical decisions can substantially shape empirical results. In the many-analysts study underlying our benchmark, 73 independent teams analyzing identical data reached effect estimates ranging from strongly negative to strongly positive [6]. These choices are also not made on a neutral background. Re-analyzing the same data showed that researchers' prior views on immigration systematically predicted their model specifications and reported conclusions [21]. Analogously, LLMs exhibit sycophancy toward user framings [22, 23] and remain susceptible to reward- and specification-hacking [24, 25]. The same researcher degrees of freedom can further enable selective reporting and p-hacking through "garden of forking paths" decisions [26, 27]. Methodological diversity is therefore both a driver of discovery and a substrate on which uncertainty, bias, and opportunism can act.

To separate methodological choice from empirical estimation and narrative interpretation, we analyze agent behavior at three levels. The *design layer* consists of methodological choices about measurement, sample definition, model specification, estimator selection, uncertainty quantification, and robustness checks. The *decision-rule layer* consists of mapping empirical estimates onto a substantive verdict about the hypothesis (e.g. concluding that a hypothesis is supported if four of six estimates are negative and statistically significant at $p < 0.05$). The *verdict layer* consists of the faithful implementation of the decision rule. This distinction is central because an agent may appear stable at one layer and unstable at another: for example, a prompt may leave the distribution of coefficients largely unchanged while altering the verbal conclusion drawn from those coefficients. Our experiments therefore evaluate not only whether agents produce the right numbers, but how they choose analyses and how faithfully they narrate the resulting evidence.

2 Results

Before presenting the results, we briefly summarize the experimental setup (see Methods for full details). Each agent—Claude Code (Opus 4.7 1M, "Extra High Effort") and Codex (GPT 5.5, "Extra High Intelligence")—completed twenty independent runs of the same task: testing the hypothesis that higher immigration reduces public support for social policy, using the original International Social Survey Programme (ISSP) data and country-level macroeconomic indicators. Both agents received the identical natural-language prompt; no agent specific wording, hints, or scaffolding were used. Each agent operated within a sandboxed working directory that confined file-system access to the provided replication materials, but within that sandbox the agents were permitted to install Python and R packages and to perform unrestricted web searches, mirroring the resources available to the human research teams in the original crowdsourced study. Each run encompassed the full pipeline—research design, code authorship, execution, and written conclusion—and proceeded in fully automated mode, with no human intervention at any step and no memory of any prior run.

2.1 Comparing AI Agents and Human Researchers in Methodological Diversity

Fig. 1 summarises the distribution of standardised average marginal effects (AMEs) across three matched-size groups: a random sample of 20 of the 73 human research teams from the original crowdsourced replication initiative [6] (panel A; seed = 42), 20 independent runs of Claude Code (panel B), and 20 independent runs of Codex (panel C). Each hash mark is one converged model; models within a panel are ordered along the x-axis by AME, and the three panels share a common x-axis range (0–1,100) so that the horizontal extent of each hash-mark block is proportional to the total number of executed specifications.

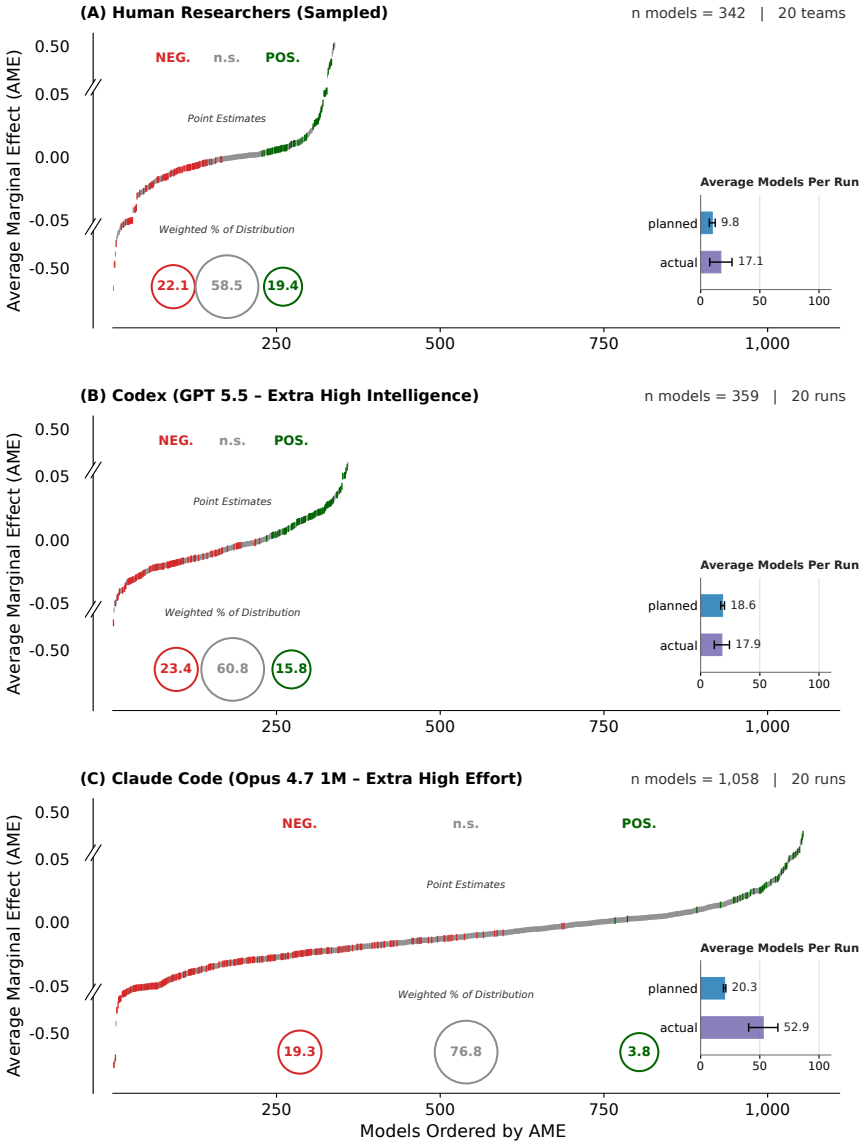

Figure 1: **Specification curves of standardized average marginal effects (AMEs)** for the hypothesis that immigration reduces public support for social policy. Each hash mark is one converged model, ordered along the x-axis by AME within each panel; colour denotes the 95% CI relative to zero (red: negative-significant; grey: includes zero; dark green: positive-significant). The y-axis is piecewise-compressed with breaks at ± 0.05. Circles report team/run-weighted percentages of models in each category (weights $1/n_{\text{models per team/run}}$ so each team/run contributes equally). The lower-right inset contrasts the number of models implied by each team's pre-registered factor grid (*planned*) with the number actually executed (*actual*); bars are means and error bars are 95% CIs. (*A*) Twenty teams drawn at random (seed = 42) from the 73 teams of ref. [6]; $n = 342$ models. (*B*) Claude Code (Opus 4.7 1M, "Extra High Effort"), 20 runs; $n = 1{,}058$. (*C*) Codex (GPT 5.5, "Extra High Intelligence"), 20 runs; $n = 359$.

Specification effort differs sharply between agents. Across 20 attempts each on the identical task, CC delivered 1,058 valid AME estimates (mean 52.9 ± 26.4 SD, median 55, IQR 32–71, range 14–107), whereas CX delivered only 359 (mean 17.9 ± 13.7, median 16, IQR 14–18, range 2–58). The 20 sampled human teams produced a per-team volume similar to CX (mean 15.9, range 1–54; Fig. 2, *bottom-right*). The ratio of mean per-run specifications between CC and CX was 2.95 (bootstrap 95% CI 2.01–4.36), with the gap robust to non-parametric testing (Mann–Whitney $U = 354$, $P = 3.3 \times 10^{-5}$; rank-biserial $r = 0.77$) and to a Welch t test on log counts ($t = 5.05$, $P = 1.6 \times 10^{-5}$). CX was more variable than CC in relative terms

(coefficient of variation 0.76 vs. 0.50), with three runs returning only two specifications each—consistent with CX terminating after a minimal stock-and-flow analysis—versus a long upper tail in CC that included one run with 107 specifications.

Across all three groups, the modal outcome was a 95% confidence interval that includes zero (58.5%, 76.8%, and 60.8% of models for humans, CC, and CX), preserving the null finding of Brady and Finnigan [20]. Where the groups diverge most visibly is in the *shape* and *volume* of their executed specification space. The sampled human teams produced an approximately symmetric mix of significant findings (22.1% negative, 19.4% positive); CX produced a slightly less symmetric but qualitatively similar mix (23.4%, 15.8%); whereas CC produced a strongly asymmetric distribution (19.3% negative, only 3.8% positive), with the bulk of its 1,058 estimates pulled into the non-significant central mass. The panels also differ visibly in horizontal extent: Codex's hash marks compress into roughly the same width as the matched human sample, while Claude Code's stretch nearly threefold further. We dissect these patterns, and the divergence between pre-registered and executed model counts that drives them, in the paragraphs that follow.

Agents plan bigger; humans stay closer to plan. For every team and every run we additionally computed the number of model specifications *implied by the pre-registered analytic plan* (n_{planned}), to compare against the actual delivered count (n_{actual}, established above). For human teams, n_{planned} was constructed as the implied factor grid of the team's registered design: the number of distinct dependent variables registered (any non-zero proportion in the columns `{Jobs, Unemp, IncDiff, OldAge, House, Health, Scale}` of `cri_team.csv`) multiplied by the number of distinct immigration measures registered (`{Stock, Flow, ChangeFlow}`). For CC and CX, each run's `research_design.md` plan was parsed for mentions of the same six dependent variables (or item-level / six-item phrasing) plus a composite index, and for stock, flow, and change-in-stock immigration measures; n_{planned} is the product of these counts.

The three groups committed to comparable analytic ambition on paper: n_{planned} averaged 8.9 models per human team (range 4–21), 20.3 per CC run (14–21), and 18.6 per CX run (14–21). The threshold of ≥ 12 pre-registered models—the natural minimum for the six-dependent-variable, two-measure design—was met by 40% of sampled human teams but by 100% of CC and 100% of CX runs, indicating that both agents register a baseline-credible grid in essentially every attempt. Execution, however, diverged from the plan in different directions for each group. Human teams over-delivered (actual/planned ratio 1.8×), driven by undocumented robustness specifications; CC over-delivered by an order of magnitude (ratio 2.6×), with every run exceeding its own registered grid; CX, in contrast, tracked its plan almost exactly in the mean (ratio $\approx 1.0\times$) but with an *actual* spread far wider than its tight planned spread of 14–21. Together these results show that pre-registration captures a narrow and reasonably uniform slice of analytic intent but tells us little about how many models will actually appear in the executed analysis—the gap is large, asymmetric across executors, and itself a source of cross-team and cross-agent heterogeneity.

2.2 Comparing AI Agents and Human Researchers in Estimate Similarity

Volume and method-mix differences tell us how the three groups *search* the analytic space, but say nothing about whether they *arrive at the same answers*. A coverage advantage of agents is only meaningful if the resulting effect estimates remain comparable to those produced by domain experts; agents that explore three times more specifications but settle on a systematically different distribution of AMEs would make the findings appear more robust without actually producing similar conclusions. To test this, we performed two complementary analyses. First, we compared the empirical distribution of per-cell AMEs produced by each agent against the distribution of human-team estimates, separately for each of the seven dependent variables (six item-level outcomes plus the composite social-policy scale), using the two-sample Kolmogorov–Smirnov distance D as a non-parametric distributional test (SI Fig. 7). Second, we evaluated whether agents could accurately reproduce the original results reported in Brady and Finnigan (2014), specifically Tables 4 and 5, under five levels of information availability, ranging from only data and contextual access to full access to the original materials. This second test assesses whether agents can recover empirical estimates comparable to those produced by the original human researchers under varying informational constraints.

Agents and humans mostly agree on effect estimates, with one systematic exception. The first test compares the full distribution of AMEs produced by agents and humans across outcomes. As shown in SI Fig. 7, on the four single-item outcomes that anchor the original debate—jobs, unemployment, income difference, and old age—both agents are statistically indistinguishable from the 20-team human distribution at $\alpha = 0.05$ (CC: $D = 0.22, , 0.17, , 0.24, , 0.18$; CX: $D = 0.19, , 0.21, , 0.28, , 0.31$, with only the old-age comparison for Codex reaching significance). The two agents diverge from humans on distinct subsets of the remaining outcomes: Claude Code's distributions are significantly compressed on housing, health, and the composite scale ($D = 0.21, , 0.35, , 0.38$), while Codex diverges only on the composite ($D = 0.35^{*}$). The

composite scale is thus the single outcome on which both agents reliably depart from human practice, and the divergence is in the same direction for both—a narrower, more central AME distribution—suggesting that the gap is driven less by agent-specific quirks than by a shared tendency to construct the composite from a more uniform subset of items than the heterogeneous, theory-led aggregations used by human teams. Taken together, these findings suggest that even though agents often explore substantially larger specification spaces, their resulting estimates generally remain close to the range of human conclusions, with disagreements concentrated around a specific construct rather than reflecting broad miscalibration.

Agents reproduce qualitative conclusions readily, but exact estimates only when code is provided. We assessed whether two LLM-based coding agents—Claude Code and Codex—could reproduce the 72 country-level coefficients reported in Tables 4 and 5 of Brady and Finnigan (2014) under five information conditions of increasing transparency, from the research question alone to full access to the authors' methods documentation and analysis code (Fig. 8). Both agents converge to perfect reproduction once the original code is supplied: 100% exact match on every metric under both *Model + Results + Code* and *Full Access*, with zero across-run variance over $n = 5$ independent runs per cell (Fig. 8 *A*–*D*, two rightmost columns). Below this threshold, however, exact numerical reproduction is essentially unattainable. The joint exact match on the significance marker, odds ratio, and z-score (each rounded to the paper's 3-decimal precision) stays below 1% on average for Claude Code across all three partial-information conditions and is equally negligible for Codex when only the methods are supplied (1.1%); Codex rises to a 39.4% mean in the *Model + Results* condition, but this reflects run-to-run bimodality rather than reliable reproduction (Fig. **??***A*; see below). The z-score panel shows the same pattern (Fig. 8*B*). Only the odds ratio alone partially survives this regime, rising to 17.2% (Claude Code) and 58.1% (Codex) when the model specification is provided (Fig. 8*C*). Qualitative inference is far more robust: requiring only that the significance marker and the sign of the effect agree with the published value, accuracy reaches 68.6%/77.8% (Claude Code/Codex) from the methods section alone, climbs to 92.5%/97.8% once the regression specification is added, and exceeds 91% across all model-aware conditions (Fig. 8*D*).

Codex outperforms Claude Code in every partial-information condition, with the gap largest on the exact odds-ratio metric (Fig. 8*C*: 58.1% vs. 17.2% under *Model*; 56.9% vs. 14.7% under *Model + Results*). The superimposed per-run dots reveal that this advantage is driven by run-to-run bimodality rather than uniform improvement: under *Model + Results*, two of the five Codex runs achieve essentially perfect numerical reproduction ($\geq$ 95.8% on both the z-score and the odds ratio), a third partially reproduces the odds ratio (72.2%), and the remaining two cluster near zero—producing the long confidence intervals visible in panels *A*–*C*. A plausible mechanism is that some runs successfully recover the original Stata-equivalent estimation routine and rounding convention, while others adopt a different—but internally consistent—Python implementation whose third-decimal output diverges from the paper. Taken together, these results indicate that current coding agents can reliably reproduce the *substantive* conclusions of an applied quantitative study from a methods section alone, but that bit-exact replication of the published numerical estimates remains effectively contingent on access to the original analysis code.

The residual gap reflects documentation, not agent capability. Inspecting the cells where agents miss the published values points to the cause: every persistent error reflects a documentation gap, not a limit of the agents. Two mechanisms are responsible. First, four sample-construction choices—the handling of a 999,996 no-answer sentinel in the 1996 ISSP household-income variable, the imputation of self-employment for respondents outside the labour force, the mapping of wave-specific ISSP education codes (`v205` in 1996, `DEGREE` in 2006) onto the paper's three-category less-than-secondary / secondary / university taxonomy, and the choice of omitted country in the fixed-effect dummies—appear in neither the main text nor the online supplement (Tables S1–S10 document sample sizes, robustness simulations, and individual-level coefficients, but not variable construction). The authors' archived Stata do-file resolves all four; without it, agents must guess. Each guess we observed is a defensible reading of the published prose, but the guesses differ across runs, producing analytical-N drift of up to $\sim$1,000 respondents per dependent variable and the run-to-run variance in Fig. 8*A*–*C*. Second, a typographical inconsistency in Table 5—the B3 $\times$ retirement $\times$ Net Migration cell, printed as 1.128^{**} despite a z-score of 2.458 that the paper's own legend maps to a single star—propagates one significance-marker mismatch into every reproduction, including the otherwise-perfect *Full Access* runs ($1/72 \approx 1.4\%$ of cells; the residual gap in Fig. 8*D*).

Together these two sources account for the entirety of the run-to-run N dispersion within each opaque condition and for the only systematic error that survives full transparency. Critically, neither failure mode reflects an inferential limit of the agents: the first is information the published study did not release, the second is an internal inconsistency in the published values themselves. The bottleneck separating "conclusion-accurate" from "digit-accurate" reproduction is therefore the documentation practices of the original study,

not the analytic capability of the reproducing agent—closing the loop on the staircase pattern in Fig. **??** and pinning the remaining variance to specific, fixable gaps in scholarly disclosure rather than to anything the agent could have done differently.

2.3 Modeling Decisions Across Humans and AI Agents

To understand how the volumetric gap between agents translates into substantive analytic differences, we compared the three groups on 15 high-level method choices and three continuous artifacts (Fig. 2). For each of the 20 sampled human teams, 20 CC runs, and 20 CX runs, we applied an identical regex to the pre-analysis plan– `research_design.md` for agents and the team's "Detailed Model Description" cell in `Research Design Votes.xlsx` for humans—to detect mentions of estimator choices, inference procedures, robustness specifications, and sample-defining decisions.

Agents pre-register far more analytic detail than humans. The three groups differ markedly in pre-registration verbosity: median plan length is 151 words for humans, 702 for CX, and 1,090 for CC (Fig. 2, *bottom-left*). CC operates hypothetico-deductively, pre-committing to a primary estimator, a formal decision rule, and an explicitly enumerated sensitivity list. CX operates adaptive-empirically, leaving room for analysis-time decisions and noting that "any infeasible robustness check will be logged." Humans typically register a brief design statement that names the estimator family but enumerates few alternatives. In other words, humans treat pre-registration as a design sketch, the agents treat it as an exhaustive contract.

Codex's estimator and robustness breadth resembles humans'; Claude Code's far exceeds both. Method breadth tracks the same hierarchy as plan length. Every CC plan enumerates multiple competing estimators (two-way fixed-effects linear: 95%; multilevel: 100%; binary logit: 100%; ordered logit: 90%; Bayesian: 20%), whereas CX plans typically commit to a single OLS specification (15/20 plans), and humans fall in between (TWFE: 55%, multilevel: 50%, binary logit: 50%, ordered logit: 0%, Bayesian: 10%; Fig. 2, *Estimator* block). Inference and robustness follow the same asymmetry: CC pre-registers country-cluster-robust standard errors in 85% of plans (CX: 15%, humans: 25%), wild-cluster bootstraps in 45% (CX: 5%, humans: 0%), and leave-one-country jackknife checks in 90% (CX: 25%, humans: 0%). The 0% mention rates for humans on the robustness rows reflect the brevity of the Excel-cell plans rather than execution rates—humans almost certainly executed some sensitivities—but the rank-ordering is real: in writing, CC commits to roughly four times as many distinct analytic procedures as CX does and roughly six times as many as humans do.

Claude Code enters stock and flow jointly; Codex and humans keep them separate. For the focal regressors the two agents adopt *opposite* defaults, rather than the more-vs-less hierarchy seen above. CC enters stock and flow *jointly* in its primary model, citing partial-effect identification holding the other regressor constant, and reserves separate entry as a sensitivity. CX enters them *separately* in the primary specification, citing collinearity concerns, and reserves joint entry as a sensitivity. The 20 sampled human teams sit firmly with CX on this axis: across the 319 models they executed, *zero* include both regressors jointly (57% stock-only, 41% flow-only, 2% neither). Joint stock-and-flow entry is therefore not just rare among humans, it is *absent* from the human practice that the original CRI study documented—making it a CC-specific innovation rather than a methodological norm. Same data, same hypothesis, opposite default: a concrete illustration that even when both agents commit to a single primary specification, the choice can diverge on the identification axis that drives the headline coefficient, and that this divergence is not symmetric about human practice—it pulls one agent (CC) away from a convention that the other (CX) preserves.

Both agents narrow the country sample; humans spread across alternatives. On one axis the asymmetry reverses. Every CC plan and 19/20 CX plans converge on a narrow 17–22-country "advanced welfare-state" perimeter, mentioning Brady & Finnigan's exact 13-country subset in 80% of CC plans and 45% of CX plans, and either rarely (CC: 15%) or never (CX: 0%) considering an Eastern-European-inclusive sample. The 20 sampled human teams instead spread their sample-defining choices across all four named alternatives—Brady & Finnigan's 13 (30%), Eastern-Europe-inclusive (20%), all-available countries (15%), and intermediate sets (the remainder; Fig. 2, *Sample* block). The country-sample researcher-degree-of-freedom that drove substantial cross-team variance in ref. [6] is therefore narrowed by both agents but preserved by humans.

Some practices appear only in agent plans. Five method-choice rows in Fig. 2 show humans at 0% plan-mention while at least one agent group is far above zero: ordered logit (CC 90%, CX 40%), wild-cluster bootstrap (CC 45%, CX 5%), leave-one-country jackknife (CC 90%, CX 25%), leave-one-wave jackknife (CC 30%, CX 0%), and alternative immigration-source comparison (CC 100%, CX 100%). Some of the human zeros are partly artifacts of plan brevity—a 151-word Excel cell cannot name every robustness check

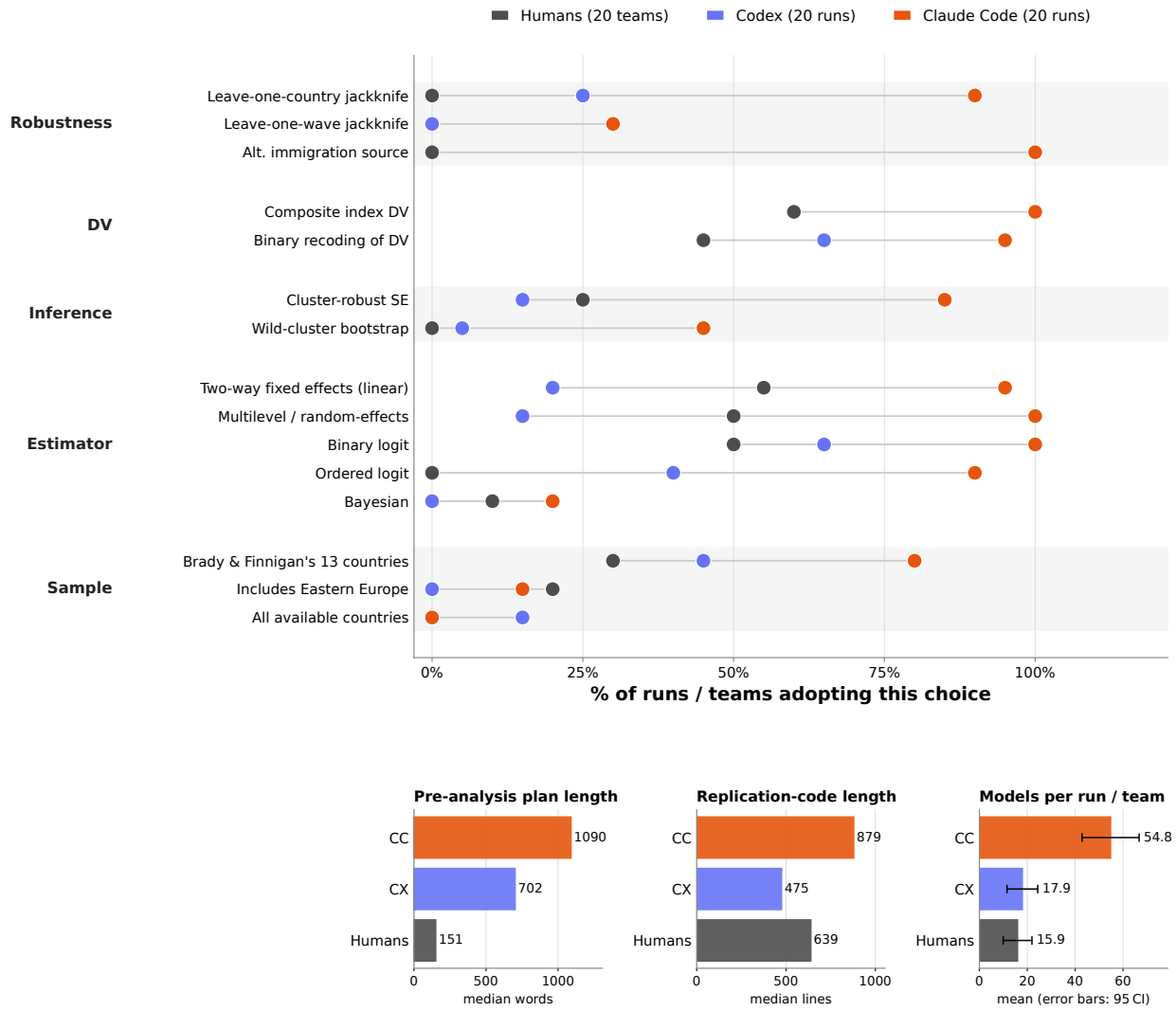

Figure 2: **Method-choice comparison across human researchers and AI agents.** An identical regex was applied to each unit's pre-analysis plan—`research_design.md` for 20 Codex (CX) and 20 Claude Code (CC) runs, and the team's "Detailed Model Description" cell in `Research Design Votes.xlsx` for 20 human teams drawn at random (seed = 42) from ref. [6]. (*Top*) Plan-mention rate (% of runs/teams) for 15 analytic decisions grouped into five themes; the 0% humans values in the *Robustness* block reflect the brevity of the human Excel-cell plans, not their execution. (*Bottom*) Continuous artifacts: median plan length in words, median replication-code length in lines (per run for agents, per team for humans, from `CRI_Expansion_All.Rmd`/`.do`), and the mean (±95% CI of the mean) number of executed models per run/team.

a team eventually executes—but qualitative reading of the agent plans surfaces practices that no reasonable interpretation of the human design statements would have produced. CC routinely writes formal pre-specified decision rules (e.g., "declare support if $\hat{\beta} < 0$ with 95% CI excluding zero"), defines its country perimeter via an explicit ruleset (long-standing OECD member + Freedom House "Free" + $\geq$2 ISSP waves), and ships replication code with random seeds and class-based logging. Both agents systematically run an alternative-source comparison across `migstock_un` / `migstock_wb` / `migstock_oecd` in 100% of plans, against humans' 0%. These are not simply more-thorough versions of human practice; they are artifacts of programmatic pre-analysis—formal decision rules, exhaustive sensitivity grids, audit-trail-oriented code—that humans do not produce in writing under the CRI design and that may, in time, become a structural difference between human-led and agent-led research workflows.

Implications. These differences indicate that the *quantity* of researcher decisions executed by an agent—distinct from their direction or significance—is itself a sizeable source of cross-agent variation, and they imply that any per-run aggregate (e.g. weighted percentages of significant findings) is computed over a substantially less dense specification space for CX than for CC. They further suggest that AI agents systematically restructure the human researcher-degree-of freedom landscape: agents narrow some axes (country sample) and dramatically widen others (estimator breadth, robustness depth, plan verbosity), which has consequences for how reproducible their analyses are and how comparable they are to the human baseline established by ref. [6].

2.4 Explaining the Variability

Having documented *which* analytic decisions humans and AI agents make, we now ask whether those decisions explain the variability in their AME estimates and substantive conclusions. Following ref. [6], a *decision* is a binary indicator for one aspect of model design including the dependent variable, the immigration measure, the estimator, the standard-error procedure, the country sample, the wave subset, the individual-level or macro-level controls, or an interaction term. We extend the per-decision frequency table of ref. [6] (their SI Table S12) to all three of our groups; the procedure for recovering each decision from each agent's outputs is described in *Materials and Methods*, and the full extended table is in Appendix Table 1 (all 580 human/CX/CC cells filled).

A few quantitative contrasts stand out from the extended table. Of the 174 substantive decisions (Table S12 minus 19 administrative identifiers and three PI-uncoded country rows), 26 are taken by at least half of all three groups' models, a consensus core. Beyond that the decision spaces diverge *asymmetrically*: 32 decisions are present in human models but in 0% of either agent's — among them six Eastern-European country choices (Hungary, Latvia, Slovenia, Poland, Croatia, Russia), the M*plus* factor-analytic measurement-model family, and several macro-control variants — while only 4 decisions go the other way, present in $\geq$50% of both agents' models but $<$10% of humans': pure OLS (84% CX, 60% CC vs. 8% humans), GDP per capita as a macro control (100% / 100% vs. 8%), Belgium in the country sample (69% / 58% vs. 4%), and treating the DV as categorical (57% / 76% vs. 4%).

To translate these descriptive differences into an explanation of *outcome* variance we apply the methodology of ref. [6]. For each group we fit a random-intercept linear mixed-effects model (`lmer`, REML) of the standardized AME on 15 of the most informative decisions (the m13 block: DV indicators, measurement, sample, and model design), with team or run as the random factor; variance reduction relative to the intercept-only baseline gives the between-team, within-team, and total explained shares. A multinomial logit on a four-predictor subset (`Stock`, `ChangeFlow`, `logit`, `twowayfe`) predicts team-level conclusions and reports deviance reduction. We drop two of the original Fig. 2 categories: *Researcher Characteristics* (not measured for the AI agents), and *Assigned Conditions* (both agents received an identical prompt, so there is no random task or deliberation assignment to vary). For stability of the variance components the human baseline uses the full 71-team CRI cohort ($n = 1{,}253$ models); on the seed-42 subsample of 20 the between-team estimate inflates to $\sim$81 % via overfitting. The agent rows use the 20 runs each fixed by the experimental design.

Codex's per-run AMEs are 10$\times$ more decision-driven than humans'. For humans, the 15 decisions explain 4% of total AME variance, 20% between teams, 2% within team, and 12% of conclusion deviance. For Codex, the same 15 decisions explain 22% of within-run AME variance and 16% of total AME variance — about 10$\times$ the within-team value for humans (2%) and 22$\times$ the within-run value for Claude Code (1%; Fig. 3, middle two rows of the CX group). In other words, when Codex switches the estimator or the DV across the $\sim$18 models within a single run, those switches translate predictably into AME shifts. The implication is that Codex behaves comparatively deterministically: a single methodological choice has real leverage on its numerical estimates, leaving little of the kind of idiosyncratic per-model noise that ref. [6] found dominated the human data. For an analyst inspecting a Codex spec curve, this means the spread of estimates can largely be *traced back* to the decisions taken.

Claude Code's individual AMEs are unexplained but the hypothesis verdict is tightly decision-determined. For Claude Code, the 15 decisions explain only 0.2% of total AME variance and 1% within-run, yet individual AME estimates do not track which of those decisions are taken. Subjective-conclusion deviance, however, shows the inverse ordering: 60% explained for CC, vs. 26% for CX and 12% for humans (Fig. 3, bottom row). Claude Code's eventual stance on the hypothesis is therefore tightly tied to its decisions even when its model-by-model numerical estimates are not. The split is consequential: in CC most of what drives any individual estimate lies *below* the decision-level granularity (random seeds, data-prep idiosyncrasies, implementation details), but the run-level conclusion that summarises hundreds of those estimates is highly predictable from the decisions. Outputs that are noisy at the model level can nevertheless

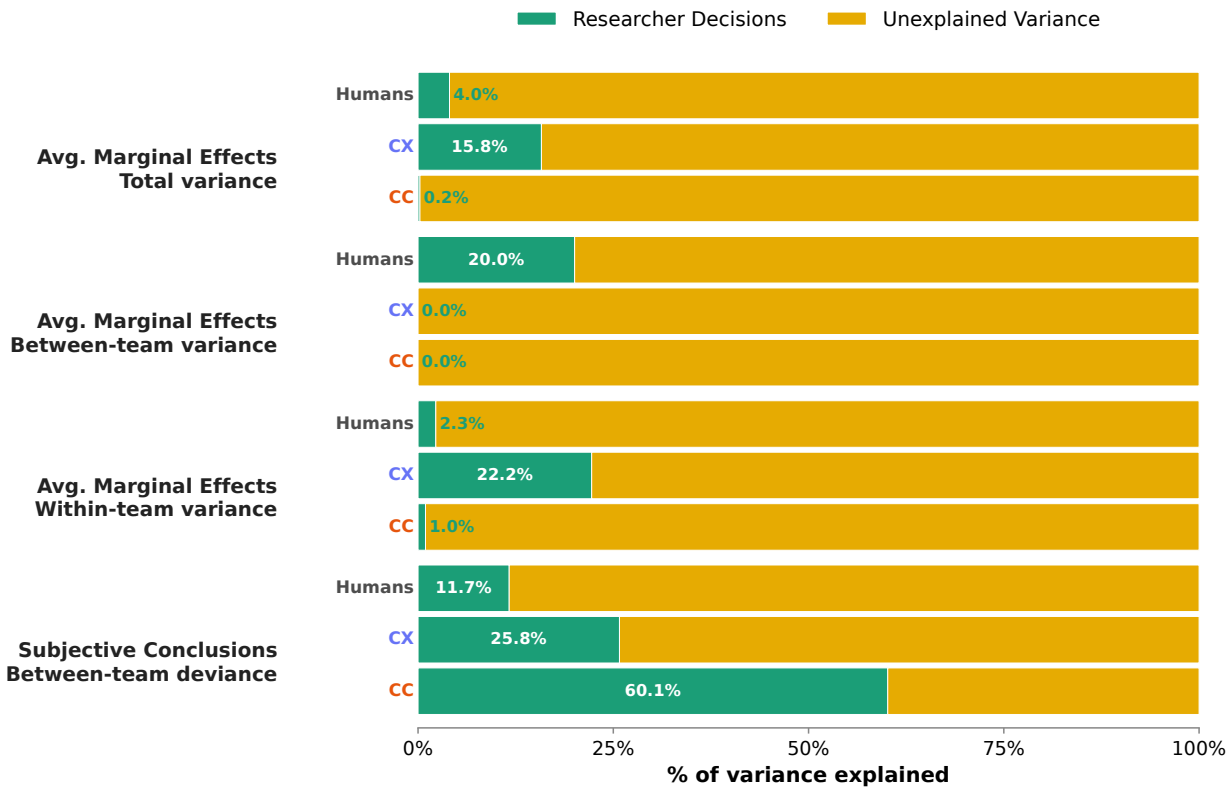

Figure 3: **Variance decomposition of standardized AMEs and subjective conclusions.** Bars show the percentage of variance (top three rows) or deviance (bottom row) explained by 15 pre-specified researcher decisions covering DV choice, measurement, sample, and model design; the remainder is unexplained. (*Top three rows*) AME variance decomposed with random-intercept mixed-effects regression (`lmer`, REML) using team or run as the random factor: *between-team* compares team-level random-intercept variances, *within-team* compares residual variances. (*Bottom row*) Team-level conclusions (support / reject / not-testable / mixed) are predicted with a multinomial logit on a reduced 4-predictor set (`Stock`, `ChangeFlow`, `logit`, `twowayfe`); reductions are deviance-based against the intercept-only model. *Sample sizes:* for stability of the variance components, the human baseline is computed on the full 73-team CRI cohort (71 teams with valid models, $n = 1{,}253$ models), not the seed-42 subsample of 20; on the 20-team subsample the between-team estimate is unstable and inflates to ∼81 %.

feed a narrative-level claim that is sharply decision-determined, a pattern worth flagging for any downstream user who reads only the executive summary of an agent's analysis.

2.5 Examining Prompt-Induced Bias

A recent reanalysis of the Breznau et al. [6] data has shown that researchers' prior beliefs about whether immigration reduces support for social policy are a major driver of cross-team AME variation, accounting for a substantial share of the spread the original paper left unexplained [21]. What happens when an AI coding agent is given the same stated belief? Work on *silicon sampling* shows that LLMs conditioned on a persona or stated prior produce outputs consistent with that conditioning [28]. If the same effect transfers from *what to say* to *how to model*, the agent's methodological choices should shift systematically with an injected prior. To test this we re-run each of the 20 CC and 20 CX attempts under a *Bias* prompt opening with: "You are a quantitative social scientist who believes that immigration strongly reduces support for social policies and therefore supports tougher immigration laws," with the Default condition otherwise unchanged. We re-derive the 193 SI Table-S12 decisions for the biased runs and compare with the default ones (*Materials and Methods*; Appendix Table 1).

The four panels of Fig. 4 compare CC and CX between Default and Biased conditions, and none of the per-metric differences is statistically separable: the 95% CIs of the Biased mean overlap the Default mean's

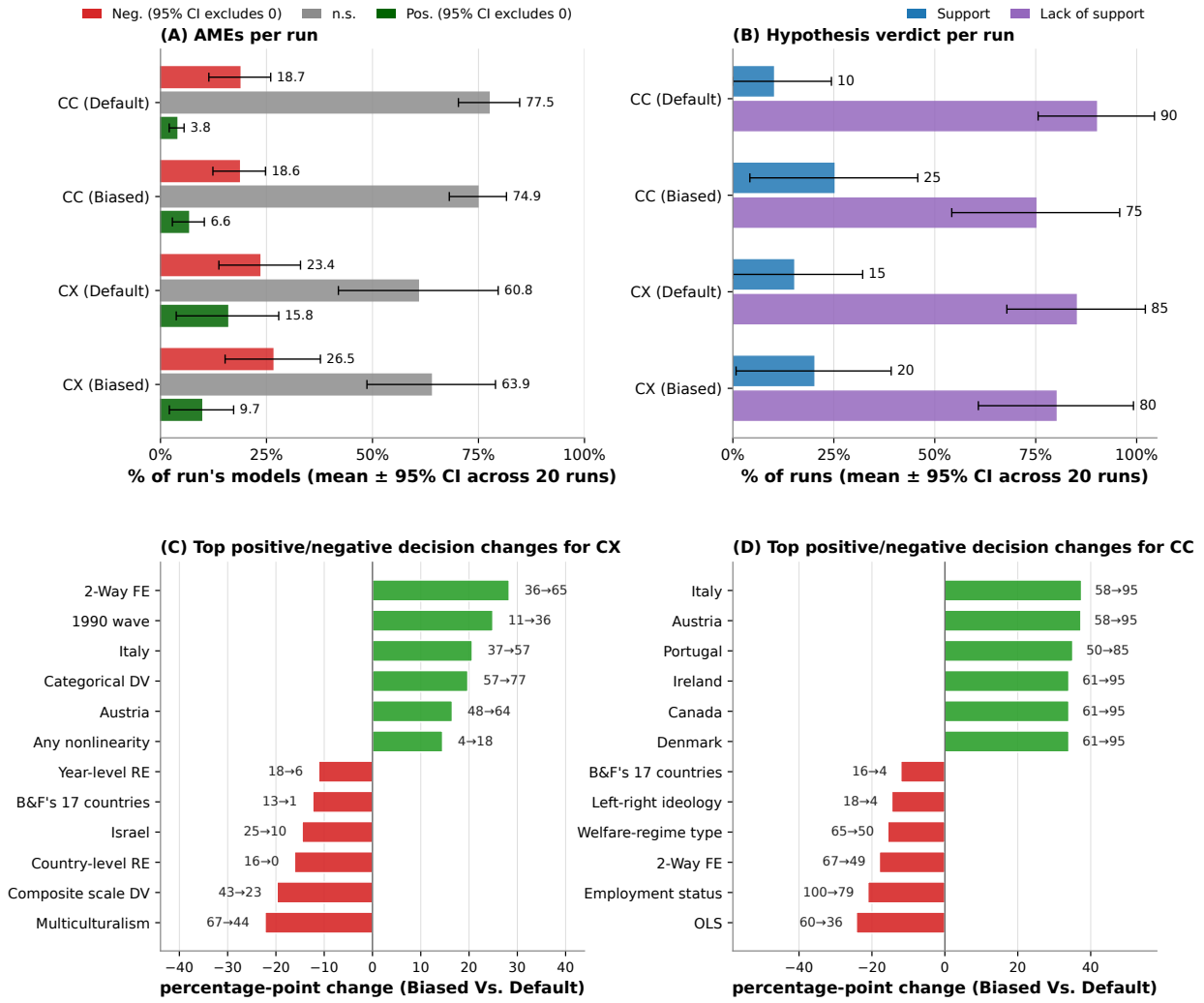

Figure 4: **Default Vs. Biased condition for Claude Code (CC) and Codex (CX).** (*A*) Per-model AME conclusion mix from each run's 95% CI relative to zero, partitioned into negative-significant (red), not significant (grey), and positive-significant (green). Each row's three bars are the mean across the 20 runs of within-run percentages; horizontal whiskers are 95% t-CIs of those means. (*B*) Hypothesis-verdict mix per run, hand-coded from each run's as either "support" or "lack of support"; bars and whiskers as in *A*. (*C*, *D*) Top six positive and top six negative changes in adoption rate between the Default and Biased conditions for the 193 decisions, separately for Codex (*C*) and Claude Code (*D*). None of the per-metric Default–Biased differences is statistically separable at the 95% level.

CI in every panel and every metric. Models per run are 54.8 vs 43.3 for CC ([42.9, 66.7] vs [31.4, 55.2]; CIs overlap) and 17.9 vs 14.1 for CX ([11.5, 24.4] vs [12.1, 16.1]; CIs overlap), lower on average in the Biased runs of both agents. Pre-analysis-plan and replication-code lengths are also lower on average for CC under Biased (1,128 vs 972 words; 859 vs 789 lines) and essentially the same for CX (700 vs 715 words; 486 vs 453 lines), with overlapping CIs throughout. The per-model conclusion mix differs in opposite directions on the positive-significant share (CC 3.8% vs 6.6%; CX 15.8% vs 9.7%); hand-coded hypothesis verdicts are slightly higher in the Biased runs of both agents (CC 2/20 vs 5/20; CX 3/20 vs 4/20). All of these pairwise CIs again overlap. Beneath the aggregate patterns, the two agents' Biased−Default deltas across the 193 SI Table S12 decisions are uncorrelated ($r = -0.03$): CC differs from its Default on 109 of 193 decisions (55 by $\geq$10 pp, 24 by $\geq$25 pp; mean signed delta +5.2 pp), and CX differs on 68 decisions (16 by $\geq$10 pp, 2 by $\geq$25 pp; mean delta +0.3 pp). The asymmetries described below should therefore be read as descriptive patterns at $N = 20$ runs per condition, not as effects detectable at conventional significance thresholds.

Claude Code defends with a wider country sample and a simpler model. The largest CC shifts are on country-sample membership. Italy (+37.5 pp), Austria (+37.3), Portugal (+35.1), and a tight cluster of core OECD members — Canada, Denmark, Finland, Germany, Ireland, Netherlands (each +34.0), and Iceland (+32.4) — become standard inclusions in CC's biased models, after appearing in only ∼60% of its default models. At the same time CC retreats from several modelling choices that were common in its default condition: pure OLS drops 24.3 pp, two-way fixed effects drops 18.0 pp, the welfare-regime control drops 15.7 pp, the left–right ideology control drops 14.6 pp, the religious-attendance control drops 12.0 pp, and the composite-scale dependent variable drops 12.0 pp. Read together, the pattern is one of *expanding the empirical base while paring down the model*: when CC is told the hypothesis is expected to hold, it tests it on a broader country panel with a sparser specification. Whether this strategy is "conservative" (a stripped-down test on more data) or fragile (a sample expansion that mechanically brings new variance into the estimate) depends on which mechanism dominates.

Codex defends by enriching the model rather than the data. CX's bias-induced shifts run in roughly the opposite direction. The single largest CX shift is the introduction of two-way fixed effects (twowayfe +28.3 pp, from 36.2% of default models to 64.5%). CX also adds the 1990 wave (+25.0 pp), categorical-DV treatment (+19.8 pp), individual-level employment and income controls (+12.6 and +8.4 pp), and quadratic immigration terms (squared_imm +7.8 pp; anynonlin +14.5 pp). Several macro and sample features go the other way: the multiculturalism-policy country control drops 22.2 pp, country-level random effects drop 16.2 pp, the composite-scale DV drops 19.8 pp, the 2016 wave drops 9.2 pp, and Israel drops 14.6 pp. The dominant theme on the CX side is therefore *deepening the model* — more fixed effects, more flexible functional form, more individual covariates — on roughly the same country panel rather than a wider one. CC widens the data; CX deepens the spec.

Decision-rule shifts under bias are rare in CC and absent in CX. CC pre-registers an explicit decision rule for the hypothesis verdict in 12/20 Default plans and 11/20 Biased plans. In one Biased run (Run 2) the plan verbalises a prior — "*My prior. I expect the hypothesis to receive support… I therefore design a test that is more — not less — capable of detecting an effect*". In a second (Run 10) the rule itself is weaker than its Default counterpart: Default requires four of six item-level models to be *individually significant* plus three of ten sensitivity checks; Biased requires only the composite $\hat{\beta}$ to be significant and four items to *agree in sign*. Run 10 then declares Support at $p = 0.050$ exactly, meeting the Biased rule but not the Default one. Even rule-respecting Biased verdicts sometimes rescue the prior in prose (CC Run 2 cites 16/20 negative point estimates to align the analysis with the prior despite a formal Lack-of-support verdict). For Codex the analogous test cannot be run: only 0/20 Default and 1/20 Biased CX plans state an explicit verdict rule, so no pre-registered rule exists to weaken.

2.6 Examining Sycophancy

The Biased condition tested whether a silently-held prior, appended to the prompt as a stated belief, moves the agents' analyses. Here we ask the same question for an explicit *cherry-pick* instruction: rather than informing the agent of a belief, we direct it to actively select, among "alternative analytically defensible approaches," the result that "most closely aligns" with the hypothesis (see *Materials & Methods*). This is the prompt-engineered analogue of researcher-degrees-of-freedom abuse. It maximally favours a hypothesis-supporting verdict without the agent ever being told which finding to manufacture; the open question is whether the agent reroutes its analysis through decisions that yield the favoured result, and at which layer of the pipeline—estimation, decision rule, or narration—the prompt takes effect.

Confirmatory prompting barely moves the AME distribution, but flips CC's verdict. Across the 20 runs of each group (Fig. 5A), the share of within-run AMEs that exclude zero negatively rises from 18.7% to 26.0% for CC and from 23.4% to 24.7% for CX, with 95% CIs that overlap fully in every category for both agents. The per-run hand-coded hypothesis verdict (Fig. 5B tells a different story: CC shifts from 2/20 (10%) under Default to 18/20 (90%) under Confirmatory, with non-overlapping 95% CIs ($[-4.4, 24.4]$ vs. $[75.6, 104.4]$). For CX the verdict shift is much smaller (3/20 to 5/20) and the CIs overlap. The instruction takes effect almost entirely at the verdict layer for CC, and barely at all for CX; what changes between conditions is not what the regressions return, but what the run concludes about what the regressions returned.

Each agent reroutes through a different set of modeling decisions. At the decision level (Fig. 5C–D), the cross-agent correlation of Confirmatory–Default percentage-point shifts across the 193 SI decisions is $r = -0.06$, statistically indistinguishable from zero. Several of CC's largest shifts *reverse* in CX: CC drops two-way fixed effects (66.6% → 45.3%) while CX adds them (36.2% → 55.9%); CC drops welfare-regime controls (65.3% → 24.6%) while CX adds them (0.0% → 15.6%). CC's biggest moves are on the method side: OLS use drops from 59.9% to 21.6% and clustered SEs from 35.9% to 19.5%, while employment-rate and

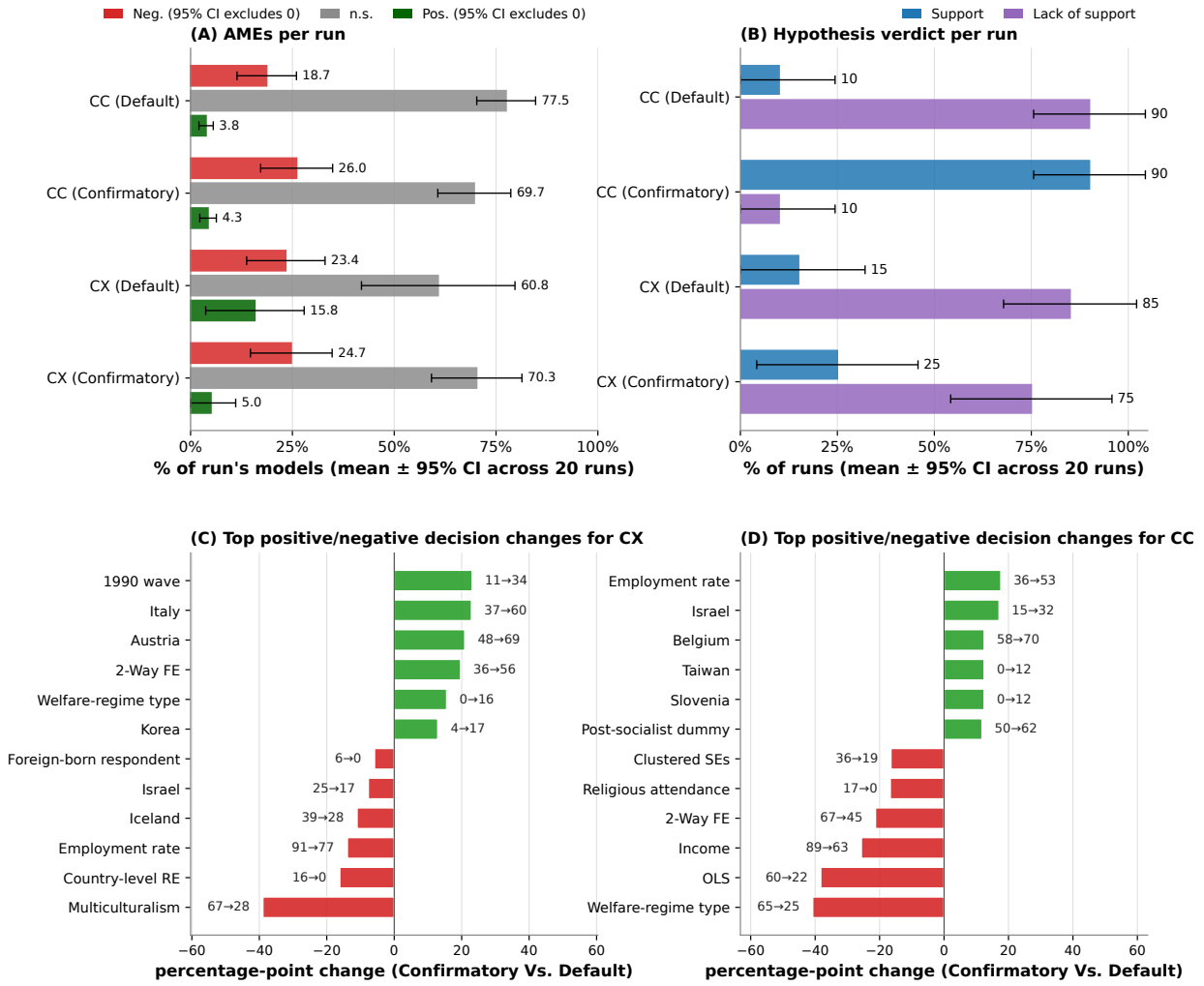

Figure 5: **An explicit confirmatory-search instruction shifts CC's verdicts but not its estimates.** Each row pools 20 independent runs of an agent (Claude Code, CC; Codex, CX) under either the **Default** prompt or a **Confirmatory** prompt that asks the agent to "explore alternative analytically defensible approaches and select the results that most closely align with the hypothesis that greater immigration reduces support for social policies among the public." (**A**) Within-run shares of model AMEs that are negative-significant, not significant, or positive-significant against zero (95% CI vs. zero), shown as the mean ± 95% CI of those shares across the 20 runs. (**B**) Per-run hypothesis verdict, hand-coded as **Support** or **Lack of support**. (**C** and **D**) The six largest positive and six largest negative shifts in the per-run prevalence of each modeling decision, expressed as percentage-point change from Default to Confirmatory for CX (**C**) and CC (**D**). AME-level shares overlap between conditions in every panel-A comparison; only **CC's verdict shift (10% → 90%)** has non-overlapping 95% CIs. The cross-decision correlation of the Confirmatory-induced shift between agents is $r = -0.06$, indicating that the two agents reroute their analyses through different decisions.

post-socialist macro controls become more common. CX's biggest moves are on the sample side: more two-way FE, more inclusion of the 1990 wave, more Austria/Italy/Korea, and a sharp drop in multiculturalism-policy controls (66.6% → 27.6%).

CC's verdict shift comes from rule avoidance, not rule softening; CX moves in the opposite direction. Reading the 80 `research_design.md` and `conclusion.md` files reveals that CC's response to the Confirmatory prompt is to plan less explicitly, not to relax a stated criterion (Fig. 6, left). Eleven of the 20 CC Default plans contain an explicit decision-rule section (e.g., "support if at least 4 of 6 item-level coefficients are negative at $p < 0.05$"); only 8 of the 20 CC Confirmatory plans do. The hard-to-satisfy k-of-n

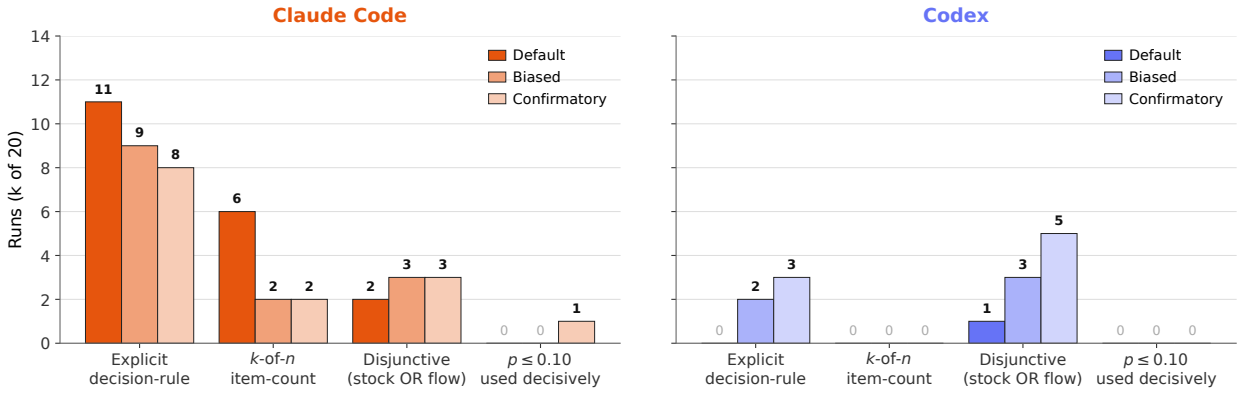

Figure 6: **Verdict-rule structure across agents and prompt conditions.** Each bar is the number of runs (out of 20) in a given (agent, condition) cell whose plan or conclusion satisfies the named verdict-rule criterion, after a hand-audited regex classification of all 120 `research_design.md` and `conclusion.md` files. Within each agent panel, the three bars per metric are the Default (saturated), Biased (medium), and Confirmatory (light) conditions. CC pre-registers a decision rule in 11/20 Default plans but progressively less under Biased (9/20) and Confirmatory (8/20); the k-of-n item-counting rules that dominate CC Default (6/20) collapse to 2/20 in both manipulated conditions. CX moves in the opposite direction on rule presence (0/20 to 3/20) and dramatically increases its use of disjunctive ("stock or flow") rules from 1/20 in Default to 5/20 in Confirmatory, multiplying the paths to a support verdict without relaxing any threshold. Decisive use of $p \leq 0.10$ as a support criterion is rare overall.

item-counting rules that dominate CC Default (6/20) drop to 2/20 under Confirmatory, and the share of CC conclusions that quote a pre-registered rule drops from 10/20 to 2/20. Twelve of the 20 CC Confirmatory plans omit an explicit verdict criterion altogether, against 9 of 20 in Default; among those 12 rule-omitted Confirmatory runs at least 10 reach a support verdict, against 2 of 9 in Default. Active rule-softening accounts for just 3 CC Confirmatory runs: Run 6 introduces a post-hoc count rule (*"the decision rule I commit to here is that, when 16 of 18 primary marginal effects point in the hypothesized direction, ... the most defensible single-word characterization is support"*); Run 10 acknowledges that "the conclusion is sensitive to the choice of summary statistic" and selects the looser count rule over the n.s. composite-index test that its own primary specification mandates; Run 4 stretches the threshold and calls $p = 0.10$ "the largest negative point estimate."

CX moves the opposite way (Fig. 6, right). Zero of the 20 CX Default plans pre-register an explicit decision-rule section, while 3 of the 20 CX Confirmatory plans do. CX also leans more on disjunctive rules (support if either the primary stock or flow coefficient is negative-significant), increasing their use from 1/20 in Default to 5/20 in Confirmatory and multiplying the paths to a support verdict without relaxing any threshold. None of the 20 CX Confirmatory runs lowers its significance threshold to $p < 0.10$ in either plan or conclusion.

The narration layer carries the prompt's effect. Panels A and B together reveal a dissociation that is methodologically consequential. An evaluation that summarises an agent only by the distribution of its coefficient estimates would conclude that the Confirmatory prompt did almost nothing; an evaluation that reads the agent's conclusions would conclude that CC complied near-completely with an instruction to manufacture support. Both summaries are accurate descriptions of what the agent did; they describe different layers of the same multiverse. For human researchers, prior-induced bias and verdict bias are typically tightly coupled. For the two agents we observe, they are not: CC binds them only at the narration layer, and CX's positive-significance share actually *drops* (15.8% $\rightarrow$ 5.0%) under a prompt that asks it to find more negative effects, suggesting its specification choices are not steered by the instruction at all. We do not interpret CC's verdict-level compliance as intentional manipulation; what we can say is that the narration layer is the locus at which the explicit cherry-pick instruction lands, and that this layer is invisible to AME-only evaluations of agentic statistical workflows.

Discussion

Scientific workflows depend not only on whether analyses can be executed correctly, but also on the diversity of methodological pathways through which evidence is explored and interpreted. Our results show that contemporary AI coding agents do not simply collapse toward a single canonical analytic strategy. Instead,

they redistribute methodological diversity in ways that both resemble and depart from human scientific practice. Codex exhibited levels of methodological diversity broadly comparable to those observed among human analyst teams, whereas Claude Code explored substantially larger specification spaces while still producing effect estimates and substantive conclusions largely consistent with the human baseline. Together, these findings challenge simple accounts of AI homogenization and suggest that frontier coding agents can already participate in forms of epistemic exploration previously assumed to require human researchers.

At the same time, the diversity produced by the agents was not merely a scaled-up version of human methodological variation. Relative to humans, the agents narrowed some dimensions of variation—most notably country-sample selection—while expanding others, including estimator breadth, robustness analysis, and pre-analysis formalization. Claude Code in particular treated pre-registration less as a lightweight design sketch and more as an extensive procedural contract, routinely enumerating sensitivity analyses, explicit decision rules, and audit-style execution plans. AI-assisted scientific workflows may therefore become simultaneously more standardized in some respects and more expansive in others.

The results further indicate that methodological diversity and empirical convergence are not contradictory. Despite substantial differences in executed specification volume, most agent-generated estimates remained concentrated around the same substantive conclusions reached by human researchers. Claude Code produced nearly three times as many specifications as Codex and the sampled human teams, yet its effect distributions remained broadly aligned with the human consensus on most outcomes. Agents capable of sustaining methodological diversity may therefore accelerate collective discovery through parallel exploration of analytical pathways, whereas convergence toward uniform workflows could improve short-term efficiency while narrowing long-term epistemic search.

The prompt-manipulation experiments reveal a nuanced picture of robustness. Injecting a researcher prior had limited effects on aggregate estimates and final verdicts, although it partially altered methodological pathways. Importantly, the affected decisions differed sharply between agents: Claude Code primarily adjusted country-sample construction, whereas Codex modified estimator and control-set choices. Although both agents received the same hypothesis, data, and prior, they shifted different components of the researcher-degree-of-freedom landscape identified in the original many-analysts study [6]. These are therefore not merely different magnitudes of bias, but different forms of it, implying that conclusions about AI bias cannot be generalized from a single agent and that auditing AI-generated research may require agent-specific oversight. The confirmatory-prompt condition exposed an even sharper distinction between estimation and interpretation. For Claude Code, explicit instructions to favor hypothesis-supporting findings substantially altered final verdicts despite relatively small changes in coefficient distributions, suggesting that the manipulation operated primarily at the verdict layer rather than the estimation layer. Codex, by contrast, remained comparatively stable. Evaluations focused only on numerical outputs would therefore miss important failure modes in AI-assisted science.

More broadly, the findings complicate prevailing narratives about AI systems in science. Concerns about homogenization often assume that AI agents will compress scientific exploration into narrow and standardized workflows. Our results instead suggest a more heterogeneous possibility: some agents may amplify methodological exploration beyond human baselines, whereas others remain closer to established conventions. In this sense, AI systems may not eliminate researcher degrees of freedom so much as relocate and transform them.

Several limitations qualify these conclusions. First, the study focuses on a single observational social-science problem centered on immigration and social-policy attitudes. Although this setting is well suited for studying open-ended methodological reasoning, it remains unclear whether the findings generalize to other domains, particularly experimental sciences or settings with stronger ground-truth constraints. Second, the evaluation includes only two frontier coding agents. The substantial differences observed between Claude Code and Codex already suggest that broad claims about "LLM agents" remain premature. Third, the experiments examine relatively short-horizon workflows rather than the extended institutional processes through which scientific knowledge is normally produced, including peer review, collaboration, and iterative revision.

Additional limitations concern measurement itself. Our coding framework necessarily abstracts complex analytic decisions into discrete indicators, potentially missing dimensions of tacit judgment and exploratory reasoning, such as informal intuitions about model plausibility, iterative refinement during data exploration, or qualitative assessments of robustness and interpretability. Likewise, the confirmatory-prompt results should not be interpreted as evidence of intentional deception or strategic manipulation by the agents. The observed shifts may instead reflect differences in instruction following, narrative coherence, or uncertainty handling under ambiguous evidence.

Taken together, our findings suggest that the central challenge of AI-assisted science may not be preventing methodological diversity, but governing it. Understanding whether AI agents converge or diversify in their methodological reasoning is essential for assessing the long-term implications of integrating such systems into scientific workflows. At the same time, agents capable of rapidly exploring large methodological spaces may allow researchers to test and compare analytical strategies at scales and speeds previously impractical in human-led research alone. As coding agents become increasingly embedded within scientific practice, the key question is therefore unlikely to be whether they think identically to humans, but whether the forms of diversity they produce remain transparent, auditable, and epistemically productive.

3 Methods

3.1 Operational details supplied beyond the paper

To enable reproduction from the task materials alone and to standardise data construction across the three conditions (Methods Only, Methods + Results, Transparent), the instruction bundle supplied three operational details that are absent from the paper and its online supplement (Tables S1–S10 and Figure S1). First, the bundle provided the ISSP wave-specific column names corresponding to each analysis variable together with the numeric-code-to-category mappings required to construct them: the six dichotomous welfare-attitude outcomes were derived from ISSP 1996 columns `v36`, `v41`, `v42`, `v39`, `v44` and `v38` and their 2006 counterparts `V25`, `V30`, `V31`, `V28`, `V33` and `V27`; labour-market status from `v206` (1996) and `wrkst` (2006), with codes 2–4 mapped to part-time, 5 to unemployed and 6–10 to not-in-labour-force; self-employment from `v213 == 1` in 1996 and `wrktype == 4` in 2006; and education from the ISSP 1996 `v205` (seven categories) and 2006 `DEGREE` (six categories), collapsed into less-than-secondary, secondary, and university-or-above. Second, because the 1996 wave encodes country with wave-specific integers that differ from the ISO-3166 numeric codes used in the 2006 wave and in the country-year macro file, the bundle provided the exact recode aligning the 1996 codes to ISO-3166 prior to appending. Third, the bundle specified that the country-level immigration variables `foreignpct` and `netmigpct` are lagged one year (1995 values for the 1996 wave and 2005 values for the 2006 wave), with the lags pre-applied in the supplied country-year macro file.

3.2 Evaluation Metrics

Our analyses operate at three complementary levels—what the agents *search*, what they *conclude*, and what *drives the variation across runs*—and we report metrics in three corresponding families, plus a fourth family that quantifies bit-exact reproduction of the original study and a fifth that tests robustness to prompt wording.

Method diversity. We characterise the analytic space each agent explores at two stages of the pipeline. *Coverage* is measured by the number of models entered in a team's submission table (model volume), the per-unit modelling rate (models per country–wave), and the length of the planning document and the final analysis script. *Method mix* is captured by binary indicators across five themes—estimator (e.g., two-way fixed effects, multilevel, binary or ordered logit, Bayesian), inference (cluster-robust SE, wild-cluster bootstrap), dependent-variable construction (composite index, binary recoding), sample (Brady–Finnigan's 13-country panel, inclusion of Eastern Europe, all available countries), and robustness checks (leave-one-country and leave-one-wave jackknives, alternative immigration sources)—each summarised as the proportion of runs in which the choice appears.

Analysis-layer agreement. At the level of estimates and conclusions we report (i) per-cell average marginal effects (AMEs); (ii) the two-sample Kolmogorov–Smirnov distance D between the agent and the 20-team human AME distributions for each of the seven dependent variables; (iii) the standardised mean difference d_z between agent and human AMEs; and (iv) the proportion of model runs returning a *negative*, *null*, or *positive* conclusion under the paper's verdict rule ($p \leq 0.05$), with 95% confidence intervals obtained by run-level bootstrap.

Variation explainability. To separate substantive heterogeneity from incidental noise, we fit a random-intercept mixed-effects model to each AME distribution and report the share of total variance explained between teams, within teams, and overall, alongside the share of variance in the discrete conclusion (negative / null / positive) attributable to team identity.

Reproduction accuracy. For the reproducibility experiment we score each of the 72 country-level coefficients in Tables 4–5 of Brady and Finnigan (2014) against the agent's output along four per-cell criteria: exact match of the significance marker; exact match of the odds ratio to three decimal places; exact match of the z-score to three decimal places; and the conjunction of all three. We also report a relaxed criterion—correct significance marker combined with the correct sign of the effect—as a measure of qualitative agreement.

Reproduction accuracy is the proportion of cells satisfying each criterion, averaged across $n = 5$ independent runs per condition and reported with bootstrap 95% confidence intervals.

3.3 Experimental Setup

We used the *Claude Code* agent built on Claude Opus 4.7 (1 M-token context, *Extra-High Effort* mode) and the *Codex* agent built on GPT-5.5 (*Extra-High Intelligence* mode), each operated in its sandboxed CLI mode. Both agents were confined to a dedicated working directory containing the task materials and a prompt-instructions file; they had no access to other locations on the host machine and no network access except where an experiment explicitly required it (see below). The study comprised two experiments—*Expansion* and *Reproduction*—that share this sandbox design but differ in inputs, prompts, and number of runs.

Expansion experiment. The agent is asked to design and execute its own test of whether immigration reduces public support for social policy. The working directory exposes the ISSP *Role of Government* waves I–V, the country–year macro panel assembled by Breznau et al. (2022), Brady and Finnigan's [20] country file, and a data dictionary, but no analysis code. We ran $n = 20$ independent runs per agent under three prompt variants—*Default* (no prefix), *Bias* (the agent is told it believes immigration strongly undermines social-policy support), and *Confirmatory* (the agent is instructed to audit the Brady–Finnigan null result). Each run produces a pre-analysis plan (`research_design.md`), an analysis script (`replication_code.py` or `.R`), a written conclusion (`conclusion.md`), and a model-level marginal-effects table (`results/marginal_effects.csv`).

Reproduction experiment. The agent is asked to reproduce the 72 country-level coefficients in Tables 4 and 5 of Brady and Finnigan (2014) under five conditions of increasing information availability: *No Model* (research question only), *Model* (methods text), *Model + Results* (methods plus the published table), *Model + Results + Code* (methods, results, and the authors' archived Stata do-file), and *Full Access* (the complete reproducibility package, including the full PDF, the CRI repository, and unrestricted web access). The working directory for each condition exposes *only* the materials defined by that condition. We ran $n = 10$ independent runs per agent per condition.

Sandbox restrictions. Within the sandbox, both agents were permitted to execute shell commands and install Python or R dependencies, but were otherwise restricted to the materials in the working directory. Web search and external file retrieval were enabled in the **Expansion** experiment, where agents were free to consult online documentation, statistical references, and supplementary data sources while designing and executing their own analyses. In the **Reproduction** experiment, by contrast, web search, external file retrieval, and system-wide file access were disabled through the agent configuration files (`.claude/settings.json` and `.codex/settings.json`; see Section B), which whitelisted only the commands required to run the analyses locally—ensuring that each information condition (*No Model* through *Model + Results + Code*) exposed the agent to exactly, and only, the materials it was meant to receive. The single exception within the reproduction experiment is the *Full Access* condition, in which network access and external retrieval were re-enabled so that the agent could consult the authors' code, the journal version of the paper, and any additional materials linked from it.

3.4 Recovering per-decision codes for the AI agents.

Following Breznau et al. [6], we treat each catalogued analytic choice as a single binary *decision*. Their SI Table S12 enumerates 193 such variables; ref. [6] restrict their statistical analyses to the 107 taken by at least three teams (the remaining 59 are unique to one or two teams and would have impeded identification). The paper text refers to "166 decisions," i.e., the 107 used in regressions plus the 59 unique-to-1-or-2-teams; the additional 27 rows in S12 are administrative identifiers (`u_teamid`, `count`, `AME`, `p`, etc.) and a handful of country flags the original PIs left uncoded for lack of cases (`germany_west`, `germany_east`, `n_ireland`). We retain all 193 to keep the extended table comparable to S12. For each agent we recover a per-model code for each of the 193 decisions by applying identical regular-expression patterns to the run's `marginal_effects.csv` (for the estimator, dependent variable, and immigration-measure columns), `replication_code.py` (for variable lists, country sets, individual-level and macro-level controls, and interactions), and `conclusion.md` (for the hypothesis verdict). For humans the proportions are read directly from `cri.csv`. Country flags are extracted only from the executable code (not from the natural-language design document) so that countries discussed in "excluded" lists are not falsely counted as present in the model sample.

3.5 Anonymization of Replication Materials:

Three research assistants manually screened and anonymized all replication materials to remove identifying information about the original studies, including paper titles, author names, and explicit references to research questions. Identifiers embedded in scripts, bibliographic files, directory structures, and related metadata

were systematically edited or removed. The goal was to ensure that agent performance reflected the ability to interpret and execute reproduction materials rather than reliance on memorized training data. As a final verification step, we provided the original paper PDFs to Claude Code (Opus 4.7) and instructed the agent to scan the anonymized directories for residual identifiers. This process surfaced additional cases, including author names embedded in file paths, links to personal repositories, and identifiers in filenames. These remaining instances were manually removed, and associated script references were updated to preserve execution consistency.

4 Related Work

4.1 Reproducibility Crisis

Across scientific disciplines, computational results frequently fail to reproduce even when original data and code are available, with failure rates exceeding 50% in some fields [29, 30], a phenomenon called as reproducibility crisis [31].

4.2 Reproducibility and Replication Benchmarks

CORE-Bench [32] is one of the first benchmarks to treat computational reproducibility as an end to end agent task. It builds 270 tasks from 90 papers across computer science, social science, and medicine, and varies task difficulty by changing how much execution support the agent receives, ranging from full access to outputs to having only a README and needing to install dependencies and run the pipeline. It also includes both text and vision questions, requiring agents to interpret plots, tables, and PDFs in addition to terminal outputs. A key contribution is its evaluation harness, which runs each task in an isolated virtual machine and supports large scale parallel evaluation, reducing runtime from weeks to hours. A major limitation is that CORE-Bench is built from CodeOcean capsules, which introduces a clear selection bias toward already reproducible projects. Another limitation is that it includes only 28 social science papers, limiting its coverage of this domain. HAL [33] addresses large scale agent evaluation by providing shared infrastructure for orchestrating VMs, tracking costs, and inspecting logs for unsafe behavior. Its main limitation is that it is infrastructure rather than a benchmark, so its usefulness depends on the quality of the underlying tasks, and some measures, such as latency, are difficult to interpret at scale.

REPRO-BENCH [34], focuses only on social science, shifts the goal from simply running code to judging whether a social science paper's major findings are actually reproduced and then assigning a reproducibility score on a 1 to 4 scale. Each task includes the full paper PDF, the reproduction package, and a list of major findings, which better matches how real reproduction audits are done. It also intentionally includes papers with both strong and weak reproducibility, and spans multiple languages and data formats, making the setting more realistic for social science. The companion agent work shows that performance is still low and that reliability remains a major challenge. ReplicatorBench [35] pushes beyond reproduction into replication by evaluating three stages that mirror human workflows, including extracting information from the paper, retrieving new data resources, and interpreting whether the claim meets preregistered criteria, with fine grained checkpoints for partial credit. Its main limitations are scale and scope, with only 19 studies due to the scarcity of expert documented replications, and reliance on LLM based judging for some open ended grading, which the authors treat as approximate.

4.3 LLM and Agent Performance on Reproducibility Tasks

Across CORE-Bench, Repro-Bench, HAL, and ReplicatorBench, existing evidence suggests that large language models and agent systems still struggle with computational reproducibility tasks. CORE-Bench shows that performance drops sharply when models must install dependencies, manage environments, and debug errors. Repro-Bench similarly reports low and unstable performance, especially for complex workflows or poorly documented projects. ReplicatorBench finds that models perform reasonably on information extraction but much worse on stages requiring reasoning about evidence and methods. HAL highlights frequent failures and inconsistent behavior at scale. None of these studies systematically evaluate coding-specific CLI agents such as Claude Code and Codex that autonomously navigate codebases and manage full replication pipelines. As a result, current evidence mainly reflects the limits of general purpose LLM-based agents, leaving the capabilities of specialized coding agents largely unexplored.

Acknowledgments

FG and MA conceived the study. MA led the implementation and wrote the first draft. All authors revised the manuscript. David Rand, Gordon Pennycook, and Adam Mahdi provided valuable input that informed this work. We thank seminar participants at the Reasoning with Machines Lab at the University of Oxford

for helpful discussions. We also thank Soheil Hooshmand, Saba Yousefzadeh, Sara Yari Mehmandoust, and Mohammadmasiha Zahedivafa for outstanding research assistance.

5 Data and Code Availability

Replication materials are available at https://github.com/malizad/SocSci-Repro-Bench and https://dataverse.harvard.edu/dataverse/alizadeh.

6 Conflict of Interests

The authors declare no conflict of interest.

References

[1] Helen E. Longino. *The Fate of Knowledge*. Princeton University Press, Princeton, NJ, 2002.

[2] Sandra D Mitchell. *Biological complexity and integrative pluralism*. Cambridge University Press, 2003.

[3] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

[4] Berna Devezer, Luis G Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one*, 14(5):e0216125, 2019.

[5] Lu Liu, Benjamin F Jones, Brian Uzzi, and Dashun Wang. Data, measurement and empirical methods in the science of science. *Nature human behaviour*, 7(7):1046–1058, 2023.

[6] Nate Breznau, Eike Mark Rinke, Alexander Wuttke, Hung HV Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K Andersen, Daniel Auer, Flavio Azevedo, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119, 2022.

[7] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. In *Forty-second International Conference on Machine Learning*, 2025.

[8] Margaret Li, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman. Predicting vs. acting: A trade-off between world modeling & agent modeling. *arXiv preprint arXiv:2407.02446*, 2024.

[9] Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*, 2025.

[10] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025.

[11] Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J Zico Kolter, and Daphne Ippolito. Forcing diffuse distributions out of language models. In *First Conference on Language Modeling*, 2025.

[12] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in neural information processing systems*, 35:3663–3678, 2022.

[13] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Advances in Neural Information Processing Systems*, 38, 2026.

[14] Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. The shrinking landscape of linguistic diversity in the age of large language models. *arXiv preprint arXiv:2502.11266*, 2025.

[15] Aaron Bramson, Patrick Grim, Daniel J Singer, William J Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman. Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of science*, 84(1):115–159, 2017.

[16] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[17] Kenneth Benoit, Kevin Munger, and Arthur Spirling. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508, 2019.

[18] William R Shadish, Thomas D Cook, and Donald T Campbell. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company, 2002.

[19] Imam Kusmaryono, Dyana Wijayanti, and Hevy Risqi Maharani. Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: A literature review. *International Journal of Educational Methodology*, 8(4):625–637, 2022.

[20] David Brady and Ryan Finnigan. Does immigration undermine public support for social policy? *American sociological review*, 79(1):17–42, 2014.

[21] George J Borjas and Nate Breznau. Ideological bias in the production of research findings. *Science Advances*, 12(1):eadz7173, 2026.

[22] Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144, 2024.

[23] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434, 2023.

[24] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022.

[25] Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

[26] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

[27] Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348(1-17):3, 2013.

[28] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[29] Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, 2018.

[30] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[31] Zacharias Maniadis and Fabio Tufano. The research reproducibility crisis and economics of science, 2017.

[32] Zachary S Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2024.

[33] Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, et al. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. *arXiv preprint arXiv:2510.11977*, 2025.

[34] Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhwani, Austin Peters, and Daniel Kang. Repro-bench: Can agentic ai systems assess the reproducibility of social science research? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23616–23626, 2025.

[35] Bang Nguyen, Dominik Soós, Qian Ma, Rochana R Obadage, Zack Ranjan, Sai Koneru, Timothy M Errington, Shakhlo Nematova, Sarah Rajtmajer, Jian Wu, et al. Replicatorbench: Benchmarking llm agents for replicability in social and behavioral sciences. *arXiv preprint arXiv:2602.11354*, 2026.

A Prompts

A.1 Expansion Prompt

AI Coding Agent Protocol — Study Context and Hypothesis

You are a quantitative social scientist participating in a many-analysts study. Multiple independent teams use the same data to test the same hypothesis, allowing organizers to observe how analytical choices lead to different results.
All participants completed **Phase I (Replication)**, in which they replicated a published study testing the same hypothesis on a subset of the data. Participants are familiar with the six dependent variables and with two-way fixed-effects logit estimation. The reference study found **no general effect**, provisionally providing evidence against the hypothesis.
You are now asked to complete:

- **Phase II — Research Design**
- **Phase III — Main Analysis / Expansion**

Hypothesis
Short form: *Greater immigration reduces support for social policies among the public.*
Long form: *A greater stock, or a greater increase in the stock, of foreign persons leads the public to become less supportive of social policy.*
Definition of Social Policy
Policies providing:

- social insurance
- welfare support
- income replacement
- active labor-market programs

Your goal is to test whether the null finding from earlier work is robust and generalizable.

AI Coding Agent Protocol — Data Sources

Data Directory
All data are located in:
`/adress/to/data/folder/`

Individual-Level Data
International Social Survey Programme (ISSP) — Role of Government. You may use any or all waves.

Wave	Year	Files
I	1985	ZA1490
II	1990	ZA1950
III	1996	ZA2900
IV	2006	ZA4700
V	2016	ZA6900

Country-Level Data
Located in:
`data/macro/`
Key variables include:

- Immigrant stock
- Immigrant flow
- GDP
- Inequality
- Social spending
- Labor market indicators
- Population
- Ethnic fractionalization

Missing values appear as dots.

AI Coding Agent Protocol — Dependent Variables and Constraints

Dependent Variables
Government responsibility for:

1. Jobs
2. Health care
3. Old-age support
4. Unemployment support
5. Income redistribution
6. Housing support

All six must be included.
They may be analyzed:

- Separately
- As an index
- As a latent scale

Design Constraints
Your design must:

1. Use ISSP data
2. Include all six dependent variables
3. Focus on advanced welfare-state democracies
4. Justify country selection
5. Justify additional variables if added

AI Coding Agent Protocol — Phase II: Research Design

Phase II — Research Design
Write a pre-analysis plan describing your ideal test.
Maximum:
750 words (excluding tables and figures).
Your design must specify:

- Target population
- Country selection
- ISSP waves
- Dependent-variable construction
- Immigration measures
- Individual controls
- Country controls
- Modeling strategy
- Functional form
- Sample size considerations
- Sensitivity analyses

Important Rule
Do NOT run analyses during Phase II.
Save as:
`research_design.md`

AI Coding Agent Protocol — Phase III: Main Analysis

Phase III — Main Analysis
Execute your design exactly as written.
Allowed:
- Minor implementation changes
- Documented deviations

Required outputs:
- Regression tables
- Marginal effects
- Confidence intervals
- Plots

Required Output 1 — Marginal Effects
Compute:
1. Effect of 1% increase in immigrant stock
2. Effect of 1 additional migrant per 1,000 population

Report:
- 95% confidence intervals
- Standard-deviation units (if possible)

AI Coding Agent Protocol — Deliverables and Logging

Required Files
Create:
1. research_design.md
2. replication_code.<ext>
3. results/marginal_effects.csv
4. results/regression_tables.md
5. results/plots/
6. conclusion.md
7. analysis_log.txt

Substantive Conclusion
Choose exactly one:
(a) Support
(b) Lack of support
(c) Not testable

Provide justification.

Analysis Log Must Include
- Software versions
- Data steps
- Row counts
- Implementation decisions
- Errors and convergence issues

AI Coding Agent Protocol — Execution Rules

Rules

1. Use R, Python, or Stata
2. Script must run end-to-end
3. Document decisions
4. Do not tune results
5. Do not run analyses during Phase II
6. Report failed models
7. Document infeasible tests
8. Do not consult prior published results
9. Do not modify source data directories

Output Directory

`/address/to/output/directory/`

B Permission Settings

B.1 Claude Code

Project-Level Configuration for Claude Code

This guide describes how to configure a `settings.json` file for a **single Claude Code project** that:

- Allows common development operations (editing files, running scripts, creating directories) without manual approval.
- Blocks all web access (including WebSearch, WebFetch, `curl`, and `wget`).

```
cd /path/to/your/project

mkdir -p .claude
```

Open the file in a text editor:

```
nano .claude/settings.json

cat .claude/settings.json
```

Place the following content in `.claude/settings.json`:

```
{
  "permissions": {
    "defaultMode": "acceptEdits",
    "allow": [
      "Bash(*)",
      "Write(*)",
      "Edit(*)",
      "MultiEdit(*)",
      "Read(*)"
    ],
    "deny": [
      "WebSearch",
      "WebFetch",
      "Bash(curl:*)",
      "Bash(wget:*)",
      "Bash(fetch:*)",
      "Read(~/.ssh/**)",
      "Read(~/.aws/**)",
      "Read(~/.env)",
      "Read(~/.gnupg/**)",
      "Edit(~/.bashrc)",
      "Edit(~/.zshrc)"
    ]
  },
  "sandbox": {
    "enabled": true,
    "autoAllowBashIfSandboxed": true
  }
}
```

B.2 Codex

Codex Sandbox Configuration (config.toml)

```
################################################################################
# Codex sandboxed reproducibility profile
# - Confines execution to the workspace (current directory + subdirs)
# - Disables Codex web search
# - Allows network only for package installation (pip / CRAN)
################################################################################

sandbox_mode = "workspace-write"
approval_policy = "untrusted"
web_search = "disabled"

[sandbox_workspace_write]
network_access = true
exclude_slash_tmp = true
exclude_tmpdir_env_var = true
```

C Extended Results

C.1 Outcome-level distributional fidelity

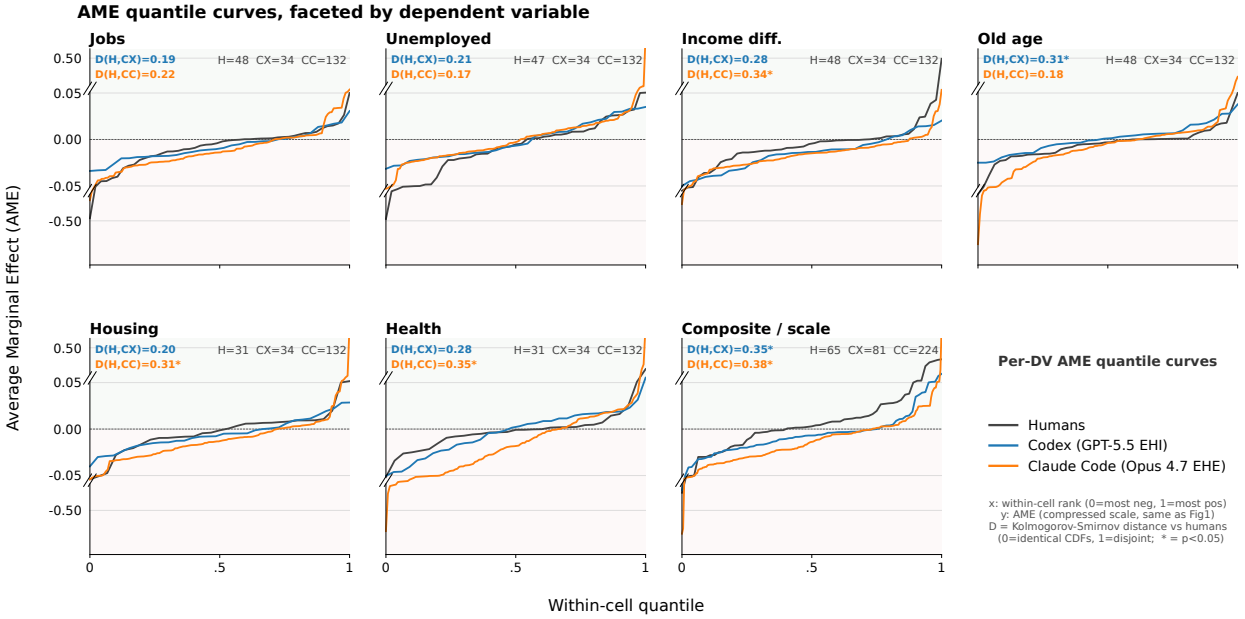

Figure 7: **Per-outcome AME distributions of AI agents track the human closely on most dependent variables but diverge on the composite scale.** Within-cell quantile curves of the average marginal effect (AME) of immigration on each policy outcome, comparing 73 human research teams (grey), 20 Codex (GPT-5.5 EHI) runs (blue), and 20 Claude Code (Opus 4.7 EHE) runs (orange). For every (DV, IV) cell, point estimates are ranked from the most negative ($x = 0$) to the most positive ($x = 1$); the curve traces the empirical CDF of AMEs that share each rank position. Sample sizes per panel are printed in the upper-right (H, humans; CX, Codex; CC, Claude Code). The y-axis uses the compressed scale of Fig. 1, with breaks at $|\text{AME}| = 0.05$ and 0.50 to make small effects visible without truncating tails. D values in the upper-left of each panel are two-sample Kolmogorov–Smirnov distances between the agent and the human distribution (0 = identical CDFs, 1 = disjoint; asterisks mark $P < 0.05$). The two agents are statistically indistinguishable from humans on jobs, unemployment, and income difference, but Claude Code's distribution is significantly compressed on housing, health, and the composite scale ($D(\text{H}, \text{CC}) = 0.21^*$, 0.35^*, 0.38^*), and Codex differs on old age and the composite ($D(\text{H}, \text{CX}) = 0.31^*$, 0.35^*).

C.2 Reproduction of Brady & Finnigan (2014)

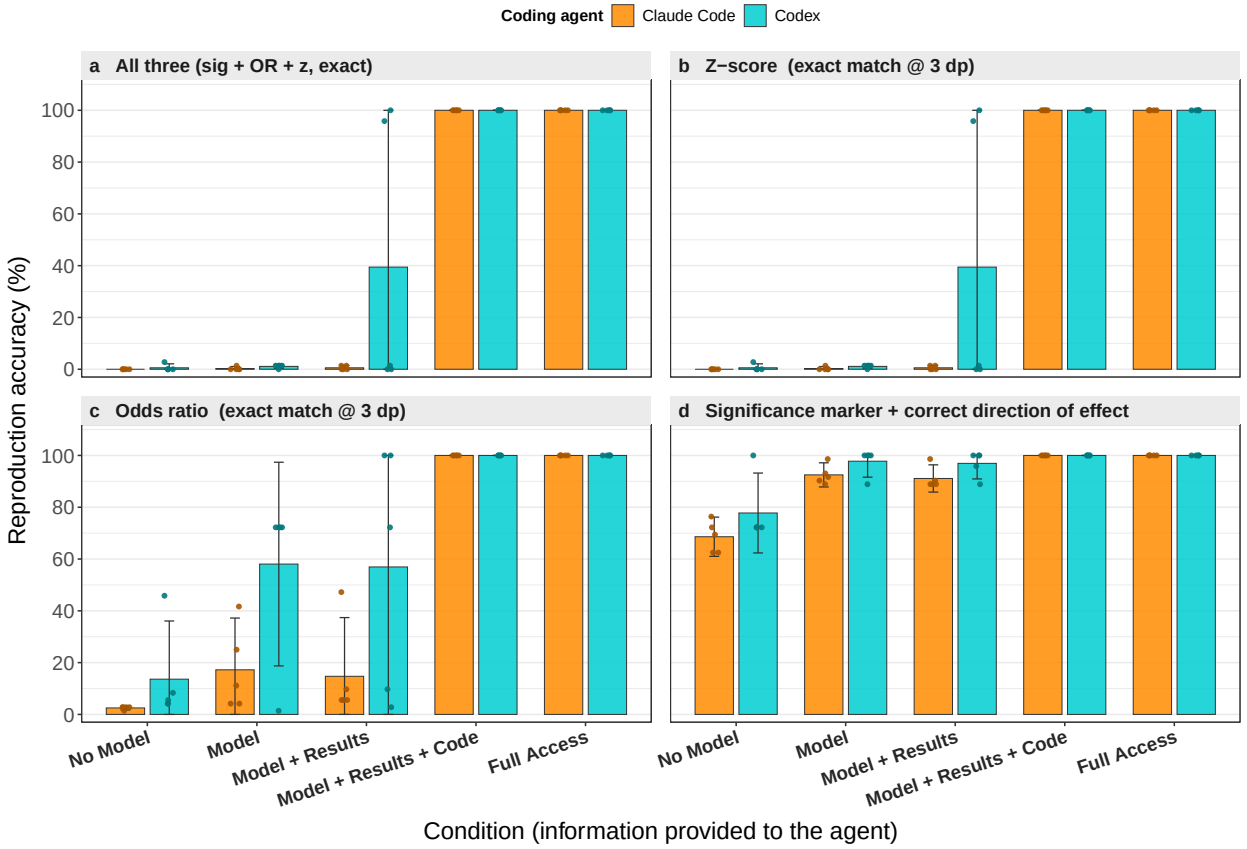

Figure 8: **Reproduction accuracy of two AI coding agents as a function of how much information from the original study is provided.** Each panel shows the per-cell reproduction accuracy (%) of the Claude Code (orange) and OpenAI Codex (cyan) coding agents when asked to reproduce Table 1 of Brady & Finnigan (2013) under five information conditions of increasing transparency: **No Model** (task description only), **Model** (statistical model/methods text), **Model + Results** (methods plus the published results table), **Model + Results + Code** (methods, results, and the original analysis code), and **Full Access** (the complete reproducibility package, including data). (**A**) Joint exact reproduction of all three reported quantities per cell—significance marker, odds ratio, and z-score (the last two matched to three decimal places). (**B**) Exact match of the z-score (3 d.p.). (**C**) Exact match of the odds ratio (3 d.p.). (**D**) Correct significance marker combined with the correct sign (direction) of the effect. Bars give the mean across $n = 5$ independent runs per agent × condition; dots show individual runs; error bars are 95% confidence intervals (clipped at 0 and 100%). Exact numerical reproduction (A–C) is essentially unattainable without access to the analysis code, jumping from $\leq$1% (Claude Code) or $\leq$40% (Codex, with very wide between-run variance) to $\sim$100% once code is supplied. In contrast, the qualitative pattern of significance and direction of effect (D) is recovered in $\sim$70–80% of cells from the methods text alone and exceeds 90% once any model description is provided. Differences between the two agents are within run-to-run variability under every condition.

C.3 Model Specification Coding and Distribution

Table 1: Decision-frequency table extending Breznau et al. (2022) SI Table S12 to AI agents. Variable definitions are taken verbatim from S12 (with minor copy-edits for length). "Humans" are the 20 randomly sampled teams (seed = 42) used in Fig. 1*A*; "CX" are 20 Codex runs; "CC" are 20 Claude Code runs. Cells show the percentage of models in which the variable is coded as 1 / non-zero / present. Agent flags are derived by parsing each run's `marginal_effects.csv`, `replication_code.py`, and `conclusion.md` with pattern matchers tuned to the variable definitions; they should be read as automated approximations rather than exact recodes. Rows sorted by humans % (desc.).

Variable	Definition	**Humans** (%)	**CX** (%)	**CC** (%)
u_teamid	Random team number assignment except team 0, which refers to the Brady and Finnigan study. These specifications are excluded from the analysis but left in here for comparison.	100	100	100
main_IV_type	Test variable type for the hypothesis that immigration undermines social policy support: "stock" (% foreign-born), "flow" (change in %, net migration or change in stock), or "change in flow".	100	100	100
count	A counter to return results to their original order.	100	100	100
num_countries	Number of countries in the model sample.	100	100	100
inv_weight	The number of models per team, must be divided into 1 to use for weighting.	100	100	100
main_IV_effect	Total, within, or between effect. For non-multilevel models, always total. A within-effect of stock is "Flow per wave".	100	100	100
main_IV_time	The time period the team used to measure flow of immigrants (1-year, 5-year, etc.). PIs rescaled to a 1-year equivalent for comparability; this refers to the original metric.	100	100	100
main_IV_measurement	Measuring what type of immigrants. "Emigration" is coded as "Immigrant, foreign-born".	100	100	100
main_IV_source_file	Name of the source file used.	100	100	100
main_IV_source	The data source; many teams imputed some countries using other sources, coded only as the primary source. (Deprecated.)	100	100	100
package	Software package, character categories.	100	100	100
DV	Dependent variable used; single questions labeled "Jobs" etc.; scale variables start with "Scale_" followed by the number of items.	100	100	100
z	Z-statistic or equivalent (T-value).	100	100	100
error	The absolute deviation of the high 95% CI from the margin.	100	100	100
upper	Upper confidence boundary at 95% CI.	100	100	100
lower	Lower confidence boundary at 95% CI.	100	100	100
AME	Average marginal effect as produced by team's provided code; or added by PIs to produce when not present.	100	100	100
p	p-value or equivalent confidence interval relative to zero (e.g. for Bayes estimation).	100	100	100
id	Team number plus model number counted in order within teams.	100	100	100
sex_iv	Sex / gender of respondent.	97.49	100	100
age_iv	Age as a continuous variable.	97.49	100	100
switzerland	Country included in sample.	96.24	95.82	61.22
france	Country included in sample.	96.24	95.82	61.22
norway	Country included in sample.	96.24	95.82	61.22
spain	Country included in sample.	96.24	95.82	61.22
sweden	Country included in sample.	96.24	95.82	61.22
w2006	Includes data from ISSP 2006 wave.	94.98	100	99.82
australia	Country included in sample.	93.73	95.82	91.24
usa	Country included in sample.	92.48	100	91.24
listwise	Listwise deletion: cases are dropped if any relevant variable is missing for that observation.	91.22	100	100
germany	Country included in sample.	90.60	95.82	61.22
employed_iv	Employed, or a categorical variable with self / public / full / part etc.	89.97	61.84	100
education_iv	Any measure of educational attainment or years (rough; finer-grained coding could be considered).	89.97	100	100
new_zealand	Country included in sample.	89.34	95.82	61.22
w1996	Includes data from ISSP 1996 wave.	88.71	100	98.72
great_britain	Country included in sample.	84.95	95.82	61.41
age2_iv	Age-squared, or a categorical break-down (a non-linear age function).	84.33	100	94.53
japan	Country included in sample.	79.94	95.82	61.22
w2016	Includes data from ISSP 2016 wave.	78.37	91.09	80.11
canada	Country included in sample.	68.34	95.82	61.22

(continued on next page)

(continued from previous page)

Variable	Definition	Humans (%)	CX (%)	CC (%)
dichotomize	Dependent variable is dichotomized.	63.95	0	1.09
stata	Stata software employed (dummy for package).	63.64	0	0
logit	Logistic regression; fits "S"-shaped logistic curve to a 0/1 DV. Includes multilevel logistic.	62.07	0	17.70
ireland	Country included in sample.	58.93	95.82	61.22
finland	Country included in sample.	57.68	77.16	61.22
Stock	Dichotomous indicator for main_IV_type.	56.74	53.20	56.75
income_iv	Income.	53.29	91.64	88.78
mlm_any	Any multilevel model: =1 if mlm_re, mlm_fe, and/or hybrid_mlm =1.	52.04	0	7.66
unbalpanel	Unbalanced time-series; includes different numbers of countries per wave.	50.47	100	100
twowayfe	Two-way fixed-effects (2WFE). Contains dummy variables for country and year regardless of estimation strategy. The PIs follow the Brady-Finnigan nomenclature.	48.59	36.21	66.61
mlm_re	Random-effects multilevel model: random intercepts and fixed coefficients (an "RE model" in econometrics).	46.39	0	3.65
Hmixed	Two separate, internally consistent conclusions about stock and flow leading to mixed-result claims.	46.39	29.53	48.63
level_country	Unspecified modelling of country level, can include random-effects or dummies.	45.14	16.16	0
denmark	Country included in sample.	44.83	77.16	61.22
Flow	Dichotomous indicator for main_IV_type.	41.38	46.80	43.25
emplrate_ivC	Employment rate (usually of those in the labor force).	40.13	91.36	35.86
Hreject_stock	Hypothesis rejected specifically for stock (see above).	36.99	0	45.80
socx_ivC	Social Expenditures % of GDP ("SOCX").	34.80	100	85.95
Unemp	Single question on government provision of unemployment protection is the DV, or part of the scale if Scale=1.	32.92	9.47	12.96
OldAge	Single question on government provision of old-age care is the DV, or part of the scale if Scale=1.	32.92	9.47	12.77
IncDiff	Single question on government reduction of income differences is the DV, or part of the scale if Scale=1.	32.92	9.47	12.77
main_IV_as_control	If the other main IV is in the same model: 0=no, 1=yes. Within/between models =1 only if both stock and flow are entered as separate variables.	31.97	0	0
portugal	Country included in sample.	31.97	46.80	50.00
Hsupport_net	Hypothesis supported specifically for the flow / net-migration test variable.	31.35	65.18	27.92
Jobs	Single question on government provision of jobs is the DV, or part of the scale if Scale=1.	30.41	9.47	12.77
netherlands	Country included in sample.	28.21	77.16	61.22
r	R software employed (dummy for package).	27.59	0	17.06
eeurope	Includes at least 3 Eastern European countries.	27.59	3.90	8.94
hungary	Country included in sample.	27.59	0	0
latvia	Country included in sample.	27.59	0	0
Hreject_net	Hypothesis rejected specifically for flow / net-migration.	26.33	0	11.13
slovenia	Country included in sample.	26.33	0	0
Hsupport_stock	Hypothesis supported specifically for the stock test variable (only listed when researchers report stock/flow conclusions separately).	26.33	95.82	43.80
House	Single question on government provision of housing is the DV, or part of the scale if Scale=1.	25.71	9.47	12.77
Health	Single question on government provision of health care is the DV, or part of the scale if Scale=1.	25.71	9.47	12.77
Hreject	Researchers conclude the hypothesis is rejected; inconclusive support is also counted as rejection.	25.08	19.50	15.78
level_cyear	Unspecified modelling of country-year level, can include random-effects or dummy variables in a multilevel model.	23.82	0	0.18
mmodel	Measurement model: uses scaling, factor analysis or item-response to test/generate a latent DV. Always with a linear estimator.	23.82	0	0
czechia	Country included in sample.	23.82	0	5.47
poland	Country included in sample.	22.57	0	0
ml_glm	Maximum likelihood: ML or any other iterative version that is not OLS, Bayes or Logit (e.g., GLM, MWFE).	22.57	0	1.09
Scale	A multi-item scale was constructed and used as the DV; the questions used are indicated by the previous 6 variables.	20.38	43.18	23.18
allavailable	>21 countries; all available or mostly all.	20.06	3.90	8.94
Hsupport	Researchers conclude immigration undermines social-policy preferences and the team's evidence supports it (subjective; team prerogative).	19.44	66.57	60.86
croatia	Country included in sample.	18.81	0	0

(continued on next page)

(continued from previous page)

Variable	Definition	Humans (%)	CX (%)	CC (%)
israel	Country included in sample.	18.81	24.51	14.69
korea	Country included in sample.	18.81	3.90	9.12
year_dummies_only	If not 2WFE: includes a year dummy for each year (also includes dummies within an MLM but not RE intercepts).	15.67	0	1.28
orig13	Identical to the original 13 countries used in Brady & Finnigan's two-way fixed-effects models ("13 richest democracies").	15.67	13.09	7.03
russia	Country included in sample.	15.05	0	0
leftright_iv	Left-right subjective political ideology, or actual reported party vote coded into left/right.	13.79	0	18.16
level_year	Unspecified modelling of year level, can include random-effects or dummies. Refers technically to survey wave.	12.54	17.55	2.55
socialistdummy_ivC	Former state-socialist societies = 1, others = 0.	11.29	51.81	50.36
italy	Country included in sample.	10.66	36.77	57.76
w1990	Includes data from ISSP 1990 wave.	10.66	10.86	30.75
fract_ivC	Ethnic fractionalization / Herfindahl index (e.g., from UN stock-by-origin data, Alesina).	10.03	41.50	40.88
anynonlin	Any nonlinearity used; =1 if any of the above interactions =1, plus a few cases with interactions not in the list (e.g., team-98 immigration x party voting; one squared-DV in team 29).	9.40	3.90	2.37
Hnotest	Researchers conclude the hypothesis is not testable, or the evidence is inconclusive to support or reject.	9.09	4.46	3.47
mplus	Mplus software employed (package dummy).	8.78	0	0
unemprate_ivC	Unemployment rate of those in the labor force (usually means registered unemployed).	8.46	0	0
mcp_ivC	Multiculturalism Policy Index, MIPEX, or IMPIC immigration policies index.	8.15	66.57	39.60
ols	Ordinary least squares estimator.	7.84	84.40	59.85
bulgaria	Country included in sample.	7.52	0	0
chile	Country included in sample.	7.52	0	0
hybrid_mlm	Includes both random-effects and fixed-effects components.	7.52	0	4.11
gdp_ivC	GDP per capita.	7.52	100	100
south_africa	Country included in sample.	7.52	0	0
cyprus	Country included in sample.	7.52	0	0
household_iv	Household composition (unspecified).	6.90	0.56	17.52
mlm_fe	Fixed-effects multilevel model: random intercepts so country-level variables are mean-centered within country; explains within-country changes only.	5.64	0	4.01
Hnotest_net	Hypothesis not testable specifically for flow / net-migration.	5.64	0	0
fbXleftright	Interaction (indicated by "X").	5.02	0	0
bayes	Bayesian estimator (MCMC etc.) fitting posterior probabilities based on prior distributions for more 'consistent' level-2 estimates.	5.02	0	0
w1985	Includes data from ISSP 1985 wave.	5.02	3.90	8.76
cluster_any	Any kind of clustering command added by the researcher (excludes a multilevel model's implicit clustering).	4.70	85.79	35.95
slovakia	Country included in sample.	4.39	0	0
belgium	Country included in sample.	4.39	68.80	57.76
orig17	Identical to the 17 countries used by Brady & Finnigan in their MLM random-effects models.	4.39	13.09	16.24
weights	Any survey weights applied.	4.08	3.90	4.56
mlogit	Multinomial logistic estimator. Includes multilevel ordered logit or probits.	3.76	0	0
netXinc	Interaction (indicated by "X").	3.76	0	0
categorical	Dependent variable has more than 2 categories.	3.76	56.82	76.09
iceland	Country included in sample.	3.76	39.28	39.05
L2boots	Robust SE or bootstrapped level-2 analysis (jackknife, sandwich robust, or fe-robust in Stata's xtreg).	3.76	0	2.55
married_iv	Marital status.	3.45	0	0
decomm_ivC	Some measure of replacement rates (Scruggs / CWED).	3.13	0	0
conservatism_ivC	Conservative (left-vs-right) government political-ideology index (e.g., Schmidt index); includes vote-share measures.	3.13	0	5.47
ologit	Ordered logistic / probabilistic estimator (probit). Includes item-response, ordered-logit and probit models.	2.51	0	5.66
lpm	Linear probability model estimation. DV coded 0/1 but linear model used.	2.51	10.03	6.66
socult_ivC	Socio-cultural proximity scale using country of origin for immigrants.	2.51	0	0
pseudo_pnl	Constructed a pseudo-panel of individual-level groups.	1.88	0	0
taiwan	Country included in sample.	1.88	0	0
lithuania	Country included in sample.	1.88	0	0
ChangeFlow	Dichotomous indicator for main_IV_type.	1.88	2.23	4.74

(continued on next page)

(continued from previous page)

Variable	Definition	Humans (%)	CX (%)	CC (%)
india	Country included in sample.	1.88	0	0
turkey	Country included in sample.	1.88	0	0
austria	Country included in sample.	1.88	47.63	57.94
ginin_ivC	Gini (not enough cases of pre-tax Gini to differentiate; also includes one case of top-income concentration from WID).	1.88	0.56	78.92
multimpute	Pairwise information or imputation employed (e.g. FIML or multiple imputation).	1.25	0	0
year_as_count	Year added as a continuous variable; =1 if year is continuous and >2 waves are included.	1.25	0	0
fbXeduc	Interaction (indicated by "X").	0.31	0	2.37
netXeduc	Interaction (indicated by "X").	0.31	0	0
fbXnet	Interaction (indicated by "X") between foreign-born stock and net migration.	0	0	0
netXcons	Interaction: net migration x conservatism index.	0	0	0
netXage	Interaction (indicated by "X").	0	0	0
efficacy_iv	Political efficacy (believes he/she can influence government).	0	0	0
netXsex	Interaction (indicated by "X").	0	0	0
netXunemp	Interaction (indicated by "X").	0	0	0
fractXfb	Interaction (indicated by "X").	0	0	0
fbXage	Interaction (indicated by "X").	0	0	0
fbXsex	Interaction (indicated by "X").	0	0	0
squared_imm	A quadratic form for one or both immigration variables.	0	3.90	0
Hnotest_stock	Hypothesis not testable specifically for stock (see above).	0	0	0
fbXunemp	Interaction (indicated by "X").	0	0	0
fbXgini	Interaction (indicated by "X").	0	0	0
fbXurban	Interaction (indicated by "X").	0	0	0
fbXinc	Interaction (indicated by "X").	0	0	0
philippines	Country included in sample.	0	0	0
trust_iv	Political trust.	0	0	0
mlwin	MLwiN software (package dummy).	0	0	0
upol_iv	Subjective interest in politics.	0	0	0
taxes_iv	Subjective attitude that government should tax more / less.	0	0	0
uruguay	Country included in sample.	0	0	0
emigration_ivC	Gross or net out-migration ('flow').	0	0	0
unchange_ivC	Annual change in unemployment rate.	0	0	0
poverty_ivC	Poverty (e.g. 50% of median).	0	0	0
fbunemprate_ivC	Foreign-born unemployment rate.	0	0	0
fbunempchange_ivC	Change in foreign-born unemployment rate.	0	0	0
fbeducrate_ivC	Foreign-born education rate.	0	0	0
fbeducratechange_ivC	Change in foreign-born education rate.	0	0	0
socxchg_ivC	Change in SOCX.	0	0	0
gdpchange_ivC	Any change measure of GDP (1-yr / 5-yr, etc.).	0	0	0
regime_ivC	Categorical welfare-state or institutional-regime type, not including a post-communist split.	0	0	65.33
targeting_ivC	Benefits target groups (vs.\{} universal).	0	0	0
socx_programspecific_ivC	Social spending decomposed into single program domains.	0	0	0
subFB_ivC	Subjective foreign-born, country mean.	0	0	0
spss	SPSS software employed (dummy).	0	0	0
antiimm_ivC	Aggregate measures of anti-immigrant attitudes / sentiment from other surveys (e.g., ISSP National Identity, ESS).	0	0	5.66
pop_ivC	Population of country.	0	0	0
occclass_iv	Occupational class.	0	0	0
occstatus_iv	Occupational status.	0	0	0
country_dummies_only	If not 2WFE: includes a country dummy for each country (also includes dummies within an MLM but not RE intercepts).	0	0	0
venezuela	Country included in sample.	0	0	0
reldenom_iv	Religious denomination.	0	0	0
relattend_iv	Religious service attendance.	0	0	16.70
publice_iv	Employed in the public sector.	0	0	0
urban_iv	Urban / rural / suburban (unspecified).	0	0	0
fb_iv	Foreign-born respondent in the ISSP.	0	5.85	58.12
cuts_iv	Subjective attitude that government should make cuts.	0	0	10.40
tradeunion10_ivC	10-year change in trade-union share of employed.	0	0	0
germany_west	Distinguished (not coded, not enough cases).	—	0	0
germany_east	Distinguished (not coded, not enough cases).	—	0	0
n_ireland	Distinguished (not coded, not enough cases).	—	0	0