

AI and the Research Team^{*}

Johan Fourie[†]

Abstract

Artificial intelligence is associated with larger research teams, yet in mathematics, among the most codifiable fields, individual researchers working with AI now produce research-grade results. A span-of-control model reconciles these observations. AI lowers execution cost, which expands laboratory scale, and automates codifiable tasks, which lowers the member share of each unit. Team size is therefore quasi-concave in AI capability, with at most one peak. The model predicts that a fully codifiable team peaks when effective automation coverage reaches a closed-form threshold, typically near complete coverage, and, among fields with shared primitives that possess an interior peak, those with less irreducibly human task content peak first. Under explicit priors, the 90 percent forecast intervals for the fully codifiable peak span 2026 to 2030.

Keywords: research teams; artificial intelligence; automation; team size; forecasting

JEL codes: D23; J24; O31

^{*}Many conversations have informed this paper. I particularly appreciate thoughtful discussions with Rulof Burger, Helanya Fourie, Willem Fourie, Jesse Naidoo, Matthew Olckers, Melt van Schoor and Marisa von Fintel, and the many reader comments on ourlongwalk.com on this and related issues. This paper was created with the help of Anthropic’s Claude Code (Opus 4.8 and Fable 5) and OpenAI’s Codex (GPT-5.5 and GPT-5.6), and was checked for errors with refine.ink; the author checked every result and is responsible for all claims and errors. Proofs, Monte Carlo details, and additional evidence are in the supplemental appendix. Cite this paper as: Fourie, Johan. 2026. “AI and the Research Team.” Working Paper, Department of Economics, Stellenbosch University.

[†]Department of Economics, Stellenbosch University. Email: johanf@sun.ac.za.

1 Introduction

Research teams combine scientific judgment with execution. A principal investigator sets direction and evaluates results while other researchers run experiments, write code, collect data, and develop proofs. Artificial intelligence performs a growing set of the codifiable tasks in the second category, with two opposing effects: first, it lets a laboratory undertake more work and, second, it does work researchers previously performed. The evidence points in both directions. AI use in science is associated with larger teams (Hosseinioun, Uzzi and Fosse, 2026), and scientists who adopt AI publish more and reach project leadership earlier (Hao et al., 2026). In mathematics, among the fields where execution is most codifiable, a language-model system produced a formal Lean proof resolving Erdős problem #728 in January 2026, and a community-maintained ledger now records dozens of full AI resolutions of open problems (Sothanaphan, 2026; Erdős Problems Wiki contributors, 2026). A single researcher directing AI can complete work that previously required collaborators.

This paper reconciles the two observations and converts them into a dated, falsifiable prediction. A laboratory pairs a fixed stock of leader judgment with execution assembled from tasks. A share ϕ of tasks is irreducibly human; AI covers an increasing share of the rest. Better AI lowers the unit cost of execution, which expands laboratory scale, and lowers the member share of each unit. Member employment is the product of the two, and the product is quasi-concave in AI capability: team size rises at most once and falls at most once, with long-run employment equal to ϕ times laboratory scale. Among fields that share the remaining primitives and the adoption path and possess an interior peak, those with smaller ϕ peak first. Capability evidence places mathematics furthest along; fields built on physical execution are least exposed.

The contribution is to make three predictions that require increasingly strong assumptions. First, a fully codifiable team peaks when its effective automated task share reaches a closed-form threshold s^* , with median 0.91 under the priors below; if effective coverage rises toward a plateau below s^* , such teams grow and never turn. Second, among fields with shared primitives that possess an interior peak, those with smaller irreducible human-task shares peak first, so the most codifiable fields form the watchlist in which the first peaks should appear. Third, under a benchmark bridge from software capability to research tasks, a Monte Carlo dates the fully codifiable peak: median 2027.2 on the capability clock, 2028.5 with a diffusion lag. Workers' lifetime earnings are nonmonotone in the horizon over which a sectoral transition unfolds (Grigsby and Zorzi, 2026), so when, and how quickly, research teams change size matters for the people inside them. Team-size data through 2025 show no detectable response to predetermined AI exposure; survey evidence dates research adoption of coding agents to late 2025, so the null is the pre-diffusion baseline rather than a rejection.

Task-based automation models separate displacement from productivity effects (Acemoglu and Restrepo, 2019), task-based models of AI in science ask which scientists gain (Agrawal, McHale and Oettl, 2026), and automation's inequality effects need not be monotone in capability (Benzell and Myers, 2026). Closest to this study is Ide and Talamàs (2025), where basic AI in knowledge hierar-

chies shrinks firms and advanced AI enlarges them; here the nonmonotonicity concerns the research team, whose long rise is documented by Wuchty, Jones and Uzzi (2007) and Jones (2021), and the model returns a dated, falsifiable turning point. The model borrows its distinction between leader ability and scalable execution from Sadun et al. (2025); the leader-quality implication is standard complementarity (Kremer, 1993); and field evidence that generative AI raises the productivity of less-experienced workers most (Brynjolfsson, Li and Raymond, 2025) is consistent with codifiable execution being what the technology transmits.

2 Model

A laboratory has a leader with judgment ability $C > 0$ and one unit of attention. Project i produces

$$y_i = E_i^\beta (Ct_i)^{1-\beta}, \quad E_i \geq 0, \quad t_i \geq 0, \quad 0 < \beta < 1, \quad \sum_i t_i \leq 1, \quad (1)$$

where E_i is execution and t_i is leader attention.

Lemma 1 (Span of control). *Given executions (E_i) with $X = \sum_i E_i > 0$, optimal attention is $t_i = E_i/X$ and laboratory output is $C^{1-\beta} X^\beta$.*

Execution requires a unit measure of tasks. A share $\phi \in [0, 1)$ is irreducibly human: a person remains necessary as software capability advances. Physical interaction is a sufficient source of this requirement but not the only one. The remaining $1 - \phi$ tasks are codifiable, with automation difficulty described by a strictly increasing distribution function $G : [0, \infty) \rightarrow [0, 1)$ that is continuously differentiable on $(0, \infty)$, right-differentiable at zero, satisfies $G(0) = 0$, and obeys $G(a) \rightarrow 1$ as capability $a \rightarrow \infty$. The automated task share is $\sigma(a) = (1 - \phi)G(a)$. A member task costs q , an AI task costs p_A , with $0 < p_A < q$, so unit execution cost is $c(a) = q - (q - p_A)\sigma(a)$, with limit $c_\infty = \phi q + (1 - \phi)p_A > 0$. The laboratory solves $\max_{X \geq 0} C^{1-\beta} X^\beta - c(a)X$, giving scale $X^* = C(\beta/c)^{1/(1-\beta)}$. Members supply the nonautomated task share, so member employment and total team size are

$$m^*(a) = X^*(a) [1 - \sigma(a)], \quad N^*(a) = 1 + m^*(a). \quad (2)$$

The added one is the leader.

Theorem 1 (At most one peak). *For every $\beta \in (0, 1)$, $0 < p_A < q$, and $\phi \in [0, 1)$, m^* and N^* are quasi-concave in a : they have at most one turning point, from increasing to decreasing, and $m_\infty = \phi X_\infty$ with $X_\infty = C(\beta/c_\infty)^{1/(1-\beta)}$. Thus $N_\infty = 1$ at the fully codifiable limit $\phi = 0$, while $N_\infty > 1$ at every $\phi > 0$.*

Corollary 1 (Peak coverage and the field threshold). *Let $r = p_A/q$. Team size rises and then falls if and only if $r < \beta$ and $\phi < \phi^*$, where the automated share at the peak and the field threshold are*

$$s^* = \frac{\beta - r}{\beta(1 - r)}, \quad \phi^* = \frac{r(1 - \beta)}{\beta(1 - r)}, \quad s^* + \phi^* = 1. \quad (3)$$

If $r \geq \beta$, team size never rises; if $r < \beta$ and $\phi \geq \phi^$, it rises to its long-run limit.*

Proofs are in the supplemental appendix. The identity $s^* + \phi^* = 1$ makes the thresholds one object. A field peaks in the interior exactly when its human share is below the nonautomated share at which teams turn. Near $a = 0$, team-size growth is decreasing in ϕ , so at low capability author counts should grow faster in work with less physical exposure; the ordering is exact at $a = 0$ and extends near zero by continuity. This is the sign tested below. These are partial-equilibrium comparative statics: the member price, AI price, and leader distribution are held fixed. A falling relative AI cost raises s^* and shrinks ϕ^* , moving positive- ϕ fields toward the monotone regime; the supplemental appendix discusses time-varying prices and substitution.

3 Evidence through 2025

Measurement precedes outcomes. I selected thirty OpenAlex subfields (Priem, Piwowar and Orr, 2022) spanning theory, computation, fieldwork, clinical work, and laboratory science, and scored 120 pre-period articles per field (2015 and 2019; 3,600 abstracts) under a written rubric: zero for theory, simulation, software, or secondary-data work; one for bench, field, clinical, or fabrication work; one-half for both. The field mean $\hat{\phi}_f$ is an ordinal proxy for the model’s human-task share. Using it assumes that physical exposure orders the broader human-task share; physical interaction is one source of irreducibly human work, not all of it. The proxy runs from zero in algebra and theoretical computer science to above 0.9 in organic chemistry (Figure 1). An external check agrees: fields mapped to a teleworkable research occupation under Dingel and Neiman (2020) average 0.13; the rest average 0.73.

At the field level, the raw post-2022 interaction of author counts with $\hat{\phi}_f$ is positive, the wrong sign, but physical fields had been trending toward larger teams for two decades, and with field-specific trends the interaction turns negative. Within fields, computational papers gain modestly on physical papers after 2022. The adjusted estimates match the rising-limb sign; none is statistically distinguishable from zero (supplemental appendix, Table 2).

At the researcher level, I follow 2,279 principal investigators, identified from 2015–2019 records, monthly across thirteen preprint servers (36,186 preprints), with each investigator’s physical exposure scored from up to twenty pre-2020 abstracts. The outcome is the log author count; the regression interacts predetermined exposure with mapped capability under investigator, field-by-month, and server fixed effects, so identification is within-field. Lemma 1 leaves the division of execution across projects indeterminate, so authors per paper is one margin of laboratory employment; the scale margin is reported alongside it. Through December 2025 the interaction is 0.19 (standard error 0.26), the wrong sign for the rising limb and indistinguishable from zero. Pre-period interactions are jointly indistinguishable from zero, though this does not establish parallel trends, and a provisional extension through June 2026 is likewise null. On the scale margin, annual papers and total credited authorships per investigator give the same null (supplemental appendix, Table 2).

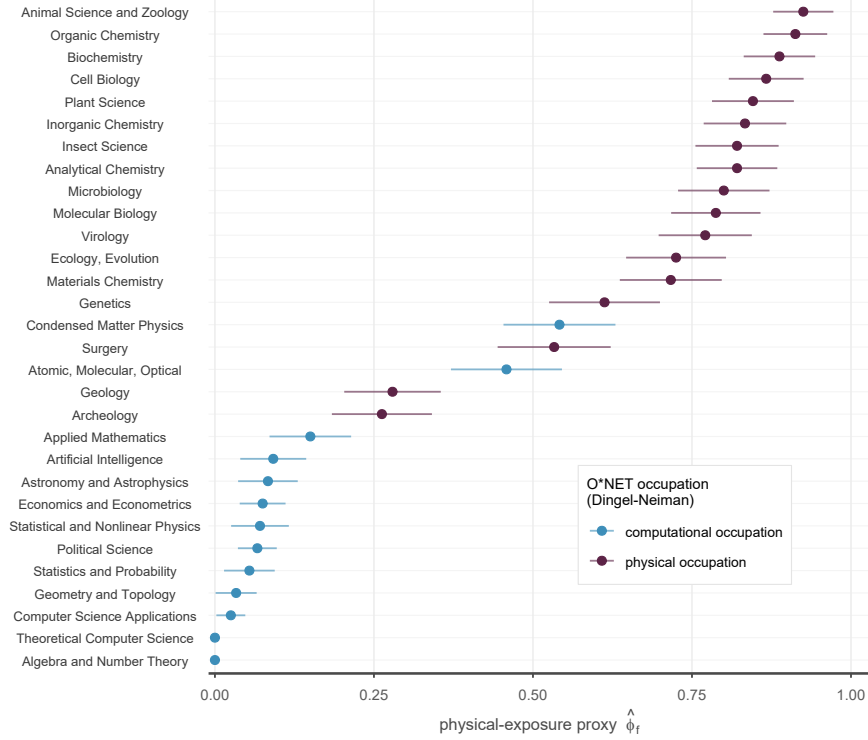


Figure 1: Physical-exposure proxy by field. Each point is the mean execution-mode score in 120 pre-period articles. Bars are 95 percent sampling intervals. Color marks teleworkability of the mapped occupation under Dingel and Neiman (2020).

In a February–March 2026 survey of 1,260 quantitative social scientists, 20 percent used coding agents regularly, and the surge in use dates to late December 2025 (Lyttelton, Massenkoff and Wilmers, 2026). Teams form before preprints appear, so the first plausible author-count response arrives in the second half of 2026 or later. With negligible effective adoption, the model predicts approximately no differential response through 2025, and none is detected.

4 What to Watch, and When

The model’s content is a hierarchy of three predictions.

Coverage. With no calendar mapping at all, a fully codifiable team peaks when its effective automated task share reaches s^* in equation (3). When $\phi < \phi^*$, a field with human share ϕ peaks at codifiable-task coverage $s^*/(1 - \phi)$; under a shared G and adoption path, such positive- ϕ fields peak later than the fully codifiable benchmark. Under the priors below, s^* has median 0.91 (90 percent interval $[0.75, 0.97]$). Effective coverage is capability discounted by adoption. Under the maintained pre-diffusion interpretation, current effective coverage remains below the threshold. Conditional on that interpretation, a peak typically requires high, often near-complete, long-run adoption. If effective coverage for a fully codifiable team rises toward a plateau below s^* , team size

never turns downward. Capability evaluations and adoption surveys in codifiable domains are the quantities to monitor.

Ordering. Among fields that share β , r , G , and the adoption path and satisfy $\phi < \phi^*$, lower- ϕ fields peak earlier. At benchmark parameters ($\beta = 0.6$, $r = 0.10$, $\phi^* = 0.074$), treating the ordinal proxy cardinally, seven of the thirty frozen proxies fall below the threshold: algebra and number theory, theoretical computer science, computer science applications, geometry and topology, statistics and probability, political science, and statistical physics. These fields form the benchmark watchlist in which the first author-count peaks should appear. Equivalently, the watchlist is the seven lowest-ranked proxies; its boundary is thin, with the eighth-ranked field at 0.075 against a threshold of 0.074, which is why the dated test below uses only the two zero-score fields. A peak in a high- $\hat{\phi}$ field before the watchlist turns would reject the joint hypothesis of the ordering, the proxy mapping, and the shared primitives, not just the date.

Calendar. Mapping the METR 50-percent software time horizon $a(t)$ (Kwa et al., 2025) into coverage by $G[a(t)] = a(t)/[a(t) + h_{50}]$, where h_{50} is the median codifiable research-task duration, a Monte Carlo with 20,000 draws propagates explicit priors: $\beta \sim U(0.4, 0.8)$, $r \sim U(0.05, 0.20)$, h_{50} log-uniform on 2 to 24 hours, and a pairs bootstrap of the post-2023 METR trend. On this capability clock a fully codifiable team peaks at median 2027.2 (90 percent interval [2026.4, 2028.1]). Adding an adoption-and-output lag of 0.5 to 2.0 years, anchored to the survey timing above and conditional on eventual full adoption, moves the median to 2028.5 ([2027.4, 2029.6]). Drawing a long-run adoption ceiling from $U(0.8, 1.0)$ instead, the peak occurs in 50 percent of draws, with conditional median 2028.6 ([2027.5, 2030.2]); the supplemental appendix reports peak probabilities at fixed ceilings (Figure 2). Benchmark performance can also misstate deployment value when task mixes differ (Gans, 2026), a further reason this calendar layer requires the most assumptions.

The through-2025 estimates predate the capability-clock window and fix the baseline for the test; the provisional 2026 extension overlaps the window’s left edge and is reported for completeness, not as part of the baseline. The test has two stages, set out in the supplemental appendix. Stage one establishes from capability evaluations and adoption surveys that effective coverage has crossed s^* . Stage two reestimates the researcher-level specification, and the watchlist fields’ author-count paths, on records through the fourth quarter of 2029, retrieved in the first half of 2030. Because a field with $\phi > 0$ peaks later than the fully codifiable benchmark, the dated test applies to the two zero-score fields, algebra and number theory and theoretical computer science. Conditional on a confirmed crossing, no author-count peak there rejects the joint hypothesis of the model, the capability bridge, the near-zero- ϕ interpretation of the proxy, and a stable mapping from laboratory employment to credited authors. It would not reject quasi-concavity itself, which is a shape restriction: under the same maintained mapping, that is rejected by a second turning point. Two qualifications bound what the test can show: employment funded by preferences or fellowships can hold headcounts above the production-side prediction, and a laboratory can grow by running more projects at constant authors per paper.

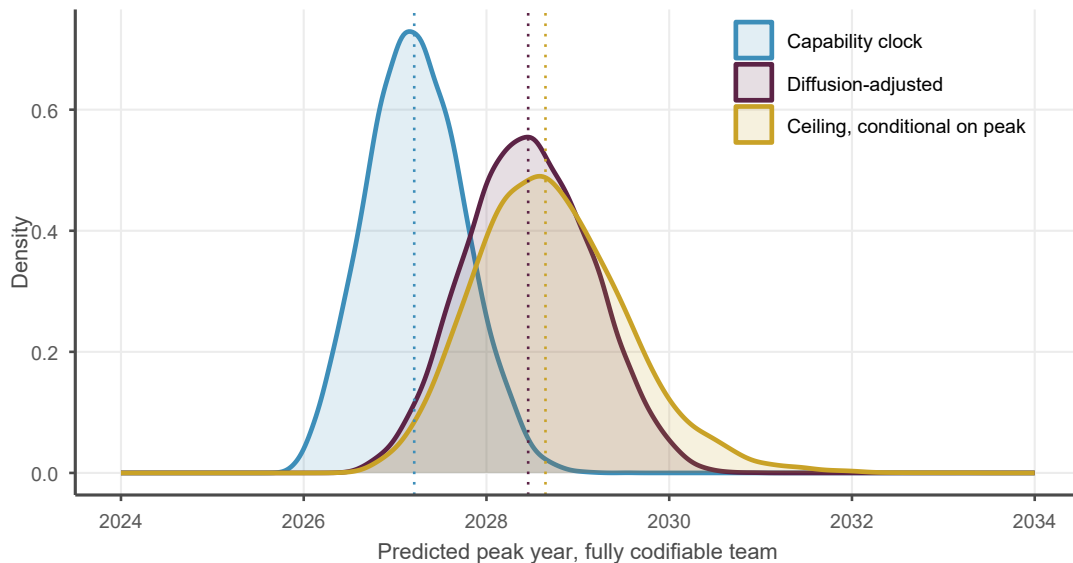


Figure 2: Predictive distribution of the peak year for a fully codifiable team. The capability clock assumes instantaneous adoption; the diffusion-adjusted scenario adds an adoption-and-output lag of 0.5 to 2.0 years and conditions on eventual full adoption; the ceiling scenario draws long-run adoption from $U(0.8, 1.0)$ and is plotted conditional on a peak occurring. Dotted lines mark medians.

5 Conclusion

AI changes research-team size through scale and substitution, and their product has at most one peak. The model reduces disagreement about AI and scientific employment to observables: coverage relative to s^* , the ordering of fields by task content, and a dated turning point. Its central forecast is that 2027 is the last year research teams grow in algebra and number theory and in theoretical computer science if capability alone governs adoption, and 2028 once realistic diffusion lags are added. The forecast is conditional on adoption. If long-run adoption converges to a level between 80 and 100 percent of codifiable tasks, the peak occurs with probability of about one half; below that range it does not occur at all. In the fully codifiable limit the laboratory is one researcher exercising judgment over machine execution. If member execution is where judgment is learned, sustained automation could erode the formation of the model’s fixed factor; Acemoglu, Kong and Ozdaglar (2026) formalize the general mechanism, AI substituting for human learning. The cost of the transition then depends on its speed and on labor-market frictions (Grigsby and Zorzi, 2026). Early cases in mathematics already exhibit that endpoint. For the lowest-exposure watchlist fields, the data needed to assess the dated prediction arrive within the next five years.

References

Acemoglu, Daron, and Pascual Restrepo. 2019. “Automation and New Tasks: How Technol-

- ogy Displaces and Reinstates Labor.” *Journal of Economic Perspectives*, 33(2): 3–30.
- Acemoglu, Daron, Dingwen Kong, and Asuman Ozdaglar.** 2026. “AI, Human Cognition and Knowledge Collapse.” National Bureau of Economic Research Working Paper 34910.
- Agrawal, Ajay K., John McHale, and Alexander Oettl.** 2026. “AI in Science.” National Bureau of Economic Research Working Paper 34953.
- Benzell, Seth Gordon, and Kyle R. Myers.** 2026. “Automation Experiments and Inequality.” National Bureau of Economic Research Working Paper 34668.
- Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond.** 2025. “Generative AI at Work.” *Quarterly Journal of Economics*, 140(2): 889–942.
- Dingel, Jonathan I., and Brent Neiman.** 2020. “How Many Jobs Can Be Done at Home?” *Journal of Public Economics*, 189: 104235.
- Erdős Problems Wiki contributors.** 2026. “AI Contributions to Erdős Problems.” *GitHub wiki*, Data through June 30, 2026; commit c8ad4309d20120c67cb97faa86daa1443acee018.
- Gans, Joshua S.** 2026. “Artificial Jagged Intelligence: When AI Benchmarks Misstate Deployment Value.” National Bureau of Economic Research Working Paper 34712.
- Grigsby, John R., and Nathan Zorzi.** 2026. “The Labor Market Consequences of Rapid Sectoral Shifts.” National Bureau of Economic Research Working Paper 34922.
- Hao, Qian Yue, Fengli Xu, Yong Li, and James Evans.** 2026. “Artificial Intelligence Tools Expand Scientists’ Impact but Contract Science’s Focus.” *Nature*, 649: 1237–1243.
- Hosseinioun, Moh, Brian Uzzi, and Henrik Barslund Fosse.** 2026. “Artificial Intelligence in Science: Returns, Reallocation, and Reorganization.” arXiv preprint 2603.27956.
- Ide, Enrique, and Eduard Talamàs.** 2025. “Artificial Intelligence in the Knowledge Economy.” *Journal of Political Economy*, 133(12): 3762–3800.
- Jones, Benjamin F.** 2021. “The Rise of Research Teams: Benefits and Costs in Economics.” *Journal of Economic Perspectives*, 35(2): 191–216.
- Kremer, Michael.** 1993. “The O-Ring Theory of Economic Development.” *Quarterly Journal of Economics*, 108(3): 551–575.
- Kwa, Thomas, Ben West, Joel Becker, et al.** 2025. “Measuring AI Ability to Complete Long Tasks.” *METR*, arXiv:2503.14499.
- Lyttelton, Thomas, Maxim Massenkoff, and Nathan Wilmers.** 2026. “Coding Agents in the Social Sciences.” *Anthropic Economic Research*.

- Priem, Jason, Heather Piwovar, and Richard Orr.** 2022. “OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts.” *OurResearch*, arXiv:2205.01833.
- Sadun, Raffaella, Rachel J. Schuh, Jonathan S. Hartley, John Van Reenen, and Nicholas Bloom.** 2025. “Management and Firm Dynamism.” National Bureau of Economic Research Working Paper 33765.
- Sothanaphan, Nat.** 2026. “Resolution of Erdős Problem #728: A Writeup of Aristotle’s Lean Proof.” *arXiv:2601.07421*.
- Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi.** 2007. “The Increasing Dominance of Teams in Production of Knowledge.” *Science*, 316(5827): 1036–1039.