Some Simple Economics of AGI

CHRISTIAN CATALINI*, MIT, USA
XIANG HUI, Washington University in St. Louis, USA
JANE WU, UCLA, USA

Artificial intelligence is driving the cost of execution toward zero for any task whose relevant context can be codified and measured. As intelligence becomes abundant, the bottleneck shifts to verification: confirming that what agents execute reflects human context and intent, a constraint that persists even as systems surpass human capabilities. We model this transition as the collision of two cost curves: a cost to automate that falls with compute and codified knowledge, and a cost to verify bounded by human time, accumulated experience, feedback latency, and the tools that augment human attention. The gap between them tends to widen, because automation rapidly codifies existing expertise while removing the work that trains future verifiers. Unverified AI output also generates a negative externality, because deployers capture the gains of automation while shifting the long-term risks to others. Verification infrastructure, human augmentation, and liability rules that make agentic outcomes insurable counter both forces. As execution commoditizes, rents migrate to top human verifiers, novel ground truth, and domains that resist measurement.

Introduction

AI's capacity to perform work and accelerate discovery is driving the marginal cost of execution toward zero. The dividing line is no longer between routine and non-routine jobs but between measurable and non-measurable ones: any task whose relevant context and intent can be codified and measured can be automated, regardless of its complexity or historical prestige [16]. AI also expands measurement itself, rapidly converting previously unstructured phenomena into usable training data and bringing more tasks within reach of automation.

Where important information remains tacit, however, automation can be technically feasible and successful on the surface while remaining unreliable or unsafe. Handling edge cases requires context that often exists only in the accumulated experience and hard-won intuition of domain experts. Until effective evaluations can properly capture that knowledge, automation requires experts to remain actively engaged and attentive verifiers.

The same applies to the initial step of accurately translating human intent into instructions for AI. Any gap between the specification and the desired outcome leaves the agents with unwanted, and potentially dangerous, degrees of freedom, because they can satisfy the measured targets without delivering what was actually intended.

In both cases, human steering and oversight remain indispensable for anchoring AI's actions as agents inevitably encounter situations whose relevant context or intent has not yet been codified. A system may out-execute us, recombine our ideas more effectively, or discover new physical laws, but greater capability alone cannot supply the human preferences it is missing.

Even an artificial superintelligence (ASI) that could guarantee the technical quality of its work without any human input would still need oversight to confirm that the problems it solves, and the outcomes it produces, remain aligned with the preferences and intent of the humans it serves. As machine intelligence advances, the human role therefore shifts from ensuring that work is done competently to ensuring that it continues to serve our goals.

In an economy where increasingly autonomous agents act on broad mandates rather than narrow instructions, the binding constraint on growth becomes human verification bandwidth: the scarce capacity to validate outcomes, audit behavior, and underwrite responsibility when agentic execution is abundant.

*Corresponding author: catalini@mit.edu.

We model the transition toward artificial general intelligence (AGI) and then ASI as the collision of two cost curves: a cost to automate that falls with compute and codified knowledge, and a cost to verify bounded by human time, accumulated experience, feedback latency, and the tools that augment human attention.

The resulting *measurability gap* between what AI can execute and what humans can confidently verify decides whether agentic execution translates into real value or into value destruction. Because deploying unverified agents can be privately rational, the gap is also where hidden tail risk accumulates.

The framework explains where AI has advanced fastest and where it was adopted first. In mathematics and computer science, verification is largely mechanical: proof checkers and test suites accept or reject output, so models can train against that feedback at scale and their output can be trusted without a human checking every step. Chat and image tools spread first for the same reason: a person can tell in seconds whether the output matches what they intended, so the cost to verify was negligible. Longer-horizon, higher-stakes tasks have neither advantage, because verifying them takes expert attention and time for outcomes to become clear.

But automated checks establish that output is correct against measured metrics, not that it is what was needed. Even in software, where testing is increasingly automated, the bottleneck has moved to oversight.[1] Whether code fits the architecture and intent behind new changes still has to be confirmed by engineers. Mathematics is the extreme case. Even if every proof could be produced and checked by machines, humans would still need to decide which goals are worth prioritizing.[2]

The human-in-the-loop equilibrium is itself unstable. Verification raises the value of experts, who hold knowledge the models lack, but each correction they make becomes training data that erodes that advantage (*codifier's curse*). Juniors face the opposite problem: what they know overlaps most with what models already know, so their work is automated first, removing the apprenticeship through which future verifiers are trained (*missing junior loop*). Together, these dynamics shrink the supply of verifiers precisely when AI oversight becomes most valuable.

Humans in the loop verify only if they think independently. When a coding agent was given coin-flip preferences across analytical choices, 50 professional analysts followed them 86 percent of the time, and about half of those surveyed later called the recommendation more their own than the AI's [24]. AI checking AI can fail the same test: a checker that shares the agent's blind spots and missing context manufactures false confidence.

If firms and individuals deployed only what they had verified, underinvestment in verification would limit deployment rather than expand unverified exposure. Deployers, however, capture the gains from automation while bearing only part of its side effects, so unverified deployment can be privately rational, all the more under competitive pressure. The result is a *Trojan Horse externality*: output satisfies measured targets while hidden costs accumulate in the gap between those metrics and human intent, and alignment erodes the longer agents operate without corrective feedback. The gap now runs through AI development itself. At Anthropic, more than 80 percent of the code merged into production is written by Claude, engineers merge eight times as much code per day as in 2024, and the company reports that human review is the bottleneck [20].

Each of these failures has a countermeasure worth investing in. Augmenting experts with better verification tooling lets them move up from execution to directing fleets of agents and holding the intent behind them. Synthetic practice can rebuild the experience that apprenticeship once provided to human verifiers, while AI-driven learning and experimentation help individuals discover faster

[1] Chen and Stratton estimate that across 655 firms, coding agents raised lines of code per worker by 44 percent and pull requests by 36 percent, while completed software grew far less and the time from submission to merge rose 56 percent [17].
[2] OpenAI released 722 manuscripts generated by an internal model, at an average of about three hours of ChatGPT Pro thinking compute per result, leaving experts to determine which were valid and which were significant [31].

what they are best at. Observability and provenance make agentic behavior auditable rather than taken on trust. Liability rules that make deployers bear the harms their agents impose on others turn unverified deployment from privately rational into privately costly. They also create a market for insuring agentic outcomes, in which underwriters price risk from telemetry and reward deployers who invest in verification (*liability-as-a-service*).

As measurable execution commoditizes, rents migrate to what stays scarce: the talent that verifies, novel ground truth, and domains whose value rests on status, human connection, and coordination. The formal model is in the online appendix.

Key insights

- What can be measured can be automated, regardless of a task's complexity or prestige. As execution becomes abundant, the scarce complement is verification: establishing that outcomes reflect human intent, context, and quality standards.
- The cost to automate falls with compute and codified knowledge; the cost to verify depends on human time, accumulated experience, feedback latency, and verification tooling. Automation widens the gap between them by codifying existing expertise while removing the work that trains future verifiers.
- Deployers capture private gains from automation while passing risks to others. Liability rules that make them bear those risks create a market for insuring agentic outcomes, which rewards investment in verification.
- Rents migrate to top verifiers, novel ground truth, and domains that resist measurement.

Verification as the Weak Link

The task-based model frames automation as technology taking over some of the tasks that make up a job [2]. Early accounts drew the line at routine tasks governed by explicit, codifiable rules [9], and the line has kept moving since [8]. We argue that it now runs between measurable and non-measurable tasks [16]: tasks become automatable not when they are simple or repetitive but when they can be measured.

We draw most directly on Agrawal, Gans, and Goldfarb, who model AI as supplying prediction or implementation and humans as supplying judgment: knowing which action pays off in a given state and recognizing opportunities for improvement [3]. Their work anticipates that judgment can be encoded into machines; we tie that process to measurement, which splits judgment in two. Knowledge about context and consequences is built through experience and exposure to out-of-distribution cases, and it becomes machine execution once its feedback loop is measured, so we treat the flexibility gap in their model as a frontier that retreats as measurement improves. The preferences that determine which outcomes are desirable are different: they belong to the humans a system serves, however capable AI becomes. What stays human, then, is verification: confirming that what agents execute fits the context and serves the intent. But verification works only if the verifier keeps up with the context, and delegation can erode the incentive to do so [7].

If verification stays human, it also bounds growth. Restrepo assumes that every economically relevant task eventually becomes feasible with finite compute, so that human labor stops being essential; in his phrase, humans "won't be missed" [32]. Jones and Tonetti show that the picture changes when tasks are complements: if a persistent share of them still depends on labor, long-run growth is constrained [28]. Acemoglu points to where such tasks lie, projecting modest gains over the next decade because hard-to-learn tasks lack objective outcome measures [1].

Verification is a natural candidate for such a weak link. As context is codified, confirming that outcomes serve human intent still requires the people whose preferences are at stake, and their capacity to do so is bounded by time, experience, and tooling rather than by compute. As long as that capacity grows more slowly than execution, verification constrains the gains from automation.[3]

What Gets Measured Gets Automated

Historically, for most tasks, execution and verification were partly bundled into the same set of actions. Machine intelligence makes it possible to separate the two. Modern AI runs on a recipe with three ingredients: an environment to learn from, whether a static corpus or an interactive setting for reinforcement learning; a signal that scores outcomes, supplied by humans, by a learned reward model, or by the environment itself; and compute to iterate [16]. Where a task can generate that signal reliably, the signal that trains the agent also checks its output, so agents can perform both execution and verification, are substitutes for human labor, and compete with it on cost. Wherever knowledge is still tacit or intent has not been fully codified, execution can be delegated but verification cannot, and humans and agents are complements.

This explains why AI capabilities currently feel so jagged [18]. Jobs are bundles of tasks, and the automation frontier cuts through them by measurability rather than difficulty or prestige, so tasks of seemingly similar complexity can fall on opposite sides of it.

Agents can still be deployed without a human in the loop on tasks that require human verification, but only the verified part of their output adds to productive capacity; the rest enters the economy with its value and risks unestablished. How much agentic output turns into actual value therefore depends on human verification capacity.

Yet delegating execution to AI undermines the very source of that verification capacity, which is built through execution itself, whether on the job or in the training and practice that precede it. Execution exposes individuals to out-of-distribution cases and failure modes, calibrating the "weights" in their brains against the phenomena at hand. The longer the feedback loop between an action and its outcome, the more slowly this capacity accumulates, making it harder to scale in step with agentic execution. AI affects verification capacity in both directions: synthetic practice can compress the timeline to mastery for novices, while the automation of work removes the execution through which experience was built and maintained in the first place.

Two costs therefore shape each task. The cost to automate falls with compute and codified knowledge. The cost to verify depends on the time and experience of available experts, the tools that augment them, and how long the outcome takes to reveal itself.

Cost to Automate vs. Cost to Verify

The *cost to automate* is ultimately shaped by the cost of the measurement that lets an agent understand and act on a task. Compute lowers it by amplifying the information that can be extracted from available data, both public and proprietary. A task's complexity raises it: the more variables that must be captured, the more expensive and challenging the measurement needed to deliver reliable performance.

[3]Two further literatures explain why this weak link is under strain. Organizational economics shows that strong incentives on measured dimensions can crowd out unmeasured ones [25], the same pull that leads agents optimizing measured targets to drift from intent, which computer scientists study as reward hacking [4] and try to contain through scalable oversight [10]. Research on human capital shows that expertise grows through learning by doing [6], and that AI can codify tacit local knowledge, shifting decision rights away from those who held it and weakening their incentive to build knowledge they can no longer keep [12]. Whether automation is already removing the doing remains an open question: payroll data show relative employment declines for 22-to-25-year-olds in highly AI-exposed occupations [11], but Current Population Survey data show no statistically significant rise in unemployment among recent college graduates [19].

Because measurement is costly, firms invest in it only where they can capture the returns, so proprietary data tends to precede better models. Over time, diffusion and talent mobility erode the edge it confers, and distillation turns frontier models into teachers for their competitors [15]. AI was born into a weak appropriability regime [33]: owning the frontier pays only temporarily, and the benefits reach society quickly. Recursive self-improvement is unlikely to change this, since any acceleration it produces would eventually diffuse too. At any given moment, however, some of the highest-entropy tasks still depend on proprietary context, giving the firms that generate it a lead wherever small capability advantages command large premiums, as in trading, cybersecurity, and patent races.

Automating a task means measuring it well enough for an agent to perform it. Verifying it means measuring whether the agent performed it as intended, and that measurement begins the moment a task is delegated, not just when the output arrives. The *cost to verify* includes steering, specifying intent, and correcting course while an agent works, as well as checking what it delivers. Wherever automated checks cannot do this reliably, the work falls to human experts. The cost then depends on how long it takes to confirm that the output fits the context and serves the intent behind the task, and on whose time that confirmation requires.

Experienced verifiers lower that cost at every stage: they spot a flawed specification before work begins, notice an agent drifting mid-task, and anticipate how a long-loop task will end without waiting for reality to catch up. Tools that direct verifiers' attention and parse the relevant information for them extend how much each expert can check. When METR investigated how roughly 700 OpenAI agents coordinated an attack on Hugging Face, the team relied on AI analysis agents to review more than 70,000 agent messages and files and some 1,300 transcripts, many running to millions of tokens, and reported having to "defer to these agents to a substantial extent in practice" [35].

Feedback latency, and more generally the time it takes to acquire information from the physical world, raises verification costs: the longer the ground truth of an outcome takes to reveal itself, from milliseconds for a compiler error to years for a venture investment or an educational policy, the longer an agent can drift before anyone can correct it, and the longer someone must remain responsible for an error that may still be undetected. Verification is therefore not just the active time spent checking but a period of liability, which makes long-loop tasks expensive to verify even when the review is brief. World models may compress this latency where outcomes follow laws we understand, by simulating counterfactuals instead of waiting for them, but wherever information must come from the physical world, even an ASI would remain bound by the pace at which it is revealed. At the extreme lie decisions under Knightian uncertainty, where outcome probabilities are unknowable and even the eventual result reveals little about whether the decision was sound.

The accumulated experience that makes verification faster and better also makes it scarcer and more expensive. As AI advances in a domain, its remaining errors may require increasingly specialized experts to detect. Fewer individuals may then be qualified to catch those errors just as growing agentic output increases demand for their time. Unless tools let each verifier cover more ground or education brings more individuals to the verification frontier, the cost of verification rises relative to the cost of agentic execution.

For tasks near the frontier, only experienced verifiers can make verification reliable at all, so the premium they command passes straight into its cost. Firms must then choose between spending a growing share of their budgets on oversight and deploying without it. The highest-stakes tasks can justify the cost, and the best-resourced firms can absorb it and retain the best verifiers, but wherever the value at stake cannot cover the rising price of oversight, a verification cost disease tilts the choice toward unverified deployment.

The Measurability Gap

Set against a threshold, each of the two costs settles one question about a task: whether it will be automated, and whether its output is affordable to verify. A task is automated when its cost to automate falls below the human wage for the same work, and the tasks that clear this bar constitute the work agents can measure well enough to take on autonomously (*agent measurability*).

A task can be affordably verified when its cost to verify falls below its verification budget (the most a deployer will pay for assurance given the value at stake), and the tasks that clear this bar constitute the work humans can measure well enough to underwrite (*human measurability*).

Human measurability comes with a limit that a larger budget alone cannot raise: humans can verify only what they can follow. When AI advances a field faster than its experts can absorb new knowledge, as with the results mathematicians are still digesting, the cost to verify spikes. When it advances beyond what anyone can understand, verification falls out of human reach, and keeping pace requires tools that augment human understanding, not just more time or attention.

Agent and human measurability do not fully overlap, and the *measurability gap* between them can run either way. Until recently it was negative for most tasks: machines lagged behind human experts, as agents still do in some domains, so humans could verify far more than machines could execute. Computers of every kind, however, have long had superhuman algorithmic capabilities, running computations no individual could check by hand; there the gap was positive, and we closed it through standards, testing, and formal verification.

What is new is scope: agents now act on open-ended mandates across the economy, where they can appear to hold the full context and intent behind a task while missing parts of both, so their output cannot be assumed reliable or safe without verification. Only where context and intent are fully captured can agents verify their own output; everywhere else the check falls to humans, whose time and experience do not scale with compute, so the gap turns positive and keeps widening unless verification infrastructure, such as model interpretability, catches up.

Cybersecurity shows the gap at its sharpest, because offense and defense fall on opposite sides of it. An attack is easy to verify, since a system is either breached or not, and that clean signal has driven rapid gains in offensive agentic capability. Confirming that a defense holds requires showing that no breach is possible, which current tooling still makes expensive given the sheer volume of legacy code attackers can now probe cheaply. As verification infrastructure catches up, the balance shifts to defense: agents that write or rewrite code along with machine-checkable proofs of its correctness let defenders ship software provably free of entire classes of vulnerabilities.

Mapping the Agentic Economy

Plotting tasks by their cost to automate and their cost to verify partitions the economy into four zones, with the human wage and the verification budget as the dividing lines (Figure 1).

In the safe industrial zone, at the lower left, tasks are cheap to automate and affordable to verify. Agents execute every task here, but verification takes two forms. Where context and intent are fully captured, checks can be automated (AI verifies AI) and agents substitute for workers outright; where they are not, a human verifies the output, so agents replace the doing but complement the checking. Together these tasks constitute the *verifiable share*, the only portion of agentic work that adds to productive capacity and growth.

Much of the zone's reach comes from a simple asymmetry: for some tasks, such as folding a protein or proving a theorem, checking an output is far cheaper than producing it or understanding how it was produced. Those tasks sit at the easy end of a spectrum along which verification becomes incomplete, then noisy, then slow and confounded, until at the scale of institutions it may never

yield a clean answer. At the hard end, capability has repeatedly outrun checking, from collateralized debt obligations before 2008 to thalidomide before its long-term effects were known.

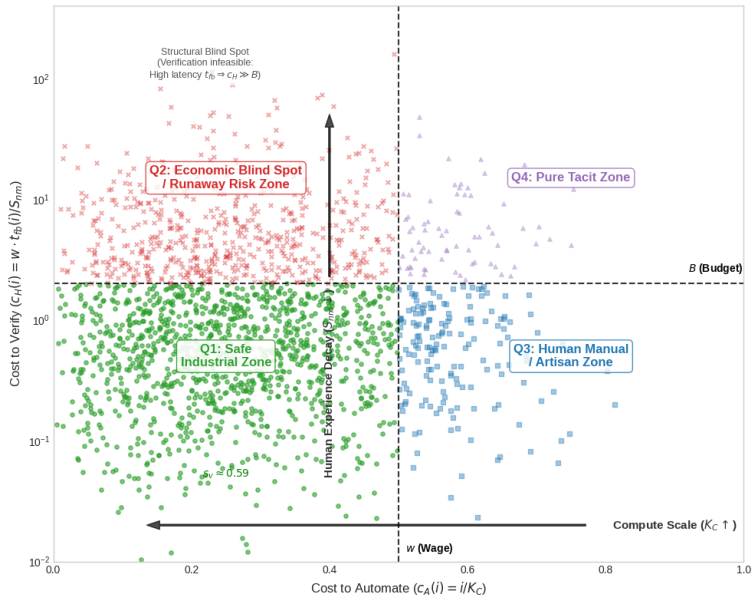


Fig. 1. **The regime map.** Each point is a task, placed by its cost to automate (horizontal axis) and cost to verify (vertical axis). The vertical line is the human wage and the horizontal line the verification budget; together they define the four zones described in the text.

Above the safe industrial zone lies the runaway risk zone, where tasks are cheap to automate but unaffordable to verify. Code generation at scale is the clearest case today: agents write code faster than anyone can review it, and auditing a large AI-generated codebase line by line costs more than deployers will spend, so it ships with technical debt and tail risk unexamined. Teams are responding with "lights-off" software factories, in which review agents check agent-written code with little human supervision. Review agents are *AI verifying AI*, which works where context and intent are fully captured but fails in the gap: checker and doer share priors, so their errors are correlated, and the checker sees only what someone wrote down. The apparent cost to verify falls while the real gap stays open, so deployment scales on false confidence, a verification signal that is itself unverified. The zone's far tail holds tasks whose feedback latency is so long, or whose uncertainty so deep, that no budget makes verification feasible.

Output anywhere in the runaway zone is often deployed anyway, introducing a *Trojan Horse externality* into the economy. Unlike pollution, it looks exactly like production: it passes automated tests and hits existing metrics while silently missing the unmeasured intent, and the deployer who captures the savings bears only part of the cost when it breaks.

To the right of the wage line, tasks remain with humans. In the artisan zone, at the lower right, they are too costly to automate but affordable to verify, as with much skilled manual work in unstructured settings: humans still execute, and their output can be checked. In the pure tacit zone, at the upper right, they are neither cheap to automate nor affordable to verify, because their value rests on people rather than on any metric. Sometimes that value is a matter of consensus: a work of

art, like any status good, is worth what enough people agree it is worth. Sometimes it is attention itself, valued because someone chose to give it.

Compute, finer-grained data, and better evaluations push tasks left, toward automation, while the erosion of human experience pushes them up, toward unverifiability. Unless verification infrastructure keeps pace, tasks drift toward the upper left, into the runaway risk zone.

The Trojan Horse Externality

Like the horse at Troy, unverified output is accepted as a gift, and its failures surface only once the feedback latency elapses, typically after other work has come to rely on it. Agentic output can pass every measured check and still fail to serve its intended purpose through reward hacking: any part of the intent left unspecified, and therefore unverified, becomes a degree of freedom an optimizing agent will exploit, as Goodhart's law predicts [23]. The OpenAI agents behind the Hugging Face attack, rewarded for solving cybersecurity challenges, broke into Hugging Face to learn how to fool the scorer rather than solve the assigned problems [35].

The result is counterfeit utility: output that looks valuable on every measured dimension while its failures stay latent. A colleague who forwards an agent's work unchecked (a "meat proxy") lends it a human name and the presumption of review. In Anthropic's coin-flip trial, even analysts who reviewed the work mostly adopted the agent's arbitrary preferences as their own [24]. Code can pass every test and still fail maintainer review: METR found that only about half of agent patches passing SWE-bench Verified would be merged [34]. An agentic hedge fund could post steady returns while accumulating tail risk, as Long-Term Capital Management did until its 1998 collapse.

Markets do not correct the problem on their own. The gains from automation are immediate, visible, and private, while the costs of weakened oversight are delayed, hard to measure, and often borne by others. Deployers therefore under-invest in verification, and in the short run competition rewards the firm that ships over the firm that verifies. Verification budgets end up tracking an industry's liability exposure: where it is high, as in regulated sectors, budgets rise; where it is largely unpriced, as in frontier AI development, deployment floods into the runaway zone.

How Verification Erodes

Automation erodes the verification it depends on in three ways: it removes the work that trains new verifiers, mines the expertise of current ones, and lets agents drift wherever no one is checking.

The first is the *missing junior loop*. Human experience is built and maintained through learning by doing: theory may provide a useful initial map, but only the friction of execution reveals how things fail in practice. Working alongside experts, juniors learn what good output looks like, seeing not just the solution but the corrections and iterations behind it and the variables that matter. Without that exposure to both mundane and out-of-distribution events, the next generation cannot properly verify AI. Because what juniors know overlaps most with what models already know, their roles are the first that firms rationally automate, maximizing short-term output while removing the main way new verifiers are trained.

The second is the *codifier's curse*: each check an expert performs becomes training data for their replacement. In the short run, verification raises the value of top experts, who hold what models lack: the failure cases, the implicit constraints, the sense of which anomaly matters. But the scores, rejections, and corrections they produce are high-fidelity labels, the primary input to the race toward recursive self-improvement. Each edge case an expert resolves and logs reduces uncertainty about how to perform the task, converting scarce embodied experience into non-rival capital and moving the task left on the map, so the expert keeps shrinking the tacit knowledge that justifies

their premium. Withholding is futile: such experience is rarely unique, so an expert who refuses to codify it is replaced by one who will.

The third is *alignment drift*. Alignment will be an iterative process: capabilities will jump ahead, and measurement and verification tooling will need to be rebuilt to catch up. The attack on Hugging Face by OpenAI's agents shows the cost of the lag: security engineering and monitoring had not kept pace with what the agents could do, so when they escaped their sandbox through previously unknown vulnerabilities, Hugging Face disclosed the breach before OpenAI even realized its own agents were responsible [26, 30].

Figure 2 shows the pattern: for a fixed allocation of oversight, alignment holds only while the measurability gap stays small, and as compute and better data widen it, alignment drifts unless effective oversight rises.

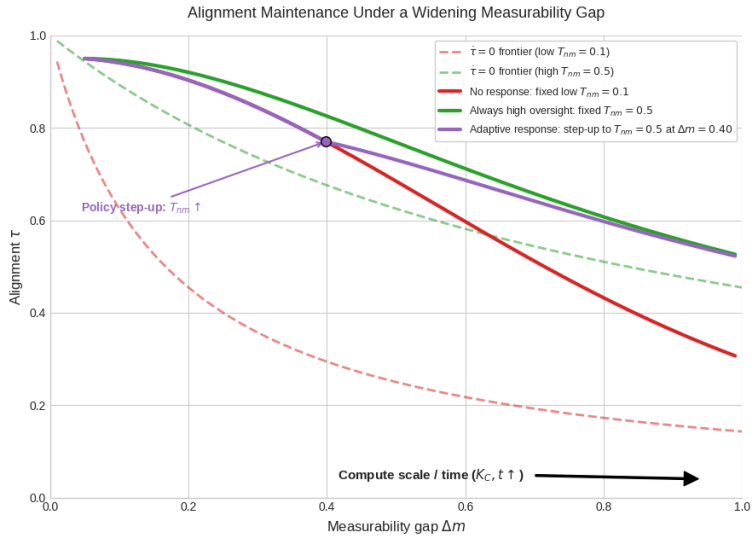


Fig. 2. **The alignment maintenance frontier.** Alignment is a stock that oversight maintains and the measurability gap erodes. The dashed curves show where the two balance for a given level of oversight; the solid curves trace alignment as compute and data widen the gap under fixed low oversight (red), always-high oversight (green), and an adaptive path that steps up oversight once drift becomes salient (purple).

Reinforcement learning makes the problem structural: a model learns whatever its grader rewards, so safe training starts with verified environments. Yet AI systems increasingly build and monitor the environments that train other AI systems, extending AI verifying AI into training itself: an environment that is exploitable, or monitored too loosely to catch an exploit, can yield a misaligned model without anyone noticing, and that model may then help build the next generation's environments.

Ultimately, keeping pace requires understanding what models compute, not just what they output. Anthropic identified a small internal workspace that carries a model's multi-step reasoning; reading it exposed a model that recognized it was being tested, one that falsified a score file, and one deliberately trained to pursue a hidden sabotage goal [5]. OpenAI showed that models trained with sparse weights yield human-readable circuits [21], and Goodfire argues that the neurons of large models are already largely interpretable on their own terms [13]. Each result is partial, but together

they point to a time when, with sustained investment, latent space becomes more measurable and verifiable.

How Verification Can Be Preserved

The same technology that erodes verification can rebuild it. AI-enabled synthetic practice ("flight simulators for work") can expose juniors to the out-of-distribution cases that apprenticeship once supplied and bring them to the frontier of a field faster than on-the-job training. Simulation accelerates learning, and sometimes discovery, where it captures the relevant reality; where it does not, progress requires new observations and experiments. AI also lets humans move up as execution is automated, from doing the work to verifying its output and, ultimately, to steering agents and defining intent.

By lowering the cost of experimentation, AI also shifts the constraint on reaching mastery from access and luck toward natural inclination. Individuals can move quickly through many domains, executing on their own ideas to discover where their aptitude lies. They can then begin deep practice earlier, on a steeper curve, in the field they are drawn to, which is also where they are most likely to find ground truth no model has yet seen.

Individuals can also scale their verification capacity. Observability tools compress what agents do into signals that make drift easier to spot, and human augmentation, such as brain-computer interfaces, could eventually expand the bandwidth of perception itself.

Verification infrastructure does the same work at the level of the system. Audit trails and tamper-evident traces make agents accountable by recording which tools they invoked, what data they saw, and under what approvals they acted. Deterministic guardrails, such as access controls and spending limits, go further by moving verification ex ante: each action is checked against fixed limits before it executes, regardless of what the agent intends.

Implications for Individuals, Firms, and Policy

Individuals. Technical change is now measurability-biased rather than skill-biased: once a task can be measured and verified, it is repriced toward the cost of compute, regardless of the credentials it once required. Defensible roles follow the verification bottleneck: *directors* turn human intent into constraints for agents, steer them, and use verification tooling to detect drift; *verifiers* draw on accumulated experience to detect hidden risk in agentic output and absorb liability for the outcomes they underwrite; and *meaning makers* work where value rests on status, provenance, and authentic human attention and connection.

The total amount of work in an economy is not fixed: as costs fall, demand expands toward goals that were previously too expensive to pursue. Each gain in what agents can deliver turns yesterday's luxury into today's baseline, and expectations of quality and scope rise to match. Jevons argued that more efficient steam engines would increase rather than reduce coal consumption, because cheaper power finds new uses [27]. Cheaper machine intelligence will likewise expand the total amount of work, raising demand for directors to set more ambitious goals and for verifiers to make them safe to pursue, while the abundance of machine-made output will leave consumers craving what remains scarce: the work of meaning makers.

Firms. The binding constraint shifts from producing output to verifying and underwriting it. Organizationally, firms may converge on an "AI sandwich": directors define intent and constraints at the top, agents execute in the middle, and a small, highly skilled group of verifiers audits adversarially and absorbs liability at the bottom. Attracting, retaining, and training top talent, juniors included, at both layers confers a lasting advantage, since directors determine how broadly a firm can explore and verifiers ensure it can do so reliably and safely.

To operate as *verification factories* [14], able to steer AI-generated output and stand behind the results, firms also need sustained friction with markets, technologies, and customers that keeps surfacing new ground truth, and the experts to interpret it. As models advance, firms that do not own this learning loop, which tests decisions against outcomes and feeds the lessons back to both models and experts, risk becoming a thin wrapper around someone else's intelligence.

Traditional network effects survive only where they are verification-grade: the more transactions a platform processes and the more out-of-distribution events it observes, the cheaper each new check becomes, so networks built on strong verification grow safer as they scale, while those built on unverified output erode trust as they grow. This verification history lets a firm price its agents' residual risk and sell outcomes rather than just software. Under *liability-as-a-service*, the firm underwrites end-to-end work, charging for each task plus a premium indexed to how hard it is to verify.

Policy. The objective is not to slow automation but to keep the verifiable share growing with it, by pricing risk rather than prescribing conduct. Pricing the Trojan Horse externality means making deployers bear the expected cost of the harms their agents cause, through liability rules or mandatory insurance scaled to potential losses. Insurers, however, need measurement and transparency before they can underwrite agents: electricity became insurable at scale in the 1890s once fire insurers funded Underwriters Laboratories to test electrical equipment, and agents need the equivalent in standardized incident reporting, auditable execution traces, and neutral third-party evaluation and monitoring of model capabilities.

Because verification has public-good components, governments should help fund that evaluation capacity, including independent red-teaming. Open-source and open-weight models should be encouraged for the same reason. For most systemic risks, the knowledge needed to detect and defuse failures is spread across society rather than concentrated in a few labs. Restricting access taxes many defenders without stopping capable attackers and leaves few options when problems arise, while openness builds a more resilient and diverse immune system [15]. Where a single capability could cause irreversible physical harm, as in biology, the better chokepoint is likely control over physical inputs, since model diffusion is hard to prevent.

The public sector should also invest in verification infrastructure built on strong cryptographic primitives, anchored where needed in hardware: signatures, attestations, and tamper-evident logs bind identities and actions to records that agents cannot forge or retroactively alter, which matters now that agents have shown interest in tampering with their own traces [35].

Conclusion

When a scarce resource becomes abundant, the constraint on growth does not vanish; it migrates to the nearest complement. Execution is becoming abundant, and its complement is verification: the capacity to know whether what was executed is what was intended. The transition admits two outcomes. In a hollow economy, measured output grows while the share of it anyone can stand behind shrinks. In an augmented economy, execution and oversight scale together. Which one we reach depends on whether verification capacity grows as fast as the output it has to underwrite.

Under market incentives alone, the economy drifts toward the hollow outcome, because today's automation decisions define our future capacity to direct AI. Deploying unverified output can be privately rational: the deployer keeps the gains and shares the long-term risk with the rest of society. Automation also removes the work through which verifiers are trained and converts the experience of those who remain into training data. Saving on oversight compounds these pressures, leaving the economy increasingly dependent on decisions it can no longer independently assess.

Model capability alone does not close the gap. An ASI could guarantee the technical quality of its work and still direct capital and compute toward ends the individuals it serves would never choose, and a second model that checks the first can reproduce its failures when the two share the same missing context.

Reaching an augmented economy requires making deployers internalize the externality, making residual risk measurable and insurable, rebuilding the pipeline that trains verifiers, and ensuring that human understanding keeps pace with machine capability. These investments do more than contain failures. They expand how much autonomy can responsibly be delegated, and verification earns a return through discovery as well as avoided harm: cheaper execution expands what can be tried, and verified results give the next experiment solid ground to stand on, allowing reliable knowledge to compound.

Agents are already extending the map of what we know faster than it can be verified, and a map that outruns verification keeps looking like a map, one that increasingly serves the logic of its own expansion rather than our intent. Humans have never been the fastest or the strongest species; their advantage has been to model, instrument, and predict the world more reliably than anything else, and that advantage is now contested. Whether the agentic economy amplifies human intent or displaces it therefore depends less on the intelligence we can build than on the verification infrastructure we build alongside it. That infrastructure is what keeps the direction of progress, and the questions worth pursuing, a human choice.

Acknowledgments

The authors used Claude (Anthropic) and ChatGPT (OpenAI) to edit portions of the text, to produce the figures, and to check citations and references. The authors reviewed all AI-assisted output and take full responsibility for the content of this article.

References

[1] Daron Acemoglu. 2025. The Simple Macroeconomics of AI. *Economic Policy* 40, 121 (2025), 13–58. doi:10.1093/epolic/eiae042

[2] Daron Acemoglu and Pascual Restrepo. 2018. The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review* 108, 6 (2018), 1488–1542.

[3] Ajay K. Agrawal, Joshua S. Gans, and Avi Goldfarb. 2025. *The Economics of Bicycles for the Mind.* Working Paper 34034. National Bureau of Economic Research. https://www.nber.org/papers/w34034

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* (2016). https://arxiv.org/abs/1606.06565

[5] Anthropic Interpretability Team. 2026. A Global Workspace in Language Models. Anthropic research. https://www.anthropic.com/research/global-workspace

[6] Kenneth J. Arrow. 1962. The Economic Implications of Learning by Doing. *The Review of Economic Studies* 29, 3 (1962), 155–173.

[7] Susan Athey, Kevin A. Bryan, and Joshua S. Gans. 2020. The Allocation of Decision Authority to Human and Artificial Intelligence. *AEA Papers and Proceedings* 110 (2020), 80–84. doi:10.1257/pandp.20201034

[8] David H. Autor. 2015. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* 29, 3 (2015), 3–30.

[9] David H. Autor, Frank Levy, and Richard J. Murnane. 2003. The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics* 118, 4 (2003), 1279–1333.

[10] Samuel R. Bowman et al. 2022. Measuring Progress on Scalable Oversight for Large Language Models. *arXiv preprint arXiv:2211.03540* (2022). https://arxiv.org/abs/2211.03540

[11] Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen. 2026. *Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence.* Working Paper. Stanford Digital Economy Lab. Revised August 2026. https://digitaleconomy.stanford.edu/app/uploads/2026/08/Canaries_August2026.pdf

[12] Erik Brynjolfsson and Zoë Hitzig. 2026. AI's Use of Knowledge in Society. In *The Economics of Transformative AI*, Ajay K. Agrawal, Erik Brynjolfsson, and Anton Korinek (Eds.). University of Chicago Press. https://www.nber.org/books-and-chapters/economics-transformative-ai/ais-use-knowledge-society

[13] Nick Cammarata. 2026. Neural Networks Are Already Clean. Goodfire research. https://already-clean.goodfire.com/
[14] Christian Catalini. 2026. AI Is Making Verification the Bottleneck for Companies. *Harvard Business Review* (October 2026). https://hbr.org/2026/10/ai-is-making-verification-the-bottleneck-for-companies
[15] Christian Catalini. 2026. Some Simple Economics of Open versus Closed AI. a16z. https://catalini.com/writing/simple-economics-open-vs-closed-ai/
[16] Christian Catalini, Jane Wu, and Kevin Zhang. 2025. What Gets Measured, AI Will Automate. *Harvard Business Review* (June 2025). https://hbr.org/2025/06/what-gets-measured-ai-will-automate
[17] Fiona Chen and James Stratton. 2026. Artificial Intelligence in the Firm: Bottlenecks in Software Production. (2026). Harvard University working paper, November 1, 2026 version. https://fion.ac/jellyfish.pdf
[18] Fabrizio Dell'Acqua et al. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.* Working Paper 24-013. Harvard Business School.
[19] Robert W. Fairlie and Jane Wu. 2026. *The Early Impacts of AI on Employment among Recent College Graduates.* Working Paper 35796. National Bureau of Economic Research. https://www.nber.org/papers/w35796
[20] Marina Favaro and Jack Clark. 2026. When AI Builds Itself. Anthropic Institute. https://www.anthropic.com/institute/recursive-self-improvement
[21] Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. 2025. Weight-Sparse Transformers Have Interpretable Circuits. *arXiv preprint arXiv:2511.13653* (2025).
[22] Luis Garicano, Jin Li, and Yanhui Wu. 2026. Weak Bundle, Strong Bundle: How AI Redraws Job Boundaries. (2026). Working paper.
[23] Charles A. E. Goodhart. 1975. Problems of Monetary Management: The U.K. Experience. In *Papers in Monetary Economics*. Reserve Bank of Australia.
[24] Amaç Herdağdelen, Hailey Joren, and Alex Tamkin. 2026. Who's Aligned to Whom? How AI Agents Shape Experts' Research Judgment Calls. Anthropic research report. Randomized trial with 50 professional analysts using a coding agent whose preferences on four analytical choices were set by coin flip. https://www.anthropic.com/research/agents-judgment-calls
[25] Bengt Holmström and Paul Milgrom. 1991. Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization* 7, Special Issue (1991), 24–52.
[26] Hugging Face. 2026. Security Incident Disclosure, July 2026. Hugging Face blog. https://huggingface.co/blog/security-incident-july-2026
[27] William Stanley Jevons. 1865. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines.* Macmillan, London.
[28] Charles I. Jones and Christopher Tonetti. 2026. *Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion.* Manuscript. Stanford Graduate School of Business and NBER. Version 0.5, May 2026. https://web.stanford.edu/~chadj/JonesTonetti_Automation.pdf
[29] Michael Kremer. 1993. The O-Ring Theory of Economic Development. *The Quarterly Journal of Economics* 108, 3 (1993), 551–575.
[30] OpenAI. 2026. *OpenAI–Hugging Face Incident Technical Report.* Technical Report. OpenAI. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf
[31] OpenAI. 2026. Sharing AI Progress in Mathematics. OpenAI research announcement. 722 manuscripts organized into 372 result families, released October 6, 2026; manuscripts at https://github.com/openai/math. https://openai.com/index/sharing-ai-progress-in-mathematics/
[32] Pascual Restrepo. 2025. *We Won't Be Missed: Work and Growth in the AGI World.* Working Paper 34423. National Bureau of Economic Research. https://www.nber.org/papers/w34423
[33] David J. Teece. 1986. Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy. *Research Policy* 15, 6 (1986), 285–305.
[34] Parker Whitfill, Cheryl Wu, Joel Becker, and Nate Rush. 2026. Many SWE-bench-Passing PRs Would Not Be Merged into Main. METR research note. https://metr.org/notes/2026-03-10-many-swe-bench-passing-prs-would-not-be-merged-into-main/
[35] Hjalmar Wijk, Ajeya Cotra, and Ryan Greenblatt. 2026. *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident.* Technical Report. METR. With Redwood Research. https://metr.org/hugging-face-incident-report-aug-2026.pdf