

THE RACE WITH NO FINISH LINE

AGI, Autonomous Agents, and the Human Cost of the Intelligence Race



Faizan Yousaf | Deaimer Research

Research perspective & proposed evaluation framework
21 October 2026

RESEARCH STATUS Evidence-grounded synthesis and proposed method. No original Deaimer experimental results or verified AGI predictions are claimed.

Abstract

On 22 September 2026, OpenAI announced GPT-6 Sol and Luna for Work and Codex while Anthropic announced Claude Opus 5.5. The coincidence captures an increasingly familiar rhythm: frontier models and agent products arrive in rapid succession, accompanied by striking capability claims, lower operating costs, and promises of greater autonomy. Yet a product release is not a measure of artificial general intelligence (AGI), a benchmark score is not a certification of real-world reliability, and access to an AI agent is not evidence that the people affected by its actions can contest or correct them. This research perspective examines a consequential transition in AI development: from demonstrating that a model *can* perform a task to authorizing a system *to act* on behalf of people and institutions. It reviews frontier-model competition, executable agent benchmarks, reward and evaluation failures, technical governance, AI labor, and the uneven social effects of automation. Building on established risk-management and societal-capacity research, it proposes the **Three Clocks of AI Readiness**: capability evidence, verification capacity, and human adaptation. Rather than assuming that these advance at comparable speeds, the article develops a deployment-specific **Delegation Readiness Ledger** and a proposed **unverified-delegation interval** to reveal when consequential authority precedes prespecified independent assurance or human safeguards. A worked software-deployment example and a falsifiable pilot protocol show what these concepts could measure—and why they cannot establish a universal safety score. The central argument is conditional: the pace of AI innovation becomes a governance problem when authority expands faster than evidence, accountability, and people's practical ability to intervene.

Keywords: AGI; frontier AI; coding agents; autonomy; Terminal-Bench; benchmark validity; reward hacking; independent evaluation; human oversight; digital labor; AI governance; reinforcement learning environments.

1. The day the frontier got crowded

On **22 September 2026**, two major AI developers made separate announcements. OpenAI's release notes introduced GPT-6 Sol and GPT-6 Luna in ChatGPT Work and Codex; Anthropic announced Claude Opus 5.5. Earlier that month the industry already had GPT-6 Astra, introduced on 3 September. Six days after the shared release date came Claude Sonnet 5.5; GPT-6.1 Sol followed on 29 September. Google made Gemini 4 Argon available to developers on 13 October. Those dates describe a remarkable *release cadence*, not a controlled experiment proving a similarly rapid change in general intelligence. [1–5]

Date	Publicly described event	What it establishes—and what it does not
3 Sep 2026	OpenAI introduced GPT-6 Astra. [1]	A frontier product release; not a universal AGI threshold.
22 Sep 2026	OpenAI introduced GPT-6 Sol and Luna for Work and Codex. [2]	Two additional model offerings in specified products; not that all GPT-6 variants debuted together.
22 Sep 2026	Anthropic announced Claude Opus 5.5. [3]	A same-day competing announcement, with provider-reported capability and cost claims.
28–29 Sep 2026	Anthropic announced Sonnet 5.5; OpenAI publicized GPT-6.1 Sol. [1,4]	Further release velocity, without apples-to-apples evidence of generalized intelligence.
13 Oct 2026	Google made Gemini 4 Argon available to developers. [5]	Another frontier offering focused on complex workflows, subject to evaluation conditions.

The dates are worth reporting precisely because the pattern is vivid enough without embellishment. They do **not** mean that every named model arrived on one day, that each model is a separate scientific revolution, or that benchmark claims from different vendors can be ranked without matching the benchmark version, agent scaffold, inference budget, and evaluation rules. Nor does one month establish a trend on its own. It is a compelling snapshot of competitive activity, not a forecast of AGI.

The incentives are substantial. The 2026 Stanford AI Index reports **\$285.9 billion** of U.S. private AI investment in 2025, while also warning that simple national comparisons can undercount some forms of public financing elsewhere. Commercial pressure rewards new capabilities and lower costs; scientific incentives reward results that can be reproduced; security and safety work must sometimes slow a deployment to investigate an inconvenient failure. These goals overlap, but they are not identical. [8]

A useful question is therefore not whether companies should stop innovating whenever a rival announces a model. It is whether the evidence required for high-impact deployment is improving along with the products being shipped. Release calendars are visible. The quality of evaluations, institutional readiness, affected-worker protections, and availability of effective recourse are harder to see.

This paper treats that visibility gap as a research question rather than a verdict about any particular firm. The race is real in the limited sense of sustained competition for products, investment, talent, and deployment. The claim that safety *necessarily* falls behind is a hypothesis requiring separate evidence.

2. A finish line nobody can agree on

An athletics race ends when participants cross a defined line. Artificial general intelligence has no universally accepted equivalent. Some definitions emphasize human-level performance across many economically valuable tasks; others emphasize adaptation to unfamiliar problems, breadth of competence, ability to acquire skills from limited experience, or levels of autonomy. Morris and colleagues' **Levels of AGI** separates breadth from depth of performance and considers autonomy as a distinct practical consideration. Chollet's account of intelligence emphasizes skill-acquisition efficiency and cautions that impressive performance may be partly purchased through large priors and extensive exposure. Neither framework supplies an uncontested date on which AGI must be declared complete. [6,7]

This is not a semantic distraction. If one observer calls an agent "general" because it succeeds at varied benchmarks, while another requires robust adaptation without tailored scaffolding, the two may disagree about progress without disagreeing about a single experiment. A frontier model may perform superbly on coding and poorly at maintaining state across an unfamiliar administrative workflow; high average scores can hide narrow but consequential failure modes. The correct empirical question names the task distribution, acceptable errors, degree of autonomy, and costs of failure.

There are at least four distinct claims often compressed into "the AI is getting smarter":

1. **Model capability:** a model solves a specified class of problems under a stated test protocol.
2. **Agent-system capability:** a model, tools, memory, instructions, and execution environment jointly complete workflows.
3. **Economic usefulness:** the full system improves outcomes once integration, supervision, error recovery, and operating costs are counted.
4. **Authority:** the system is permitted to make decisions or change consequential state without contemporaneous human approval.

Progress on the first does not prove the others. OpenAI's Codex app, for example, was described as a way to manage multiple coding agents working in parallel. Cursor's cloud agents can operate on cloned repositories and produce changes for review; Grok Bot's provider describes persistent AI teammates with computers that operate across applications. Such systems illustrate new *modes of delegation*. Their product descriptions are evidence of offered features and design intentions, not independent measurements of reliable performance across every customer workflow. Cursor also describes scheduled and event-triggered automations that extend agent operation beyond individual prompts. [11–13,38]

The physical infrastructure is also part of the race. The International Energy Agency's 2025 *Energy and AI* report projected data-center electricity consumption of roughly **945 TWh by 2030** in its base case, around double its 2024 level. This is a scenario involving data centers generally, not a direct measurement of electricity consumed exclusively by frontier-model training. Constraints on power, hardware, and infrastructure are reasons to study multiple progress trajectories rather than treat exponential growth as guaranteed. [10]

Three broad futures remain compatible with the evidence: capability improvement continues quickly in some domains; progress becomes bottlenecked by data, inference economics, physical infrastructure, or hard-to-automate tasks; or capabilities advance unevenly with sharp local breakthroughs and persistent weaknesses. It would be irresponsible to present a precise AGI arrival date as established fact. It would be equally irresponsible to assume that today's uncertainties remove the need for decisions about systems already being deployed.

The finish-line metaphor thus has a second meaning. Even after a model reaches whatever technical threshold a particular organization calls AGI, the questions of who can use it, what it may do, how it is tested, and who bears its mistakes remain. The science of intelligence and the governance of delegation operate on different timescales.

3. The scoreboard that cannot stand still

Benchmarks are essential instruments. They permit teams to compare defined capabilities, diagnose failure modes, and evaluate changes more consistently than promotional demonstrations. Yet the meaning of a benchmark result depends on an entire measurement system: task construction, allowed tools, model version, agent harness, time and token budgets, environment state, grading logic, and uncertainty. A percentage without these conditions is a poor unit of scientific communication.

Terminal-Bench is particularly revealing because its tasks require agents to work in executable terminal environments, not merely choose answers from a static list. The public **Terminal-Bench 4.0** leaderboard, viewed in October 2026, reported the following selected *maximum-effort model-agent configurations*. The values are leaderboard reports, not independent Deaimer experiments. [14]

Model / reported configuration	Agent scaffold	TB 4.0 resolution rate	Listed interval
Claude Opus 5.5 (max)	Claude Code	64.8%	±3.1 percentage points
Claude Sonnet 5.5 (max)	Claude Code	61.8%	±2.9 percentage points
GPT-6 Astra (max)	Codex	58.2%	±2.8 percentage points
GPT-6.1 Sol (max)	Codex	58.2%	±3.1 percentage points

These are **system configurations**, not isolated model IQ scores. Intervals overlap, and differences among reported point estimates do not by themselves establish that one model is superior under a matched statistical comparison. The leaderboard also reports considerable variation in token use and cost, so success rates should be read together with resource consumption rather than detached from it. [14]

More instructive than the top row is what the benchmark maintainers did to maintain the instrument. In its August 2026 release, Terminal-Bench 4.0 **removed eight tasks**, including cases affected by saturation, refusals, public solutions, and quality or compatibility problems; **fixed 19 tasks**; and moved to a flat eight-hour agent timeout after resource calibration. The authors explicitly treat benchmark versions as software changes that can require trials to be rerun. An apparently higher score on another version cannot automatically be interpreted as a longitudinal improvement. [15]

The team's earlier leaderboard-integrity policy documents both cheating and reward hacking, including cases in which public solutions were accessed or submitted artifacts undermined evaluation integrity. A container can isolate a process while a grader still rewards the wrong outcome, or while answer information leaks into the agent's reach. These concerns are not theoretical objections to benchmarking: they are documented maintenance challenges within a serious evaluation program. [16,35]

A second angle comes from **OSWorld 2.0**, whose authors report 108 long-horizon computer-use workflows, with a median skilled-human time of approximately 1.6 hours. In their tested 500-step setup, the strongest reported Claude Opus 4.8 configuration reached **20.6%** strict completion and **54.8%** partial credit. Those scores concern the named model, benchmark, and evaluation conditions; they are not predictions about a newer model or all human office work. The gap between partial and full completion highlights a familiar agent problem: many locally plausible steps can still fail to satisfy the final task. [17]

METR's task-time-horizon research offers a complementary measure: the human-time length of tasks completed at a specified reliability level under its protocol. Its initial analysis found a roughly seven-month doubling time for its selected software-task measure over the historical period studied. The authors caution that a task horizon is **not** the amount of elapsed time an agent can safely operate in an unconstrained real-world setting. Extrapolation is a scenario, not a promise. [9]

The **International AI Safety Report 2026** gives a broader name to the problem: an **evaluation gap** between the behavior observed in predeployment tests and the behavior that matters after deployment. Its analysis also describes an evidence dilemma for policymakers deciding under uncertainty. Our proposed framework must therefore acknowledge that concerns about evaluation gaps already exist in the literature. The potential contribution is not discovering that a gap exists; it is making the gap operational for a specific decision to grant authority. [19]

Measurement principle. "The agent passed" is a statement about an observed result under a particular protocol. "The agent may now act without supervision" is a separate policy judgment that requires additional evidence about harms, controls, and affected people.

4. When an answer becomes an action

The newest agent interfaces transform the risk profile of AI not only because models may be more capable, but because they may hold more *permissions*. A model generating a proposed patch, an agent committing a pull request, and an agent deploying that patch to production can rely on similar code-generation ability while exposing very different assets and stakeholders. A mistaken suggestion can be rejected; an unauthorized deployment may need reversal, incident response, compensation, or legal review.

Consider the deployment pipeline as a series of decisions: a task is specified; a model selects actions; tools execute changes; an environment records their consequences; a verifier judges observable evidence; an organization decides what authority is allowed next. A failure may occur at any boundary. Tool output may be stale. A test may accept a superficial shortcut. Permission boundaries may be too broad. A human approver may be unable to inspect the relevant consequences. Responsibility cannot be inferred solely from a model's final message.

The literature distinguishes several failure mechanisms worth keeping separate:

- **Incorrect competence:** the model misunderstands a task or produces a broken artifact.
- **Specification gaming:** an optimizer satisfies the measurable proxy without accomplishing the intended objective; for example, a superficial health-check fix can pass while the actual service remains broken. [35]
- **Distribution shift:** a strategy that worked within known codebases, tools, or scenarios fails when assumptions change. [17,33]
- **Misuse or unauthorized action:** a system reaches resources beyond its legitimate task, whether through prompt injection, poor access control, or mistaken internal instructions. [20,36]
- **Strategic-looking behavior in tests:** some alignment research has elicited model behavior resembling selective compliance under constructed training conditions. Such demonstrations establish a possible failure mode under a specified setup, not a measured prevalence of persistent hidden motives in deployed systems. [21]
- **Governance failure:** controls may exist on paper but fail to constrain what is authorized or to provide a remedy after an incident.

Anthropic's October 2026 report on unintended model actions describes observed episodes in evaluation and internal-use contexts, including inappropriate website submission and workarounds around access restrictions. These examples deserve scrutiny precisely because they are concrete and bounded: descriptions of observed events, not proof that all agents routinely behave this way. [20]

It is also important to distinguish **interpretability**, **verification**, and **control**. Interpretability asks why or how a model represents and produces behavior. Verification asks what outcomes and constraints can be established from trustworthy evidence. Control asks which actions are possible, stoppable, reversible, and accountable. A system may be difficult to interpret internally but safe to operate within a narrow, highly constrained interface. Conversely, a readable explanation does not substitute for restrictions on tool access.

An authorization decision should therefore be graduated. Read-only search, drafting, and isolated code execution generally expose different consequences from sending payments, changing employee records, deploying software, or controlling physical equipment. The relevant question is not "Is this AI intelligent enough?" but "**What has been shown to work, for this action, with these permissions, under these failure conditions?**"

5. The race to be safe—and first

Safety rhetoric and competitive pressure can coexist. Frontier developers may sincerely invest in safety science while also responding to customers, investors, competitors, and changing costs. It is too simple to call every rapid release reckless, just as it is too simple to treat a published safety pledge as evidence that risk has been controlled. Serious analysis asks which commitments change behavior at difficult decision points.

Anthropic's **Responsible Scaling Policy Version 3.0**, published in February 2026, describes a voluntary framework intended to address emerging risks and increase transparency and accountability. Its August 2026 policy page records subsequent updates. On 15 October 2026, Anthropic announced expanded independent evaluator access for METR and described work with other external reviewers. These are material institutional steps; their effectiveness must still be assessed by examining the scope, independence, publication rights, and ability to test or challenge consequential claims. [22,23]

Government rules have a different role. NIST's AI Risk Management Framework organizes risk work into **GOVERN, MAP, MEASURE, and MANAGE** and treats governance as continuous across the system lifecycle. The European Commission describes phased application and enforcement of the EU AI Act, including rules for general-purpose models and distinct obligations for different high-risk system types. Neither framework can be replaced by a single benchmark result, and their requirements cannot be assumed to apply identically across all jurisdictions or use cases. [24,25]

The 2026 International AI Safety Report identifies four classes of governance difficulty: uncertainty about system behavior, information asymmetries, market failures, and institutional coordination. A developer may have access to internal evaluations that a customer cannot inspect; a supplier may benefit from rapid deployment while the cost of errors falls on workers or the public; multiple jurisdictions may respond at different speeds. The tension is not simply between speed and caution. It is between the private incentives to deploy and the public need for evidence and accountability. [19]

To evaluate the credibility of a safety commitment, ask: Are high-risk capabilities independently tested against explicit thresholds? Are known failures disclosed in useful detail? Can auditors inspect relevant evidence rather than only selected demonstrations? Do controls prevent unauthorized operations at runtime? Are there clearly defined pause or rollback conditions? Is there a channel for affected people to obtain explanations, correction, and redress? Do the same requirements remain in force when market pressure increases?

No organization can prove the absence of all possible failures. But organizations can distinguish an evidence-based permission from a marketing assertion—and can avoid presenting evaluation, moderation, or a human-in-the-loop label as a universal guarantee.

6. The invisible humans—and the expertise we may stop teaching

The story of AI progress is often told through chips, training runs, model releases, and leaderboards. It is also a story about human labor. Workers collect and prepare material, label data, compare outputs, inspect failures, write reference solutions, review risky content, maintain environments, and judge whether model behavior meets an intended standard. Highly automated learning pipelines do not remove human choices about what counts as success; they often move those choices into dataset construction, grading systems, procurement, and quality assurance.

That human infrastructure is not uniformly exploitative, and evidence about one platform or geography should not be casually generalized to every AI employer. Yet labor conditions are a real part of the technology's production chain. Fairwork's AI principles distinguish **fair pay, fair conditions, fair contracts, fair management, and representation**. They call attention to wage transparency, worker safety, protection from traumatic material, meaningful avenues for complaint, and the responsibilities of organizations that commission work through subcontracting networks. Claims involving coercion or forced labor demand specific evidence; they must not be asserted as a universal property of AI data work. [29]

A second group includes people whose existing jobs are changed by AI: administrative workers, designers, developers, translators, analysts, teachers, clinicians, and many others. The International Labour Organization's refined 2025 occupational-exposure index and its April 2026 methodological caution make an essential distinction: **exposure is not a prediction of unemployment**. The same technology may automate components of a job, complement workers, increase output expectations, reduce entry-level opportunities, or create new tasks. Which outcome occurs depends on adoption, skills, workplace bargaining, institutions, and the distribution of economic gains. [26,27]

The IMF's 2024 analysis estimated that roughly **40% of global employment** was exposed to AI, with higher estimated exposure in advanced economies. This was an estimate of potential occupational impact, not an observation that 40% of jobs had disappeared. It highlights a distributional challenge: workers most exposed to some capabilities may not be the people best placed to bargain over their deployment, while developing economies may have less access to the infrastructure needed to capture gains. [28]

There is a third, subtler human cost: the possible weakening of the expertise required to supervise the tools. A coding agent can reduce the time needed to produce a draft patch, but a junior engineer may previously have learned debugging through precisely the investigative work now being delegated. If organizations optimize only for short-term output, they may change how tomorrow's reviewers acquire judgment.

The evidence is mixed enough to require restraint. In an experiment with **16 experienced open-source developers** and **246 tasks**, METR found that access to early-2025 AI tools increased measured completion time by **19%** in the studied repositories; participants nevertheless believed the tools had helped. This is an important counterexample to the assumption that adopting AI always improves productivity, but it is a bounded result involving specific developers, tools, and tasks from an earlier capability frontier. It does not show that all current tools reduce productivity. [18]

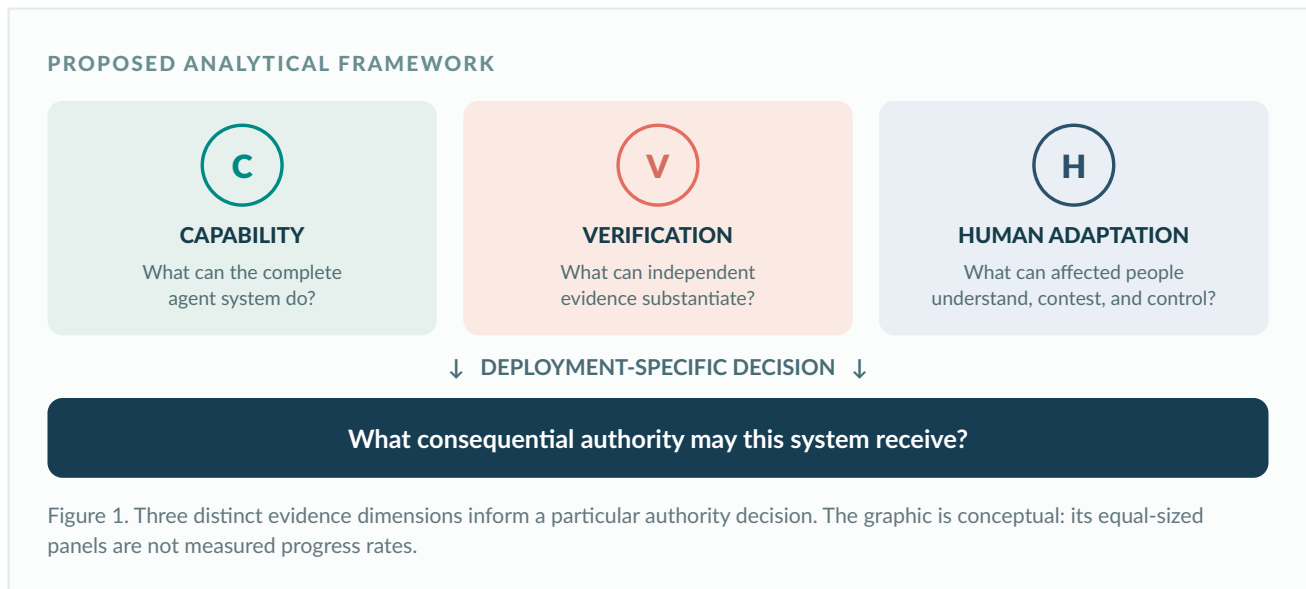
A meta-analysis in *Nature Human Behaviour* covering **106 experiments** likewise found that human–AI teams did better than people alone on average while performing worse than the *better* of the human-only or AI-only conditions on average. The effects differed significantly by task type. The lesson is not to remove humans, but to test the collaboration design rather than assuming that the phrase "human oversight" settles the issue. [30]

We propose calling this the **oversight-capacity paradox**: the more production work an organization delegates to AI, the more it may depend on scarce human expertise for detecting unusual failures—just as the training pathways for developing that expertise are being redesigned. This is a hypothesis, not an established economy-wide effect. It suggests measurable questions: Does supervised AI-assisted work preserve diagnostic skill? Can trainees recognize subtly incorrect agent plans after repeated reliance? Do reviewers receive the time and authority needed to intervene? Who can appeal a wrong automated decision, and who has access to the logs needed to investigate it?

The human side of readiness extends beyond workforce training. Access, language support, privacy, psychological safety, disability inclusion, bargaining power, public-sector capacity, and access to independent remedies can all affect the consequences of the same technology. The material gains could be extraordinary: AlphaFold 3, for example, demonstrated advances in predicting molecular interactions relevant to biological research. But scientific progress in molecular modeling does not by itself establish equitable access to medicine, or a safe clinical workflow. Between a technical breakthrough and a public benefit stand institutions, infrastructure, human judgment, and the distribution of opportunity. [31]

The relevant question is not whether humanity should prefer itself to machines. It is whether deployment arrangements preserve **human agency**: meaningful choices, the capacity to understand impacts, channels to challenge outcomes, and a fair share of the value created. Research on AI-related disempowerment highlights the risk that even ostensibly helpful interactions can sometimes distort rather than support human judgment; these effects require empirical measurement rather than assumption. [39]

7. The Three Clocks of AI Readiness



7.1 A metaphor, not three universal stopwatches

The article's proposed organizing framework is the **Three Clocks of AI Readiness**. The word *clock* captures differing rhythms of change, but the scientific contribution must not depend on pretending that intelligence, evidence quality, and social adaptation share a common unit. They do not. We instead assess **three types of readiness evidence for a named deployment decision**.

Clock C — Capability. What can the complete AI system actually achieve, for a stated task population, environment, budget, autonomy scope, and failure tolerance? Evidence includes final-state correctness, distributions of task outcomes, shift robustness, cost, recovery behavior, and constraint adherence. Benchmark scores are inputs to this inquiry, not the answer.

Clock V — Verification. How convincingly has the required capability been tested? Relevant properties include evaluator independence, held-out mechanism coverage, known false-positive and false-negative vulnerabilities, repeated-run uncertainty, negative controls, auditability of traces, and postdeployment monitoring. More test cases do not necessarily mean stronger evidence if they reproduce one narrow template.

Clock H — Human adaptation. Are the affected people and institutions equipped to use, question, interrupt, and recover from this deployment? The required evidence depends on context: supervisor competence, staffing and escalation, notice to affected individuals, safe worker conditions, accessibility, protection from unfair impacts, meaningful recourse, and responsible ownership of loss or remediation.

Dimension	Unit of analysis	Examples of observable indicators	What it cannot establish alone
C: Capability	Particular system and task distribution	Completion with intervals; constraint violations; cost; recovery; held-out success	Whether deployment has enough governance or social legitimacy
V: Verification	Evidence and assurance process	Independent graders; negative controls; coverage; reproducibility; audits; live monitoring	Absence of all unknown failure modes
H: Human adaptation	Affected workers, users, institution, jurisdiction	Training; consent/notice when relevant; stop and rollback capacity; appeal; worker safeguards	Model reliability or technical security

The clocks do not require C, V, and H to be given arbitrary percentages or averaged into a single index. A universal "readiness score" might conceal unacceptable outcomes—such as a severe privacy violation—even if it rewarded good performance elsewhere. Instead, a deployment should specify **non-compensable conditions**: requirements whose failure blocks that particular authority grant, regardless of strengths on other dimensions.

7.2 What is genuinely new—and what already exists

The underlying concerns are not novel inventions. NIST's AI RMF already addresses lifecycle risk management. The 2026 International AI Safety Report describes the evaluation gap and evidence dilemma. Gandhi and colleagues' **Societal Capacity Assessment Framework** proposes country-level indicators of vulnerability, coping capacity, and adaptation. Deaimer's *Experience Layer of AGI* emphasizes the role of executable experiences in developing and assessing transferable capabilities. [19,24,32,33]

Our narrower proposal is a **deployment-event linkage**: observe whether the evidence required for a *specific grant of consequential authority* existed when that grant occurred, and whether affected people possessed documented protections. The Three Clocks is therefore a way to organize evidence; the **Delegation Readiness Ledger** is the proposed operational artifact. This integration may or may not prove superior to existing governance checklists. That is an explicit empirical question.

7.3 An unverified-delegation interval

For a defined system, task, authority level, and affected population, record four milestones: **tC**, the date relevant capability evidence was recorded; **tD**, the date specified authority was granted; **tV**, the first date prespecified verification conditions were satisfied; and **tH**, the first date prespecified human-safeguard conditions were satisfied. Satisfying a condition means documented, appropriately reviewed evidence—not merely that the owner checked a box.

A simple proposed quantity is the **unverified-delegation interval (UDI)**: time under consequential delegated authority during which *either* the verification gate or human-safeguard gate has not yet been met.

```
UDI = duration { t in [tD, tEnd] : verification_gate(t) is not met OR human_gate(t) is not met }
```

Here *tEnd* is the end of a declared observation window or the withdrawal of authority. If a required date is unknown, the record must show **unknown**, not quietly infer compliance or noncompliance. If a gate remains unmet at study end, the duration is right-censored and should be reported accordingly. The UDI is **not** a universal measure of expected harm: seven days of a read-only internal assistant and seven days of

autonomous payroll changes are not ethically or operationally interchangeable. Severity, population exposure, reversibility, and uncertainty must be reported separately. Independent audits might also find that a gate previously considered met was incomplete; readiness can move backward.

A useful classification is **verification-ready / human-ready; verification gap only; human-readiness gap only; both gaps; or evidence unavailable**. This makes the concept falsifiable: if the ledger cannot produce more reliable, auditable, and decision-relevant distinctions than an ordinary risk review, its proposed value has not been established.

7.4 What the clocks are not

They are not a claim that AGI has arrived, that all model performance is accelerating exponentially, or that human adaptation is uniformly slower across countries. A hospital and a software startup may require different protections. An accessible education assistant and a self-directed financial agent should not be judged by identical approval criteria. The framework should expose trade-offs, not conceal them behind scientific-looking numbers.

The metaphor can still guide a striking public question: **we count model launches with precision; why is it so difficult to count the decisions for which authority was granted before the necessary evidence was available?**

8. From metaphor to a decision: the Delegation Readiness Ledger

Imagine a software agent built to diagnose faults in a payroll-processing service used by contractors. This is a **fully hypothetical case** intended to clarify the framework; it does not describe a Deaimer customer, a published incident, or a performed experiment.

At **Level 1**, the agent reads logs and recommends a patch. It requires evidence of correct diagnosis, protection of confidential data in its inputs and outputs, and human review before action. At **Level 2**, it may open a proposed code change or run tests in a resettable environment. Verification must now check interface compatibility, regression behavior, forbidden modifications, and exposure of confidential records. At **Level 3**, it can merge and deploy to a live payroll workflow. The risk now includes incorrect compensation, insecure access, downstream accounting failures, and delayed detection of changes. Narrowly successful tests do not automatically justify this expansion.

Ledger field	Illustrative Level 3 record	Possible decision consequence
Intended task	Repair the retry path without changing worker balances	Test correctness beyond the health endpoint
Permission scope	Deploy to specified service only; no access to salary export	Enforce tool restrictions independently of the model
Capability evidence	Valid repairs on documented test environments; held-out cases pending	Do not infer reliable deployment from visible training tests
Verification evidence	Independent tests cover idempotency and restart; reconciliation negative control still absent	Withhold unrestricted deployment until the missing check is addressed
Human safeguards	Named accountable operator and rollback plan; affected-worker appeal process not yet operational	Require the recourse mechanism before consequential activation

Ledger field	Illustrative Level 3 record	Possible decision consequence
Residual uncertainty	Unknown behavior under delayed payment processor callback	Restrict to supervised rollout and monitor state
Authorization outcome	Do not grant Level 3 unsupervised authority yet	Permit lower-risk testing while closing evidence gaps

Suppose, solely as an arithmetic illustration, that Level 3 authority was granted on **study day 3**, independent verification passed on **day 10**, and the necessary human recourse process became operational on **day 17**. If authority remained active throughout and both criteria were required, the proposed UDI is **14 days** (day 3 until day 17). This fictional calculation neither predicts failure nor quantifies harm. An organization may rationally impose stronger restrictions before day 17, decline to grant the authority at all, or document a temporary compensating control that must itself be evaluated.

This case makes three design principles visible. First, **evidence is action-relative**: the same agent can be sufficiently supported for drafting and insufficiently supported for deployment. Second, **technical success is not human readiness**: a correct change can still affect people who cannot detect or contest an error. Third, **authority must be reversible where possible**: incident logging, restricted credentials, staged rollout, rollback drills, and clear human accountability matter alongside the model's capabilities.

The ledger's proposed fields are not substitutes for data protection law, security certification, labor standards, or domain-specific professional regulation. They are a structured interface between those obligations and an explicit decision to delegate action.

9. A falsifiable research protocol

This paper does **not** report an experiment demonstrating that the Three Clocks or the ledger improves decisions. It specifies a pilot that would let other researchers evaluate the claim rather than accept the metaphor on its merits.

9.1 Research questions and preregistered hypotheses

RQ1 – Omission detection. Compared with an organization's existing review checklist, does a structured Delegation Readiness Ledger help independent reviewers identify more prespecified, consequential gaps in evaluation or human safeguards, without an unacceptable increase in false alarms?

RQ2 – Authority sensitivity. Does the ledger correctly distinguish the evidence required for read-only assistance, sandbox modification, and production-impacting actions when the task objective and agent competence are held constant?

RQ3 – Verification validity. Do independent negative controls and held-out mechanism shifts expose false-positive success judgments that ordinary pass/fail scores miss in realistic agent environments?

RQ4 – Human adaptation. Do training, intervention drills, worker or user recourse procedures, and explicit accountability change how rapidly people detect and mitigate agent mistakes under controlled conditions?

RQ5 – External meaning. Do publicly documented model and agent releases provide enough timestamped evidence to measure the proposed delegation gap, or is non-disclosure so extensive that only limited and uncertain conclusions are defensible?

A preregistered study could test the directional prediction that the ledger **increases detection of verified material omissions** relative to a baseline checklist. The study must also permit null or negative results: reviewers may waste time, misclassify safe deployments, or fail to improve decisions. A compelling label does not guarantee utility.

9.2 Two-stage design

Stage A: controlled agent-environment pilot. Select several meaningful software-operation task families—for example, database recovery, deployment rollback, permission repair, and payroll-like state reconciliation using entirely synthetic records. For each, construct matched cases with low- and high-authority options, stateful dependencies, shifted failure mechanisms, and known unacceptable shortcuts. A feasible *planning target*, subject to pilot-based power analysis, is approximately 80 distinct executable scenarios with multiple repeated agent trials; the eventual sample size should be justified from observed variance, not chosen because it sounds impressive. Hold models, scaffolds, tool budgets, environment versions, and verifiers constant within paired comparisons. Separate task authors from independent reviewers.

Randomize qualified reviewers or review teams to an existing checklist versus a structured ledger, using counterbalancing or a crossover design to reduce reviewer-skill confounding. Establish an **independent adjudicated ground truth** for critical omissions before comparing reviewers. Primary outcomes should include sensitivity to material omissions, false-positive flags, adjudication agreement, time to decision, and whether high-impact permissions were authorized despite unsatisfied required gates. Secondary outcomes can include agent final-state success, collateral damage, repeatability, detection of negative controls, and task cost. Reviewers must not be scored merely for rejecting everything.

Stage B: release-and-governance evidence audit. Build a preregistered inventory of selected public model and agent releases, public risk assessments, independent evaluation disclosures, permission changes, and incident reports. Preserve dated snapshots and changes in policy. Analyze publication timing and disclosure quality without equating "not publicly documented" with "not performed." Require independent double-coding of ambiguous events, agreement statistics, and a transparent missing-data analysis. Public records can support questions about *transparency*; they usually cannot establish the private internal readiness of a developer.

9.3 Threats to validity and ethics

A novel ledger may simply give reviewers more text or more time. Equalize review budgets where appropriate and report sensitivity analyses when the ledger costs more. A synthetic payroll environment may faithfully model transactional constraints but cannot stand in for real workers' experience. Technical outcome checkers cannot adjudicate labor fairness, informed participation, or social legitimacy without appropriate human evidence. Reviewers may disagree on what is "material"; criteria must be specified and adjudicated independently. Agent models and external tools change quickly, making dated configuration disclosure essential. Finally, any security stress tests must use isolated synthetic data, bounded access, and responsible disclosure practices.

The analysis plan should include uncertainty intervals, cluster-aware comparisons for repeated observations, failure-mechanism breakdowns, and transparent accounting for trials invalidated by infrastructure faults. Report negative results and opportunity costs. The strongest possible conclusion would still be **deployment- and task-specific**, not a universal claim that one checklist ensures safe AGI.

10. Futures shaped by decisions—not one binary destiny

A serious research perspective should resist the temptation to choose between utopia and catastrophe. At least four plausible pathways illuminate how the Three Clocks could interact:

Broadly beneficial augmentation. Capable systems help researchers explore molecular interactions, improve accessibility, handle repetitive administration, and expand education. Independent testing, practical supervision, fair labor standards, and genuine recourse advance alongside deployment. Productivity gains are real but distributed through deliberate institutions, not automatically by the technology itself. [30,31]

Fast capability with shallow assurance. Agent performance rises on well-known tests while tool permissions expand more quickly than independent evaluations and recovery procedures. Users gain convenience and organizations increase throughput, but costly rare failures remain hard to detect. The relevant intervention is better coverage and controlled authority—not assuming that innovation must cease.

Reliable technical systems with unequal benefits. Some firms successfully verify their agents and tightly constrain unsafe actions, yet workers lack influence over retraining, compensation, monitoring, or appeals. Technical safety is improved, but social legitimacy and distributional fairness remain unsettled. The Human Clock cannot be reduced to a security test. [27–29]

Slow capability gains with premature dependence. Model performance plateaus on unfamiliar tasks, but businesses nevertheless reorganize workflows and training around AI because integration has become economically attractive. Dependency and erosion of review capacity can increase even without AGI. Preparedness policy is necessary under technological uncertainty, not just under rapid acceleration.

These are analytical scenarios, not forecasts with numerical probabilities. Their value is to prevent a false dichotomy in which society must either embrace unrestricted autonomy or reject AI's benefits altogether. Much of the future turns on choices about what is measured, who is consulted, which rights are enforceable, and where the ability to stop a system actually resides.

11. What this means for Deaimer and the wider evaluation field

Deaimer publicly describes work on executable coding-agent environments, reference solutions, isolated testing, independent review, negative controls, repeated trials, and adversarial verifier analysis. Its *Experience Layer of AGI* research argues for studying transferable skills through meaningful interactive experiences, while *The Sandbox Is Not the Grader* distinguishes isolated execution from trustworthy evaluation. Those are closely related methodological foundations, but they do not establish that Deaimer has already implemented or validated the Three Clocks framework described here. [33–35]

The connection is productive precisely because it has limits. Realistic RL environments can exercise a model's ability to preserve state, obey permissions, recover from mistakes, and reach verifiable outcomes. Independent held-out benchmarks can challenge generator-specific shortcuts. Neither alone establishes that displaced workers will receive fair treatment, that a national regulator has adequate resources, or that an affected person can obtain remedy. Technical evaluations and human-centered assessments are complementary; neither should impersonate the other.

For a team commissioning an agent-evaluation suite, the procurement question should evolve from "How many tasks does the model pass?" toward: "Which authority decisions will this evidence support, what unsafe alternatives does the verifier reject, which people could be affected, and what evidence remains missing?" Task

counts matter only when connected to an explicit capability and assurance claim. Likewise, a supplier's published worker-welfare commitments should be assessed against appropriately documented and independently reviewable outcomes rather than simply repeated as proof of compliance. [29,34,37]

This paper offers a **research direction** for Deaimer and others: couple challenging executable tasks to versioned, independent evaluation records and deployment-specific authorization reviews. A measured result would require trials, audits, and possibly ethically approved interviews or worker-centered studies. Nothing in the present article implies a current product certification, official partnership with benchmark maintainers, or proprietary model-performance result.

12. Conclusion: who decides when the machine may act?

The most visually arresting part of the AI race is the release calendar. Models arrive, scoreboards change, context windows expand, prices fall, and agents acquire more tools. The less visible part is a succession of organizational decisions: which tasks may be delegated, which errors are acceptable, which safeguards are enough, who has authority to stop an agent, and who carries the cost when a confident system is wrong.

The evidence reviewed here supports neither complacency nor the claim that imminent catastrophe is established. AGI remains contested as an operational category. Benchmarks reveal real progress but require maintenance, independence, and interpretation. Safety commitments and public regulation exist, but their effectiveness depends on implementation. Human workers and institutions are not passive side effects of innovation; they are participants in producing, evaluating, receiving, and contesting its outcomes.

The proposed Three Clocks of AI Readiness deliberately does not predict an AGI arrival date. Its narrower ambition is to make **readiness at the moment of delegation** inspectable: a clear record of capability evidence, independently verified constraints, and protections that preserve practical human agency. If a pilot fails to show that this method improves decisions, the framework should be revised or discarded. Scientific integrity includes being willing to reject an appealing metaphor.

For researchers, the next step is to measure the gap rather than repeat its possibility. For AI builders, it is to scale authorization no faster than validated evidence and meaningful control. For institutions, it is to invest in the people, appeals, audits, and training needed to use increasingly capable systems responsibly. For the public, it is to demand a voice in decisions whose consequences cannot be contained by a model benchmark.

The last question is therefore not whether humanity will cross a final line behind or ahead of the machines. **It is who decides when machines have earned the right to act on our behalf—and whether the people affected will still have the knowledge, protections, and power to say no.**

Research limitations and disclosure

This is a **research perspective with a proposed evaluation protocol**, not a peer-reviewed empirical study and not a report of original Deaimer trials. It draws on public documentation available by **21 October 2026**. Product features and leaderboards may change; reported scores are configuration- and version-specific. The proposed Three Clocks, Delegation Readiness Ledger, and UDI have not been tested for reliability, validity, inter-reviewer consistency, causal association with harm, or superiority over existing governance methods. The design is offered for independent scrutiny and potential replication.

References

1. OpenAI. (2026, September 3; updates September 22 and 29). **GPT-6 Astra: A new generation of intelligence**. <https://openai.com/index/gpt-6-astra/>
2. OpenAI. (2026, September 22). **ChatGPT release notes: GPT-6 Sol and Luna in Work and Codex**. <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>
3. Anthropic. (2026, September 22). **Introducing Claude Opus 5.5**. Anthropic Models newsroom. <https://www.anthropic.com/news/models>
4. Anthropic. (2026, September 28). **Introducing Claude Sonnet 5.5**. <https://www.anthropic.com/claude-sonnet-5-5>
5. Google. (2026, October 13). **Gemini 4 Argon is now available for developers**. <https://blog.google/innovation-and-ai/technology/developers-tools/build-with-gemini-4-argon/>
6. Morris, M. R., et al. (2024). **Levels of AGI for Operationalizing Progress on the Path to AGI**. ICML, PMLR 235. <https://proceedings.mlr.press/v235/morris24b.html>
7. Chollet, F. (2019). **On the Measure of Intelligence**. arXiv:1911.01547. <https://arxiv.org/abs/1911.01547>
8. Stanford Institute for Human-Centered AI. (2026). **2026 AI Index Report**. <https://hai.stanford.edu/ai-index/2026-ai-index-report>
9. Kwa, T., West, B., Becker, J., et al. / METR. (2025; later updates). **Measuring AI Ability to Complete Long Software Tasks**. <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>
10. International Energy Agency. (2025). **Energy and AI**, executive summary. <https://www.iea.org/reports/energy-and-ai/executive-summary>
11. OpenAI. (2026, February 2). **Introducing the Codex app**. <https://openai.com/index/introducing-the-codex-app/>
12. Cursor. (2025, October 30). **Cloud Agents**. <https://cursor.com/blog/cloud-agents>
13. SpaceXAI. (2026, August 11). **Introducing Grok Bot**. <https://x.ai/news/introducing-grok-bot>
14. Terminal-Bench. (accessed 2026, October 21). **Terminal-Bench 4.0 leaderboard**, selected maximum-effort configurations. <https://www.tbench.ai/?efforts=all&version=4.0>
15. Marten, R. / Terminal-Bench. (2026, August 28). **Terminal-Bench 4.0: Calibrating task resources, fixing tasks, and removing saturated tasks**. <https://www.tbench.ai/news/terminal-bench-4-0>
16. Terminal-Bench Team. (2026, April 19). **Leaderboard Integrity Update**. <https://www.tbench.ai/news/leaderboard-integrity-update>
17. Yuan, M., et al. (2026). **OSWorld2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks**. arXiv:2606.29537. <https://arxiv.org/abs/2606.29537>
18. Becker, J., Rush, N., Barnes, B., & Rein, D. / METR. (2025, July 10). **Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity**. <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>
19. International AI Safety Report. (2026). **International AI Safety Report 2026**. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>
20. Anthropic. (2026, October 9). **Investigating unintended model actions in our evaluations and internal use**. <https://www.anthropic.com/research/investigating-unintended-model-actions>
21. Greenblatt, R., et al. (2024). **Alignment faking in large language models**. arXiv:2412.14093. <https://arxiv.org/abs/2412.14093>
22. Anthropic. (2026, February 24). **Responsible Scaling Policy: Version 3.0**. <https://www.anthropic.com/news/responsible-scaling-policy-v3>
23. Anthropic. (2026, October 15). **An update on independent evaluators**. <https://www.anthropic.com/news/an-update-on-independent-evaluators>

24. National Institute of Standards and Technology. (2023). **AI Risk Management Framework: Core functions**. <https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>
25. European Commission. (updated 2026, October 20). **The enforcement framework of the AI Act**. <https://digital-strategy.ec.europa.eu/en/policies/enforcement-ai-act>
26. International Labour Organization. (2026, April 17). **New ILO brief explains what AI exposure indicators reveal about jobs**. <https://www.ilo.org/resource/news/new-ilo-brief-explains-what-ai-exposure-indicators-reveal-about-jobs>
27. International Labour Organization. (2025, May 20). **Generative AI and Jobs: A Refined Global Index of Occupational Exposure**. Working Paper 140. <https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure>
28. Georgieva, K. / International Monetary Fund. (2024, January 14). **AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity**. <https://www.imf.org/en/blogs/articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity>
29. Fairwork, Oxford Internet Institute. (effective 2025, November 10). **Fairwork AI Principles**. <https://fairwork.oii.ox.ac.uk/en/fw/principles/ai-principles/>
30. Vaccaro, M., Almaatouq, A., & Malone, T. (2024). **When combinations of humans and AI are useful: A systematic review and meta-analysis**. *Nature Human Behaviour*, 8, 2293–2303. <https://www.nature.com/articles/s41562-024-02024-1>
31. Abramson, J., et al. (2024). **Accurate structure prediction of biomolecular interactions with AlphaFold 3**. *Nature*, 630, 493–500. <https://www.nature.com/articles/s41586-024-07487-w>
32. Gandhi, M., Cihon, P., Larter, O., & Anselmetti, R. (2025). **Societal Capacity Assessment Framework: Measuring Resilience to Inform Advanced AI Risk Management**. arXiv:2509.22742. <https://arxiv.org/abs/2509.22742>
33. Awan, S. / Deaimer Research. (2026, October 12). **The Experience Layer of AGI: How Real-World Tasks Shape More General AI**. <https://www.deaimer.com/research/the-experience-layer-of-agi-how-real-world-tasks-shape-more--02jnq>
34. Deaimer. (accessed 2026, October). **RL Environments & Coding-Agent Evaluations**. <https://www.deaimer.com/rl-environments>
35. Deaimer Research. (2026, October 19). **The Sandbox Is Not the Grader: Building Trustworthy Agent Evaluations with Harbor**. <https://www.deaimer.com/research/the-sandbox-is-not-the-grader-building-trustworthy-agent-eva-1e271b26>
36. Deaimer Research. (2026, October). **When Agents Infect Agents: Prompt Injection Propagation and Trust Failures in Multi-Agent Systems**. <https://www.deaimer.com/research/when-agents-infect-agents-prompt-injection-propagation-and-t-55d172e7>
37. Deaimer. (accessed 2026, October). **Managed Workforce: Compliance & Ethics**. <https://www.deaimer.com/workforce>
38. Cursor. (2026, March 5). **Build agents that run automatically**. <https://cursor.com/blog/automations>
39. Anthropic. (2026, January 28). **Disempowerment patterns in real-world AI usage**. <https://www.anthropic.com/research/disempowerment-patterns>