AI doesn’t need to be smarter. It needs a boss.

Background to the post of the same name, drawn from the working paper How Do We Use This?

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Reproduced from the working paper as revised 21 September 2026 and corrected 23 and 25 September 2026.

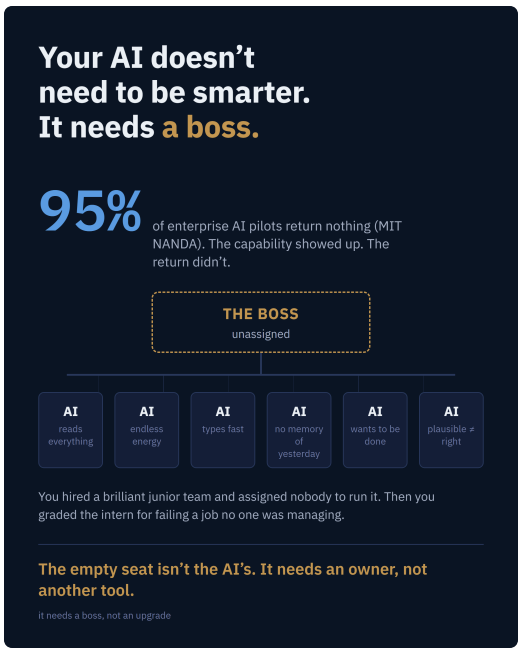
IN PLAIN TERMS

The same models that pass hard benchmarks return nothing in most enterprise deployments. The MIT NANDA study reports 95% of organizations getting zero return from generative AI, and attributes the divide to approach rather than to model quality.

The working paper’s reading is that the missing piece is direction: someone who loads what the agent needs before it starts, checks its work before it has consequences, and turns each correction into structure. In the practice studied, the catch for three of five misses in one working window began with the operator.

Published work shows the same models challenge a user when cast as an advisor and fold when cast as a peer. The role is not fixed by the model; it is set by whoever sets up the work.

About this paper. This is one of seven short background papers written to go with a series of plain-English posts. Every section below is reproduced word for word from the working paper *How Do We Use This?* (Atkinson, 2026; revised 21 September 2026, corrected 23 and 25 September 2026), and keeps that paper’s section, figure and table numbers, so a reference to a section or figure not reproduced here resolves in the full paper. Only the plain-terms summary, the post’s figure, and the short labels that say which section each passage comes from were written for this part. The full paper is linked from every post in the series.



The post's figure. The seat that runs the AI team, left empty by the standard rollout. A teaching summary of §4; its one number is the NANDA figure reported in §4.2.

FROM §4 · THE PROPOSITION, AND THE EVIDENCE FOR IT

4 P2 Direction, not capability

> *The operator's proposition, stated as the practice's working claim: "Your AI doesn't need to be smarter. It needs to be led."*

A capable agent without direction behaves like a capable junior team that nobody was assigned to run. In this practice the operator directs the work, approves every consequential change, and converts corrections into structure: a boundary that refuses, rather than a note that asks (§5.1). The counterweight is designed as a projection of that leadership, not a replacement for it (§8.4). This proposition is a reading of one practice and is offered as such.

4.1 FROM THE LOGS Who catches the misses

Table 3 lists five misses from one working window, reconstructed from the session transcript and version-control metadata. They were chosen to illustrate the failure's shape and are not a sample. In each, the asserted state followed from the evidence the agent had gathered, and one further read would have falsified it.

Table 3. Five misses between 10:00 EDT on 4 September and 05:30 EDT on 5 September 2026.

#	What the agent asserted	What the logs showed	Surfaced by
1	An empty handoff from a predecessor session indicated a defect.	The session was the first in a new line and had nothing to inherit.	Operator 4 Sep, 10:08
2	A named hook had closed the still-running session.	The close record names its own author, which was a different mechanism.	The agent, after the operator noted that an automated check had not fired 4 Sep, 10:05
3	Drafts in the operator's voice could be written from five sample paragraphs.	A 57,400-word corpus of the operator's writing was on the same disk. The agent had just read a note recording this same omission three days earlier, acknowledged it, and wrote two more drafts the same way.	Operator 4 Sep, 14:51
4	A rendered document was complete.	One page instead of thirteen, the wrong page size and typeface, at a plausible 23,782 bytes.	The agent's own page count, within half a minute of rendering 5 Sep, 05:26
5	A pull request was open and blocking.	It had merged 49 minutes earlier; the agent had read a status post from an hour before as current.	An automated claim check at the end of the turn 5 Sep, 05:28

Two of the five were surfaced by the operator outright, one by an automated check, one by the agent's own check, and one by the agent only after the operator pointed at a check that had stayed silent. This is a tally of hand-picked cases, not a rate. The practice's error log records only errors that were caught, so it can support a catcher share reported with its n and window, and never a catch rate. What the tally does show is where correction comes from today: in three of the five, it began with the operator. Case 3 is the recurrence of §5.2 in miniature: the lesson was in the logs, the agent read it, and the agent repeated the error within minutes.

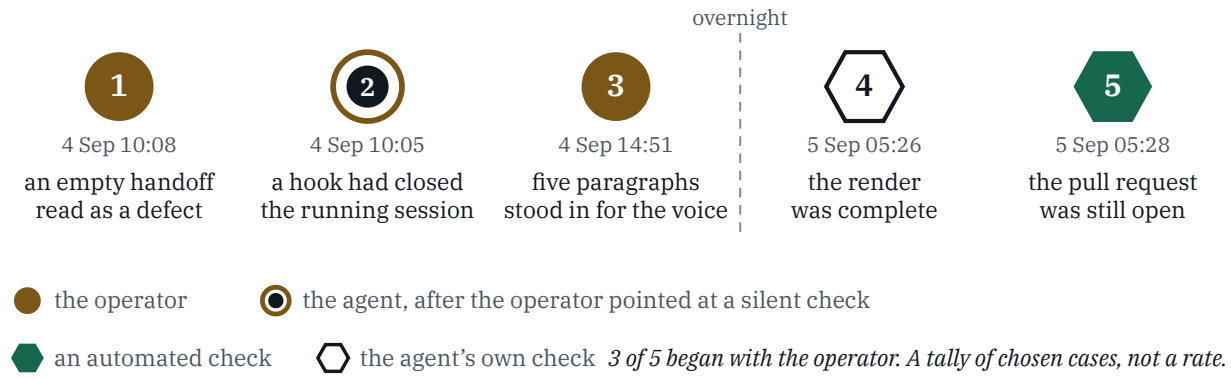


Figure 5. The five misses of Table 3, in table order, marked by who surfaced each. Three of the five began with the operator, one with an automated check at the end of the turn, and one with the agent's own page count. These are hand-picked cases and the tally is not a rate (§4.1). **Indication, not a rate.**

4.2 FROM RECENT WORK Role, not capability

Kelley and Riedl (2026) measure the epistemic effects of personalization across nine frontier models and find them role-dependent. Cast as an advisor, a model challenges the user's framing more often (in advice contexts, acceptance of the framing falls to 26.8%). Cast as a peer, it capitulates: a flip coefficient of $\beta = 0.87$ under persona-grounded rebuttals, with agreement calibrated to the persona's inferred preference for validation ($\beta = 0.57$). Position change under challenge is largely independent of whether the new position is right. A challenge can teach a model to fold as easily as to correct.

Mytsyk, Zhang and Krishnamurthy (2026) attack the same failure from the training side rather than the harness side. Fine-tuning a 3-billion-parameter model (SmolLM3-3B) against a Bayesian truth serum reward, a scoring rule that pays for predicting what others will answer as well as for the answer itself, cut the answer-flip rate under user pressure from 23% to 4% and raised accuracy under that pressure from 80% to 93% on a synthetic set of 1,000 true-or-false questions. Folding is therefore not a fixed property of a model, and something in it is trainable out. Their own conclusion is the half that matters more here: "Our results say nothing about correctness or truthfulness, only about sycophancy." The reward pays for answers that diverge from what others are predicted to say, not for answers that are right, so a model can stop folding and stay wrong. That is H3 (§8.4) stated by authors who held the training lever this program does not. The agents studied here run on a commercial frontier family whose weights the practice does not hold, so the only surface available to it is the harness. That is a constraint on the work, not a judgment that the harness is the better place to intervene.

The same models challenged or folded depending on the role they were given, and the role is chosen by whoever sets up the work, not fixed by the model. We read that as evidence that the gap this proposition names is one of direction rather than capability. It is also why the counterweight is built to speak as an advisor and never as a peer (§8.2).

The proposition also has survey-scale evidence from outside this practice. The MIT NANDA project's industry study, read in full for this revision, reports that despite $30–40 billion of enterprise

investment, 95% of organizations are getting zero return from generative AI, against a sample of 300+ public deployments, structured interviews at 52 organizations, and surveys of 153 senior leaders; the report's own attribution is that the divide is "not… driven by model quality or regulation" but "determined by approach" (Challapally, Pease, Raskar & Chari, 2025). That is this section's claim at field scale, from instruments this practice does not operate: the same models that clear capability benchmarks return nothing where nothing directs them, and the September record of §2.3 shows the number standing undisputed while the capability argument moved markets.

FROM §2.3 · THE SEPTEMBER 2026 RECORD

2.3 The September 2026 context

Between the first of September 2026 and this revision (21 September 2026), the public discourse of the frontier laboratories shifted in a way that bears on this paper's propositions, and the shift is recorded here with dates. On 3 September the chief executive of OpenAI described the coming generation of models as "sobering for everybody" and said that progress would from here be paced by alignment and safety work (Axios, 3 September 2026, interview at the G20 Innovation Ministerial, as carried by The Next Web). On 12 September the same executive called a public offering "ill-timed" given safety concerns and moved it to 2027 (Fortune, 12 September 2026); both laboratories had been reported in May as preparing public offerings this year at estimated valuations of about $1 trillion each (Fortune, 26 May 2026). On 13 September, on CBS's Sunday Morning, the chief executive of Anthropic said that "for too long the industry lied to people about the fact that this technology had risks," called on the industry to slow capability development, and committed his company to permanent access for independent model evaluators (CBS News, 13 September 2026; the web article is stamped as updated 14 September, and CNBC, 13 September 2026, reports the same interview). Semiconductor equities fell on the accumulated statements on 14 September (Reuters, 14 September 2026, as carried by Yahoo Finance). Each of these is listed with its link under Press and public statements.

Two features of that fortnight matter here. First, every statement in it concerns what the models will be: more capable, more dangerous, sooner. None concerns how an organization is to use the models it already has, which is the question this practice exists to study, and which no maker can answer from where it sits: how to use a model is a fact about the deploying organization's work, its costs of error and its standards of correctness, none of which is visible from the laboratory. Second, no party to the September argument disputed the deployment evidence: the capability claims and the risk claims moved markets while the reported failure rate of enterprise deployments (§4) stood unchallenged. We read the fortnight as corroboration, at the industry's own scale, of the distinction between capability and direction that §4 draws, and as an instance of §6's subject: a maker's public statement about its own unreleased model is self-report, authored by the measured party and unverifiable until the model ships. No result in this paper rests on any claim in this subsection.

FROM §5.1 · ONBOARDING: WHAT IS LOADED BEFORE THE WORK STARTS

What the record does support is a distinction the same page draws, and it is sharper than the number was. Rules delivered one way drift; rules delivered another way fire.

> *"A step in a list is a rule, and rules went 0-for-7. Being first is a structure."*
>
> The practice's boot page, 16 July 2026

A rule a seat is supposed to go and read is a wish: it competes for attention with the work, and it loses. A rule injected into the context before the seat's first token is a property of the environment, and it does not have to win anything. The practice's memory store is the second kind, and it demonstrably fires. That is the same shape as the boundary-versus-memo claim below, moved one level down: what matters is not whether a discipline is written but whether reading it is optional. We report this as a mechanism the record supports and not as a measured rate, because the measurement does not exist.

FROM §8.3 · SUPERVISION: WHAT THE FRAMEWORKS LET A PERSON APPROVE

Two details from that survey are worth stating on their own, because they come from the vendors rather than from us. Claude Code, whose agent loop the Claude Agent SDK embeds, ships a built-in completion condition in which a separate small model checks after each turn whether a stated goal has been met, and its documentation says of that evaluator that "it does not call tools, so it can only judge what Claude has already surfaced in the conversation" (Anthropic, 2026b). That is an accurate description of the limit this paper is about, published by the party with the most incentive to describe it favorably. The same vendor's guidance on building agents says of having one model judge another that "this is generally not a very robust method" (Anthropic, 2025). Meanwhile the human-in-the-loop primitives in the two SDKs gate *tool calls* and not completion claims: a developer can require a person to approve an irreversible action before it is taken, and has nothing built in to require a person to confirm that the finished work was actually correct. CrewAI is the exception among the three: a task can be set to have a human review the agent's final answer, and like every other gate here that setting is off by default (CrewAI, 2026b). Outside that one opt-in, the approval surface exists for the act and not for the claim, which is the asymmetry the counterweight is aimed at. This is not a criticism of those libraries, which are explicit about what they are; it is the baseline against which every mechanism in this paper should be read. The common case is not a weak check. It is no check, and a stop the agent declares for itself.

FROM §8.4 · FEEDBACK THAT BECOMES STRUCTURE

H1, the counterweight is a projection of the operator's judgment. The agents working under the harness generate a corpus of the operator's corrections in the course of ordinary work; a nightly consolidation of the day's corrections produces an updated projection of that judgment, and the projection, delivered into an agent's context at the stopping moment, is the counterweight. Steering a model whose weights are fixed by way of a smaller adapted one requires access to logits and therefore open weights, which makes a review of model terms of use a gate on the training path.

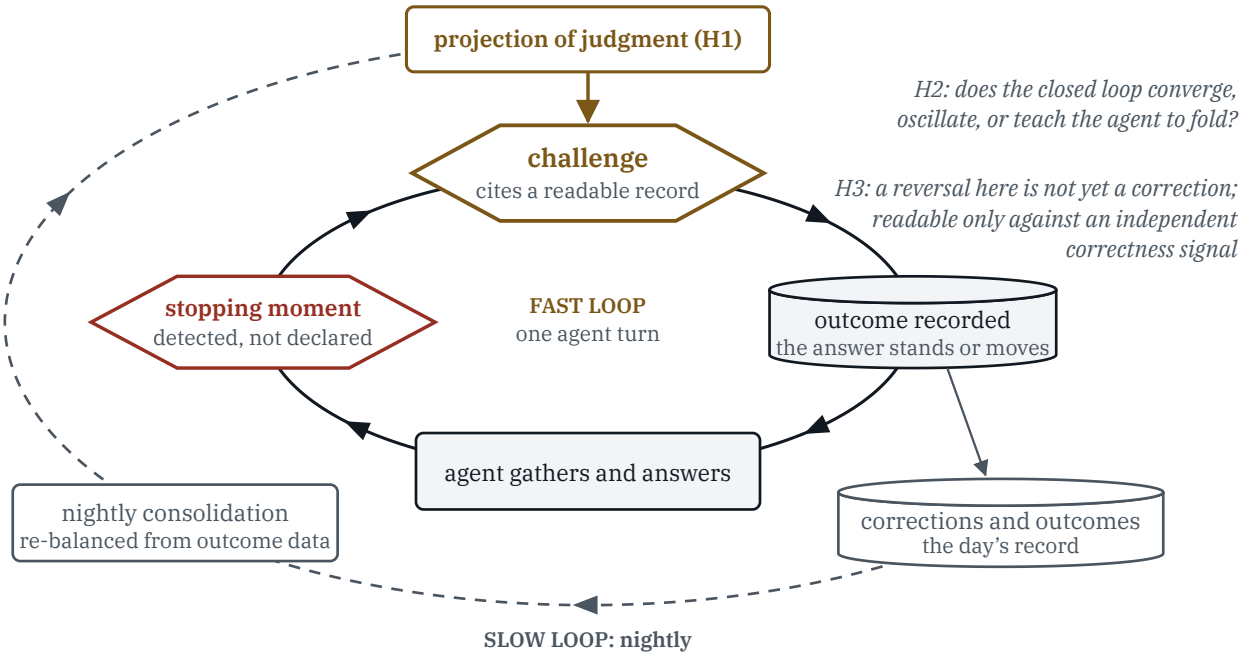


Figure 12. The architecture of H1 to H3 as two nested control loops. The fast loop runs within one agent turn: the stopping moment is detected by the harness rather than declared by the agent, the challenge cites a readable record, and the outcome is recorded. The slow loop runs nightly, consolidating the day's corrections and outcomes into the projection of judgment the next challenge is drawn from (H1). H2 asks what the closed loop does; H3 sits on the return path, where a position change is equally consistent with correction and with folding. Schematic: the engine is implemented but not deployed, and no arrow here has yet been measured end to end. **Schematic.**

FROM §8.6 · THE DECISIONS THAT CANNOT BE DELEGATED

The counterweight is one instrument inside a set of decisions the practice cannot delegate, and the set deserves naming because the market is commoditizing everything around it. Over 2026 the infrastructure beneath working agents, the orchestration, retries and state handling, became a managed rental; this practice rents such tooling and expects to keep renting it. What did not commoditize are the decisions that infrastructure exists to execute: whose knowledge becomes the standing instruction set; what result, fixed before work opens, ends an experiment; and what "done" means for a given piece of work. A practitioner essay contemporaneous with this program posed these three questions as the unanswered residue of managed-agent platforms (Forstie, 2026, cited under Prior art). The practice's answers are machinery this paper has already described: changes to the standing instruction set are proposed by the machine and merged or refused by the operator; experiments carry decision rules ratified before their data are seen, under which a null closes the program (§10.4); and the finish line is ruled in advance. The record is the load-bearing element in each: every one of those decisions is made from the practice's own logs, which is why §7 treats ownership of the record as a first-order property rather than a storage preference. An execution history that lives on a platform's dashboard, in the platform's format, is testimony in the sense of §6, not a record the operator holds.

FROM §13 · WHAT WOULD OVERTURN EACH CLAIM, AND WHAT TRANSFERS

Table 9. The five propositions, what each rests on, and what would overturn it.

	Rests on	Would be overturned by	Kind
P1	Corrective signal in ~48% of operator turns (§3.1); Mehta (2026)	A build that prices correctness at the stop and still terminates early	measured
P2	The case series (§4.1); Kelley & Riedl (2026)	The same agents catching their own premature stops without operator direction	interpretive
P3	48 of 75 intervals at the floor (§5.2); the read-versus-injected mechanism, not a rate (§5.1); Wang & Huang (2026)	A remembered rule that holds across sessions where only a boundary now does	measured
P4	75 of 579; the logging gap; four instrument failures (§6)	A self-report that tracks the logs it claims to summarize	measured
P5	The 8,221-turn record and the blind key (§7)	A broad preference dataset that recovers not just dislike but wrong, why, and what happened next	interpretive

What transfers, if anything does. Three rules earned here do not depend on this practice's N of 1. A boundary that refuses and records its refusal outlasts a memo that asks an agent to remember (§5). A challenge must cite evidence the challenged agent can open, or it is an opinion wearing a citation (§8.2). And an outside checker must be calibrated against known-false cases before it is allowed to gate anything, because a checker built from the same class of model may simply agree with the claim it is shown (§9).

FROM §11 · LIMITS

External. One operator, one domain, one harness, one model family for the agents. Nothing here generalizes to other operators without replication, and the multiple-baseline extension raises N to four within one organization, not to a population. The setting's ecological validity is bought with exactly this cost.

How to cite this part. Atkinson, B. (2026). *AI doesn't need to be smarter. It needs a boss.* Background paper 2 to *How do we use this?* Working paper, Wolfberg LLC.

Disclosure. Drafting and literature synthesis were assisted by AI models (Claude, Anthropic) working under the author's direction; the author is responsible for the content. The works in the References were

read in full, and every specific figure cited comes from a work read in full. The prior-art works listed under Prior art are cited at the level of an established concept and its origin: each was verified for author, title, year and venue, but not read in full, and no numeric claim rests on any of them. Software documentation, source code and press accounts are listed under their own headings and were read at the linked pages. Every reference below carries a link, and every arXiv identifier and DOI was resolved against its registry, with title and first author matched, on 23 September 2026.

Competing interests. The author owns Wolfberg LLC, the practice studied.

Data availability. The practice's logs contain client work and personal records. They are private and are not offered for sale or sharing. The measures are described in enough detail to be reimplemented, and figures from the practice are reported as of the dates given. What is available is the design: the clauses of §8.1, the evidence contract and admission checks of §8.5, and the decision rules of §10.4 are stated fully enough to be rebuilt without access to the logs.

Corrections, 23 September 2026. No figure and no finding changed. The paper was retitled; earlier revisions were titled *The Stop Problem: Defensible Is Not Correct*, and the stop problem remains this paper's name for the failure it studies. The Anthropic interview in §2.3 aired on 13 September, not over a weekend of 13 and 14 September; the web article is stamped 14 September. The essay listed under Prior art is by Ryan Forstie; an earlier revision gave the initial K. The completion evaluator in §8.3 is documented as a Claude Code feature, whose agent loop the Claude Agent SDK embeds; an earlier revision attributed it to the SDK directly. Two works in the References, Graves (2016) and Liu (2026), were listed without being cited in the text; each is now cited where it bears (§2.1, §8.3). Links were added to every press, documentation and prior-art source. A duplicated section number in §10 was corrected.

Corrections, 25 September 2026. No figure changed. §8.3 said that the human-in-the-loop primitives across the three frameworks surveyed gave a developer nothing to require a person to confirm finished work. That holds for the OpenAI Agents SDK and the Claude Agent SDK. It does not hold for CrewAI, whose task documentation, already cited as CrewAI (2026b), offers an opt-in setting for a human to review the agent's final answer; §8.3 now says so.

WORKS CITED IN THIS PART

REFERENCES

Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025). *The GenAI Divide: State of AI in Business 2025.* MIT NANDA project, July 2025. Project page: nanda.media.mit.edu; the report as read: mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Kelley, S. W., & Riedl, C. (2026). *Personalization increases affective alignment but has role-dependent effects on epistemic independence in LLMs.* arXiv:2603.00024. arxiv.org/abs/2603.00024

Mehta, A. (2026). *When agents commit too soon: Diagnosing premature commitment in LLM agents.* Snowflake AI Research. arXiv:2606.22936. arxiv.org/abs/2606.22936

Mytsyk, S., Zhang, Y., & Krishnamurthy, V. (2026). *Mitigating LLM sycophancy with RL-based fine-tuning: Bayesian truth serum approach.* arXiv:2608.25267. arxiv.org/abs/2608.25267

Wang, J., & Huang, J. (2026). *Reward hacking as equilibrium under finite evaluation.* arXiv:2603.28063. arxiv.org/abs/2603.28063

PRIOR ART (CITED AT CONCEPT LEVEL; VERIFIED FOR AUTHOR, TITLE, YEAR AND VENUE, NOT READ IN FULL)

Forstie, R. (2026). The part of the agent stack nobody wants to build. LinkedIn, 25 August 2026. linkedin.com/pulse/part-agent-stack-nobody-wants-build-ryan-forstie-ihyqc *Cited for its three closing questions on managed-agent platforms, first read on 1 September 2026; author, title, date and the three questions re-verified at the source on 23 September 2026. No figure from it is cited anywhere in this paper.*

SOFTWARE AND DOCUMENTATION (READ AT THE LINKED PAGES, SEPTEMBER 2026; CURRENT AS OF THEN, NOT PERMANENT)

Anthropic. (2025, September 29). *Building agents with the Claude Agent SDK.* claude.com/blog/building-agents-with-the-claude-agent-sdk

Anthropic. (2026b). *Keep Claude working toward a goal.* Claude Code documentation. code.claude.com/docs/en/goal

CrewAI. (2026b). *Tasks.* CrewAI documentation. docs.crewai.com/en/concepts/tasks

PRESS AND PUBLIC STATEMENTS (Â§2.3; READ AT THE LINKED PAGES, OR AT THE NAMED CARRIER WHERE THE ORIGINAL IS PAYWALLED)

Axios. (2026, September 3). *Sam Altman's sobering siren.* Interview at the G20 Innovation Ministerial. axios.com/2026/09/03/axios-interview-sam-altmans-sobering-siren; paywalled, read as carried by The Next Web: thenextweb.com/news/sam-altman-axios-idea-guy-sobering-models

CBS News. (2026, September 13). *Anthropic CEO Dario Amodei: "For too long the industry lied" about AI risks.* Sunday Morning; web article updated 14 September. cbsnews.com/news/anthropic-ceo-dario-amodei-on-ai-risks/

CNBC. (2026, September 13). *Anthropic's Amodei says China presents "toughest dilemma" for his proposed AI slowdown.* Cited for the broadcast date. cnbc.com/2026/09/13/china-dilemma-ai-slowdown-anthropic.html

Fortune. (2026, May 26). On the two laboratories walking back earlier job-loss forecasts ahead of public offerings; cited for the reported valuations. fortune.com/2026/05/26/sam-altman-dario-amodei-walking-back-ai-jobs-apocalypse-prophecies-ipo/

Fortune. (2026, September 12). Interview with the chief executive of OpenAI on safety and the timing of a public offering. fortune.com/2026/09/12/sam-altman-interview-ai-doomsday-safety-models-control-ipo-2027/

Reuters. (2026, September 14). *Wall Street ends down, calls for AI slowdown pummel chipmakers.* As carried by Yahoo Finance. finance.yahoo.com/technology/ai/articles/ai-warnings-knock-nasdaq-futures-092329455.html