# Everyone Is a Manager: AI Use as an Act of Management*

Benjamin T. Leyden† Daniela Scur‡

September 7, 2026

**Preliminary, please do not cite.**

**Abstract**

We propose that the use of AI is an act of management. When you give a knowledge worker an AI assistant, they stop executing tasks and start delegating and checking the output of a fallible subordinate they were never trained to manage. We model the return to AI as moderated by two managerial abilities: direction (how well they specify what they delegate) and verification (how reliably they catch the tool's errors). The return to AI use falls in the worker's own task skill and rises in both managerial abilities. As AI becomes more reliable, the skill with the higher marginal return shifts from verification to direction. To the extent that use builds skill, workers improve their managerial abilities in the normal course of work. However, as the tool improves, verification becomes the exception to this. Once checking is no longer worth the effort, the practice that sustains it stops, and the skill collapses. Workers will reach that point sooner if they only bear part of the cost of their errors. Finally, we show that in equilibrium, the market sends its strongest verifiers to the jobs where that collapse costs the most. We recast the concern about AI and human capital as a management problem: AI turns its users into managers, and while direction builds itself through use, verification is the ability a good tool stops teaching and the one firms must screen for at hire.

*We thank Kevin Donovan, Daniel Rock, John Van Reenen, Simon Quinn, and participants at the 2026 Schmidt Sciences AI at Work Convening and the Empirical Management Conference 2026 for helpful comments.

†Cornell University, leyden@cornell.edu.

‡Monash University, dscur@monash.edu.

*Above all, the new technology will not render managers superfluous ... it will demand many more managers. It will greatly extend the management area; many people now considered rank-and-file will have to become capable of doing management work ... to see and think managerially.*

—Drucker (1954), *The Practice of Management*, ch. 3

# 1 Introduction

AI is fundamentally changing how people work in myriad ways and, crucially, people who were not hired into a managerial role are now effectively performing managerial tasks—often without having been screened for managerial ability or having the skills for the role. Consider, for example, an engineer who is given an AI assistant. The tasks they are charged with do not change—a part to design, a system to ship—but the way they complete their work is now different. Rather than conducting most of it themselves, they now describe what they want, answer clarifying questions, evaluate what the tool has produced, and decide whether it is good enough to ship. Whenever the engineer misses a flaw, the mistake goes out under their name. This is the work of a *manager*, yet no one selected or trained the engineer for it.

Our scenario is not hypothetical: in OpenAI's Codex usage logs, the share of users delegating a task an experienced human would need more than a full workday to complete rose from 2.1% to 25.6% in the first half of 2026, and more than a quarter of OpenAI's own employees managed five or more concurrent agents at some point in a single week in June 2026 (Johnston et al., 2026).

In this paper we model a knowledge worker—importantly, *not* a manager—who has been assigned a portfolio of tasks and has access to an AI tool. For each task they can do the work themselves (*Solo*), hand it to the tool and ship what comes back (*Trust*), or check the AI's output before it ships (*Vet*). We characterize the worker by a task skill and two distinct managerial abilities: *direction*, how well they specify and decompose the work they delegate, and *verification*, how reliably they catch the tool's mistakes.[1] We characterize the AI tool by its *reliability*, how often its output is right, and its *scope*, which captures the range of tasks it can handle. The distinction between task skill and managerial ability has empirical support. Riedl and Weidmann (2025) measure a worker's "AI collaboration ability," which captures their ability to know when to delegate, how to phrase the request, and whether to accept the answer, as a trait distinct from their solo task skill, and find that the two are only weakly correlated.

We consider the implications of this framework in three stages. First, we hold the two skills fixed and ask how each one moderates the return to AI. We then let AI use build the skills themselves

[1] Of course, managing is much more than this. Managers rely heavily on interpersonal skills (e.g., motivating, developing, and leading the people who do the work), which we set aside. In this analysis, we restrict our attention to the part of management that concerns the work itself: directing what gets done and verifying that it was done right. This restriction is appropriate for our setting because an AI subordinate has no morale and no agenda of its own, and so the interpersonal side of management has little relevance here. We pick up the possibility of spillovers into people management in concurrent work.

(i.e., learning by doing), so that the returns of the first stage become the incentives that determine which skill is further developed. Finally, we document which jobs the two skills flow to and what each is paid in a competitive labor market.

When the worker's skills are held fixed, the return to AI falls in task skill and rises in each managerial ability. This two-part split reconciles a conflict in the empirical literature over whether AI *compresses* or *expands* the skill distribution: the compression studies (Brynjolfsson et al., 2025; Noy and Zhang, 2023) are measuring substitution along task skill and the expansion study (Otis et al., 2024) the management channel along managerial ability—two skill axes rather than a single contested one. As AI reliability increases, the leading skill crosses over from verification to direction. Errors become rarer and the ones that survive a near-reliable tool are the subtle ones, while every harder task the tool can now address is one whose value depends on how well the worker delegated it. Where the surviving errors are rare but catastrophic (so that the cost of an error climbs as errors grow scarce), verification keeps paying however reliable the tool becomes. Finally, as the AI's scope widens, the worker's own time shifts from doing tasks to directing and checking them.

Skills, though, are not fixed. The same activities that earn the return to AI are also practice.[2] That is, workers build managerial skills in the normal course of their work (a proposition we discuss in Section 4). Because the work builds whichever skill the job requires, the crossover in returns becomes, over time, a crossover in the skills themselves: as the tool improves, the worker improves their direction while losing their ability to verify once checking no longer pays. The same logic operates at the level of the firm: Babina et al. (2026) trace the productivity gains from AI investment to organization capital, the durable knowledge that accumulates as a by-product of operations. They measure a firm-level stock, while the abilities we model belong to the worker—a distinction that matters for who has the incentive to invest in each.

Of course, workers almost never bear the full cost of the errors they ship, since pay is not docked flaw by flaw. The gap is a feature of nearly every employment contract, and one of the oldest problems in organizational economics (Holmström, 1979; Holmström and Milgrom, 1991). We introduce an "accountability wedge" between what an error costs the firm and what it costs the worker.[3] As a result, verification is practiced less and the skill settles at a lower level. Ultimately, the worker can fall into a "deskilling trap," in which the verification the worker could engage in is no longer worth the fixed cost of beginning to check, so they stop verifying and their skill collapses. Importantly, this wedge does not directly affect *direction* because a better-specified job is an outcome the worker is paid for in full, while a caught error saves a loss, of which they keep only a fraction. Thus, AI use builds a "lopsided" manager, strong at directing and weak at checking.

The nature of the work determines whether that skill erosion matters. Across most of the economy the erosion is benign because the errors that survive it are rarely costly enough to matter. The danger arises where errors are rare and catastrophic—medicine, aviation, automotive

[2]Shen and Tamkin (2026) find that engaged AI use, working through the tool's output, preserved skill formation, while pure delegation eroded it.

[3]We use "firm" going forward for ease of exposition, but our model and insights are generalizable to any sort of organization that has knowledge workers—public or private.

engineering—because in those fields the stakes rise as the errors thin out. While the worker in those settings still needs to catch errors, the skill for catching them was built and maintained by routine verification that is no longer occurring. We read this as a targeting rule: where errors are common and/or cheap, the decay of the checking skill is potentially acceptable, and where they are rare and catastrophic it is the central risk. In Section 5, we show that firms may optimally choose to allow verification to collapse rather than pay to prevent it. Indeed, as the tool improves, the case for preventing it only weakens.

Finally, in a competitive labor market, workers sort along each job attribute by the skill it rewards: across jobs of a common scope, and holding direction skill fixed, stronger verifiers take the higher-stakes jobs, and across jobs of a common stakes scale, holding verification skill fixed, stronger directors take the wider scope. Verifiers therefore end up in the high-stakes jobs, where an uncaught error costs the most and where the accountability wedge bites. While the assignment is efficient the day it is made and a high-stakes job even induces more checking, the market only delays the collapse.

Our model connects several literatures. The most natural reference point is the knowledge-hierarchy framework of Garicano (2000) and Garicano and Rossi-Hansberg (2006), where a worker handles the problems within their competence and refers the rest upward. AI fundamentally alters that mechanism because an AI assistant will not always tell the worker when a task is beyond its ability.[4] Instead, the tool may return confident output regardless of its accuracy, replacing the upward flow of exceptions with a verification problem the worker has to handle themselves.[5] Moreover, the problem is not contractable in a principal-agent sense, because the tool has no agenda of its own. Thus, any misalignment is a failure of specification on the worker's part (Ellis and Huang, 2026), not a conflict of interest. This motivates our decision to separate a worker's innate managerial ability into direction and verification.

Several recent papers place AI directly inside this hierarchy. Ide and Talamàs (2025) model AI as agents that substitute perfectly for a human of equal knowledge, and derive winners at both tails of the skill distribution as workers re-sort across roles. In their setting, AI is never wrong, so the oversight margin we study does not arise, and as a result our equilibrium findings diverge, as workers re-sort on managerial ability rather than task knowledge. Relatedly, Xu et al. (2026) consider a generative AI agent that can hallucinate and thus requires human validation, but their worker is constrained by their own expertise, so verification is inseparable from task skill. We instead model verification as a graded detection technology, which is therefore distinct from task skill. Huang et al. (2026) develop a concurrent framework in which the worker also chooses among manual work, pure delegation, and verified delegation, with a misaligned-evaluation layer doing the work of our accountability wedge. Quispe and Xu (2026) put a similar menu of work modes to work

[4] While this is not unlike human subordinates, it is a different interaction when the human component of uncertainty is not present. For example, a manager might be able to tell a worker isn't sure of their ability for a task by seeing how they react when assigned the task, or by hearing the hesitancy in their voice as they present their work.

[5] Garicano et al. (2026) study how AI redraws the boundary of a job by unbundling its tasks, with AI entering as effective machine services and no role for error or escalation.

in an empirical study of the programming languages developers take on once they have an agentic assistant, though direction and verification are a single ability in their worker rather than two.

Direction and verification each have an antecedent on the human side of the AI literature. Agrawal et al. (2018) show that cheaper machine prediction raises the return to judgment. Direction extends that ability from valuing a decision to specifying and owning a delegated task.[6] Ours is also a delegation problem, though not the classical principal-agent version (see, e.g., Aghion and Tirole (1997) and Dessein (2002)) because our AI agent is non-strategic. Athey et al. (2020) consider an AI inside such a problem and demonstrate that the (human) agent's effort and the AI's quality are strategic substitutes, so the principal may prefer weaker AI. But, as with Agrawal et al. (2018), the effort their AI crowds out is spent producing information before the decision, whereas ours is spent checking an output once it is in hand.

In terms of verification, Caplin et al. (2024) show that calibration (i.e., how well a worker knows their own judgment) predicts who benefits from AI even after task ability is held fixed, and verification formalizes that calibration. Bastani and Cachon (2026) reach the opposite conclusion: oversight grows *more* costly to sustain as the AI improves. But their inspector's wage explodes as errors become less frequent, while our worker is their own inspector and so the value of catching errors falls with the likelihood of errors.[7] Catalini et al. (2026) make human verification the binding constraint on growth in an agentic economy, where aggregation removes the skill heterogeneity and the direction channel that a worker-level treatment recovers. Acemoglu et al. (2026) find a pair of stable steady states, a skilled state and a collapse state separated by a threshold. The skilled state disappears once AI quality passes a critical level, leaving only collapse. Our model has the same structure, though theirs arises from an uninternalized externality in a community's knowledge while ours arises from the accountability wedge at the level of an individual worker.

Finally, managerial ability is a real and measurable primitive. Management practices vary widely across firms, predict productivity and survival (Scur et al., 2021), and behave like a technology rather than only a style (Bloom et al., 2017). Weidmann et al. (2025) recover the same kind of variation in leadership skill at the level of the individual by having people direct teams of AI agents. Weidmann et al. (2024) randomly reassign people to lead fresh teams and recover a stable, causal manager effect worth roughly twice a good worker's. That individual-level primitive is the one our model uses. Alekseeva et al. (2026) find the firm-level counterpart of our claim in the labor market: AI adoption is associated with greater demand for managerial expertise in firms' hiring. Bick et al. (2026) attribute the bulk of the gap in AI adoption between the United States and Europe to management practices, and in particular to whether a firm encourages its workers to use the tool and gives them access to it. Ours is the within-worker version, the same pull toward management operating on individual workers rather than on the firm's demand for managers.

[6] By judgment Agrawal et al. mean the assessment of what a decision's payoffs actually are. Their fallible signal is a prediction whose unreliability the decision-maker prices in advance. In our case, its errors can only be caught ex post—verification rather than forward-looking judgment.

[7] In Siderius et al. (2026), the worker's diligence covers both, catching today's errors and preserving the expertise they would otherwise lose. Similar to this paper, they find that the firm may tolerate the erosion rather than pay to prevent it.

Read as a management problem, the concern about AI and human capital is more tangible: the primary issue is that it *informally* reassigns AI users to management roles and builds only half of the manager it needs. Specifically, continuous use lets verification decay precisely where a rare error could be most costly to miss. This changes what firms should screen for. Task skill, the typical entry-level screen, no longer captures the return to AI, and while direction can be left largely to the job, verification must be selected for at hire, because a worker who starts below the trap's threshold cannot rebuild it through use.

The remainder of the paper proceeds as follows. Section 2 sets out the model. Section 3 derives the static returns to the two managerial skills. Section 4 lets the two skills accumulate. Section 5 introduces the accountability wedge. Section 6 studies sorting and wages. Section 7 concludes, taking up the model's assumptions and extensions, what our results imply for hiring, and the empirical agenda that follows.

## 2 The Model

We model how much a worker gains from using AI, and how that gain is moderated by their innate managerial ability. Throughout the remainder of the paper we ground our model in the example of an engineer who has been provided an AI assistant that can draft components, run simulations, and write code, so any task it attempts comes back as a finished piece of engineering work the worker can adopt or reject. However, our framework applies to knowledge work more broadly, such as a consultant turning an AI-drafted analysis into client advice, a paralegal checking a generated brief, or a physician working through a diagnosis with an AI tool.

The engineer in our example faces a portfolio of tasks, each described by its difficulty $d \in [0, 1]$ and its stakes $L \geq 0$.[8] A correctly done task is worth $V$ and a wrong answer that ships costs $L$. Their time carries a constant shadow value $\lambda > 0$, so the tasks are separable.[9]

Our engineer is defined by three capacities. Their task skill $a \in [0, 1]$ is the probability they complete the task correctly on their own. Their *direction* $m_d \in [0, 1]$ is how well they specify and orchestrate any work they delegate to AI, and it scales the value a delegated task delivers through a multiplier $\varphi(m_d)$ with $\varphi > 0$ and $\varphi' > 0$.[10] A capable AI that is badly directed will produce confident slop at scale, while when well-directed it will produce valuable output. Finally, the worker's *verification* $m_v \in [0, 1]$ is how reliably they catch the tool's mistakes, through a detection technology $\rho(m_v, e; q)$. Direction and verification are distinct competencies—e.g., the engineer can be sharp at catching the machine's errors yet clumsy at steering it—and together they are what we mean by

[8] Across the portfolio the pair $(d, L)$ is drawn from a joint distribution whose difficulty marginal $F_D$ has a continuous, strictly positive density $f_D$ on $[0, 1]$, and whose stakes marginal is atomless with bounded support, $L \leq L_{\max}$. Wherever our analysis aggregates over the portfolio rather than considering a single task, we further assume that difficulty and stakes are independent (Assumption 2).

[9] In appendix B.3, we relax that separability by replacing the price of time with a shared attention budget.

[10] This multiplicative form is essentially the span-of-control device of Rosen (1982), where a director's talent scales the work done under them rather than their own output. Garicano and Rossi-Hansberg (2006) use the same device in their knowledge hierarchy. Importantly, we impose no normalization such as $\varphi \leq 1$, so directed output can be worth more than the worker could produce alone.

the worker's innate managerial ability.

AI in our model is defined by two characteristics. Its *reliability* $q \in [0,1]$ is the probability that an attempted output is correct, and its *scope* $\kappa \in [0,1]$ is the breadth or difficulty of work it can take on. It can attempt a task whenever $d \leq \kappa$. We treat scope as exogenous.

For a given task the AI can attempt ($d \leq \kappa$), the engineer has three options. They can do it themselves (*Solo*), they can hand it over and take the result as it comes (*Trust*), or they can conduct a costly review of the output (*Vet*). We assume that for tasks beyond the AI's scope ($d > \kappa$) only Solo is available. Each option carries an expected payoff to the engineer:

$$\text{Solo:} \quad u_{\text{Solo}} = aV - (1-a)L - \lambda t_H, \tag{1}$$

$$\text{Trust:} \quad u_{\text{Trust}} = qV\varphi(m_d) - (1-q)L - \lambda t_A, \tag{2}$$

$$\text{Vet (at effort } e\text{):} \quad u_{\text{Vet}}(e) = V\varphi(m_d)\big[q + (1-q)\rho\big] - (1-q)(1-\rho)L - \lambda t_A - \lambda f - \lambda c(e), \tag{3}$$

where $t_H$ is the time a task takes done themselves and $t_A$ the time it takes delegated, with delegation faster ($t_A \ll t_H$), and a convex effort cost $c(e)$ ($c' > 0$, $c'' > 0$). Vetting also carries a setup cost $f > 0$ that trusting does not, representing the fixed amount of time, valued at $\lambda$, the worker spends getting ready to properly judge the output. This might include pulling the task back into mind, reconstructing what a right answer should look like, or forming an independent expectation to hold the tool's against. The setup cost creates a friction that makes it optimal to ship a reliable tool's output unchecked when the stakes are low. Note that Solo and Trust are symmetric on errors, since when the worker produces and ships a wrong answer themselves it costs $L$, just as the AI's error does.

The engineer selects whichever action pays most, task by task. The *return to AI* on a task is how much using AI is an improvement on doing it themselves,

$$R = \max\{u_{\text{Solo}},\, u_{\text{Trust}},\, u_{\text{Vet}}(e^*)\} - u_{\text{Solo}} \;\geq\; 0. \tag{4}$$

Because $R$ nets out what the worker would have produced alone, we measure how task skill, direction, and verification moderate the return to AI by how each moves $R$.

The value of catching errors is then (3) − (2):

$$u_{\text{Vet}}(e) - u_{\text{Trust}} = (1-q)\,\rho(m_v, e; q)\,\big(V\varphi(m_d) + L\big) - \lambda c(e) - \lambda f. \tag{5}$$

Each caught error turns a $-L$ outcome into a $+V\varphi$ one (a swing of $V\varphi + L$) weighted by how often the AI errs, $(1-q)$, and how often they catch it, $\rho$. The optimal vetting effort balances that gain against the cost,

$$(1-q)\big(V\varphi(m_d) + L\big)\,\rho_e(m_v, e; q) = \lambda c'(e), \tag{6}$$

where $e^*$ is the effort they put in once they have chosen to vet. The worker thus balances the

marginal benefit of scrutiny (the value of a caught error, weighted by how quickly effort produces catches and by the likelihood that there is an error to catch) against its marginal cost (the shadow value of their time times the marginal cost of effort). Whether they vet at all is a separate, discrete choice. By equation (5), they vet only when the value of the errors caught clears the setup cost $\lambda f$, and otherwise, if they delegate at all, they ship unchecked.

Because the AI does not refer its exceptions upward, the upward flow that makes a human hierarchy efficient is replaced by the engineer's own detection technology $\rho$.[11] Detection is increasing in both the worker's verification skill and their scrutiny effort, with scrutiny running into diminishing returns. As reliability rises, catching errors gets harder, because the ones that survive are the subtler ones. Skill and scrutiny are complements ($\rho_{m_v e} \geq 0$), so an added hour of checking yields more catches if the verifier is skilled. And a checker with no skill catches nothing however hard they look, which is what makes zero verification skill self-perpetuating in Section 4.

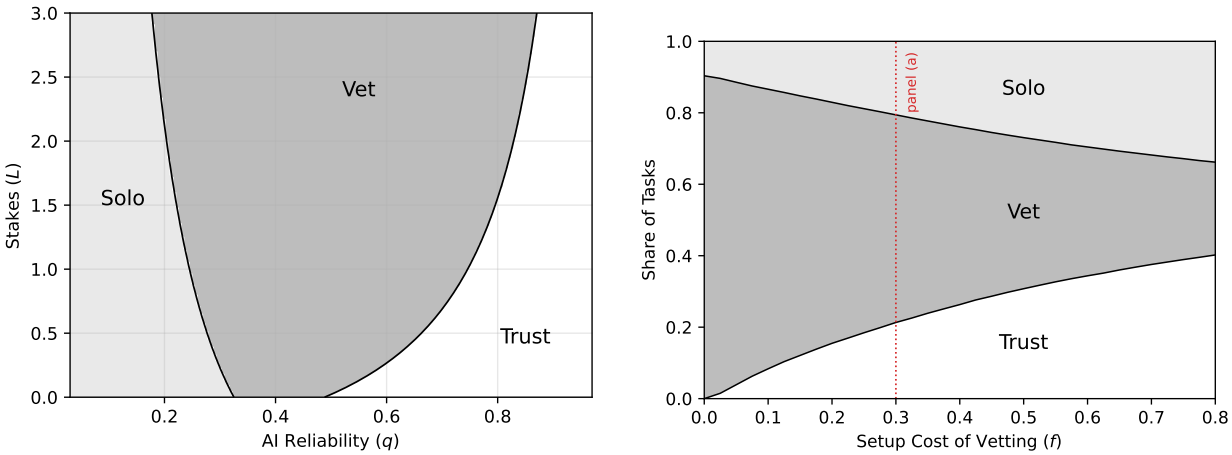
Notably, catching an error does not require being able to produce the right answer yourself, that is, detection does not depend on task skill $a$. In our running example, the engineer can flag a flaw in a design they could not have produced from scratch. We view this as consistent with how managing has always worked: a conductor cannot out-play the orchestra, a coach cannot out-run the sprinters, an editor does not out-write the writers: each directs, and answers for, people more skilled at the task itself. The one case in which task skill substitutes for verification is re-derivation, redoing the task to check it. Indeed, that is exactly the case that becomes infeasible once the AI is better at the task than the worker. What remains is sense-checking, which lets a worker add value managing a tool more capable than they are.

We assume that trusting the tool is never worse than doing the task Solo. As a result, whenever checking is not worth its cost the engineer's fallback is Trust. We refer to this assumption as "Trust dominance," and we formalize it in Assumption 3. Under Trust dominance, the tool's speed and its directed value alone must repay delegation, before any accounting for what an error costs, and that comparison must survive full exposure to the costliest error in the portfolio. This assumption would fail for a worker that is more accurate than the tool on the tasks at hand, since that worker would rationally do the work themselves.

Further, we assume that delegating pays on at least some tasks even for an engineer with no direction skill. Such an engineer still draws $qV\varphi_0$ from a delegated task ($\varphi_0 \equiv \varphi(0) > 0$), and delegation is faster ($t_A \ll t_H$). Formally, this "delegation floor" (see Assumption 4) requires that there exists a set of tasks of positive mass for which the undirected tool beats Solo. This is therefore a milder condition than Trust dominance. The floor likewise fails for a worker that the undirected tool cannot beat, i.e., where $a$ is large relative to $q\varphi_0$ and the time savings are too small to close the gap.

In what follows, we occasionally rely on a parameterized example, setting $\varphi(m_d) = \varphi_0 + (1-\varphi_0)m_d$ (so $\varphi' = 1 - \varphi_0 > 0$), $c(e) = \frac{1}{2}e^2$, a setup cost $f > 0$ (here $c'(0) = 0$, so whether they vet at all

[11] See appendix A.1 for a formal statement of our assumptions on $\rho$.




(a) Action partition

(b) Setup cost

Figure 1: The action partition and the setup cost

*Note:* **(a)** The three actions over a task's reliability $q$ and stakes $L$: Solo where the tool is too unreliable to use, Trust where it is reliable and the stakes are low, Vet in between. **(b)** The same three actions as shares of the reliability–stakes plane, as the setup cost of vetting $f$ varies (shaded as in panel (a), the three shares summing to one). As $f \to 0$, the Trust band, where the tool's output ships unchecked, vanishes. As $f$ rises, the Vet share falls and both the Trust and Solo shares grow. The dotted line marks the baseline $f$ of panel (a).

depends entirely on $f$), and

$$\rho(m_v, e; q) = m_v\,(1 - \delta q)\left(1 - e^{-\gamma e}\right), \qquad \delta \in (0,1),\ \gamma > 0, \tag{7}$$

together with difficulty $d \sim U[0,1]$, so the AI's reachable share of tasks is exactly $\kappa$. This form satisfies the assumptions of appendix A.1 and makes verification skill and effort complements everywhere, $\rho_{m_v e} = (1 - \delta q)\gamma e^{-\gamma e} > 0$.[12]

Figure 1a shows how the three actions partition the task space in this example. The engineer does it themselves when the tool is too unreliable to delegate (low $q$); they trust it where it is reliable and the stakes are low; and they vet in between, wherever errors are frequent or costly enough to be worth catching.[13]

## 3 Baseline Results

The engineer can do the task themselves, trust the machine, or check its work. As the machine improves, the primary role of the engineer becomes management. We show in this section that the return to management persists as AI improves, though its source shifts from verifying the machine

[12]This complementarity matters for the uniqueness of the crossover in Section 3 but not for its existence. Without it the decline in the return to verification could potentially stall, in which case a crossover could happen more than once.

[13]Huang et al. (2026) identify a similar trichotomy in concurrent work. Their three regimes partition the space of worker abilities while ours partition the space of tasks.

to directing it. The persistence rests on direction's marginal value staying strictly positive.

### 3.1 The Cross-Sectional Return Decomposes

Given a task the engineer chooses to vet, we now ask how its return to AI moves with their task skill and managerial abilities. The answer separates into three channels.

**Proposition 1** (Decomposition and the two skill axes)**.** *On an in-scope task in the Vet region, the return to AI* (4) *moves with the three capacities as*

$$\underbrace{\frac{\partial R}{\partial a} = -(V+L) < 0,}_{\textit{substitution}} \qquad \underbrace{\frac{\partial R}{\partial m_d} = \big(q + (1-q)\rho\big) V \varphi'(m_d) > 0,}_{\textit{direction}} \qquad \underbrace{\frac{\partial R}{\partial m_v} = (1-q)\big(V\varphi + L\big)\rho_{m_v} > 0\,.}_{\textit{verification}} \tag{8}$$

(i) *The return to AI falls in task skill at every level of stakes, and because* $\partial^2 u_{\text{Solo}}/\partial a \partial L > 0$*, higher stakes* deepen *the decline.* (ii) *On a vetted task, the return to each managerial skill rises with stakes,* $\partial^2 R/\partial m\, \partial L > 0$ *for* $m \in \{m_d, m_v\}$*.*

Conceptually, a higher task skill raises the value of going Solo, so fewer tasks clear the bar for delegation and the return to AI falls. While one might expect higher stakes to reverse this, since a skilled worker has more to protect, they deepen it instead. Better direction raises the value of everything the worker delegates, and better verification raises the value of catching what the AI gets wrong to the extent that errors are both frequent and costly. Holding the task, the tool, and the worker's task skill fixed, the dispersion in the return to AI across workers is therefore entirely managerial.

This decomposition reconciles conflicting results in the empirical literature. Some studies have found that AI *compresses* the skill distribution, with lower-skill workers gaining most (Brynjolfsson et al., 2025; Noy and Zhang, 2023). Others have found the opposite, an *expansion* in which higher-skill workers pull ahead (Otis et al., 2024). In our model, compression is substitution along *task* skill $a$ (Proposition 1(i)), while expansion is the management channel along *managerial* ability, both direction $m_d$ and verification $m_v$, amplified by the stakes (Proposition 1(ii)). In the call center context of Brynjolfsson et al. (2025), the AI assistant is trained on the firm's highest-performing workers and thus provides their knowledge and expertise to everyone else, so we read the variation it acts on as task productivity, and the stakes on any one call are low. As a result, the gains concentrate where the substitution channel predicts, largest at the bottom of the skill distribution and slight at the top. The entrepreneurs of Otis et al. (2024) asked the AI similar questions and got similar advice, and what separated them was which advice they chose to act on, with a real business riding on the choice. There the high performers pulled ahead and the low performers fell back. Our engineer is the within-worker version of the same split. *Skill*, therefore, has two components, and what AI adoption does to a skill distribution depends on which of them separates the workers in question.

### 3.2 The Crossover

We next let the AI improve along its two dimensions at once, reliability $q$ and scope $\kappa$, and ask which of the engineer's two managerial skills is then more valuable. Verification will be more valuable when the tool errs often, but as reliability improves verification becomes less important simultaneously on two fronts. First, errors get rarer, so each catch is worth less, and the ones that survive a near-reliable model are the subtle ones, so catching gets harder—and the level of scrutiny the engineer chooses falls with both. At the same time, every harder task the tool can now reach is one whose value depends on how well it is delegated. Thus, direction rises as verification fades.

We value an in-scope task with the current AI model at $U = \max\{u_{\text{Trust}}, u_{\text{Vet}}(e^*)\}$. Under Trust dominance (Assumption 3) the engineer never falls back on Solo, so the maximum runs over these two actions alone. The gap between the two managerial returns is what a marginal unit of verification skill buys the engineer, less what a marginal unit of direction skill buys them:

$$\Delta^*(q) \;=\; \frac{\partial^+ U}{\partial m_v} - \frac{\partial^+ U}{\partial m_d}. \tag{9}$$

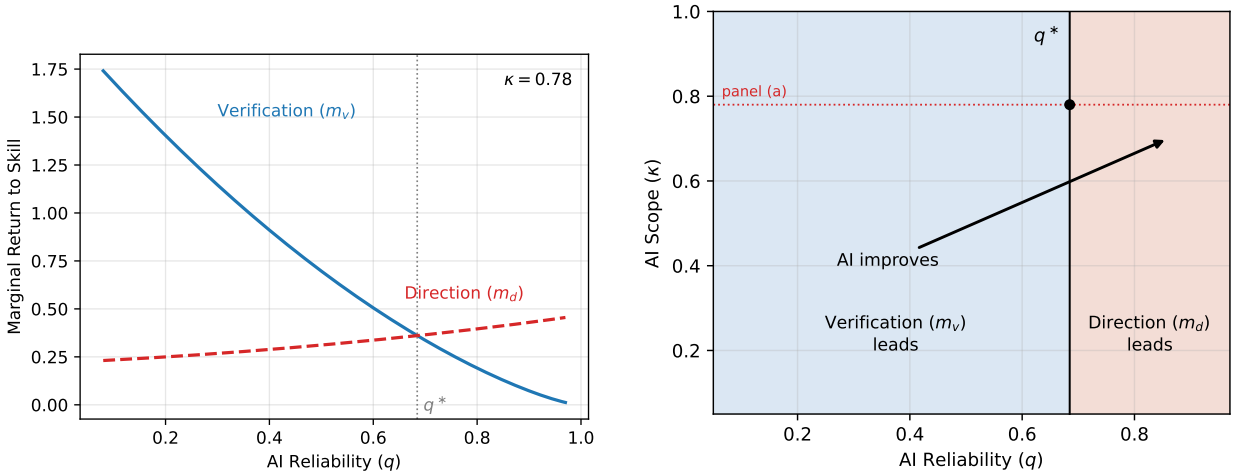
Both are total derivatives, taken at the optimum, so each includes the vetting effort the skill induces. Where Trust and Vet tie, $U$ is not differentiable and we take the one-sided derivative from above in each skill, which is what a worker acquiring skill faces. The skill with the higher marginal return *leads*: verification leads where $\Delta^* \geq 0$ and direction where $\Delta^* < 0$. We call the last reliability at which verification still leads the crossover reliability $q^*$. The following proposition establishes that the leading skill does in fact cross over as the tool improves.

**Proposition 2** (The crossover)**.** (i) *Optimal vetting effort falls as reliability rises, $de^*/dq < 0$.* (ii) *The return to verification falls with reliability,*

$$\frac{\partial^2(\mathit{value})}{\partial q\, \partial m_v} = -\rho_{m_v}(V\varphi + L) + (1-q)(V\varphi + L)\Big[\rho_{m_v q} + \rho_{m_v e}\, \tfrac{de^*}{dq}\Big] < 0, \tag{10}$$

*where the final term is the effort margin, signed by $\rho_{m_v e} \geq 0$ (Assumption 1) and part (i).* (iii) *Under Assumption 2, the return to direction rises with scope: $\partial^2(\mathit{value})/\partial\kappa\, \partial m_d = f_D(\kappa)\, \mathbb{E}_L\big[\partial R/\partial m_d \mid d = \kappa\big]$, which equals $V\varphi'(q + (1-q)\rho)$ where the marginal task is vetted, $qV\varphi'$ where it is trusted, and zero where it stays Solo. It is nonnegative always, and strictly positive whenever the marginal task is delegated. (See Proposition A1 for the full statement and its hypotheses.)*

When verification leads at low reliability, raising reliability at fixed scope makes the returns to the two managerial skills converge. In the parameterization pictured in Figure 2a, they cross exactly once. For a weak enough verifier in our parameterized example, though, vetting stops covering its setup cost before the returns meet, and the crossover arrives as an exit rather than an equality (Proposition A1(v)). If Trust dominance holds at $q = 0$ and verification leads strictly near $q = 0$, the crossover reliability $q^*$ exists, lies strictly inside $(0, 1)$, and marks the last reliability at which




(a) The crossover, sliced ($\kappa = 0.78$)

(b) Which skill leads over $(q, \kappa)$

Figure 2: The crossover from verification to direction

*Note:* **(a)** Holding scope fixed at $\kappa = 0.78$ and raising the AI's reliability, the marginal return to verification skill $m_v$ falls while the return to direction skill $m_d$ rises, and the two cross once at the reliability threshold $q^* \approx 0.69$. The crossing's existence and its order (verification $m_v$ first, then direction $m_d$) hold whenever verification leads at low reliability (Proposition 2); the level of $q^*$ is parametric. **(b)** The same crossing over the whole capability space. Because both marginal returns scale with scope, the boundary is the vertical line $q = q^*$. Scope changes how much management is worth without changing which skill leads. The arrow marks AI improvement (reliability and scope rising together), crossing the threshold as it climbs. The dotted horizontal line marks the scope at which panel (a) slices, and the dot the crossing it shows.

verification still leads (Proposition A1(iv)).[14] Both returns scale with the volume delegated, so under Assumption 2 widening scope raises them proportionally and the crossover does not move with scope (Figure 2b).[15]

Once reliability passes $q^*$, the engineer's value-add is through direction. Of course, whether they use the tool at all is a separate question, which does not depend on whether the tool is better at the task than they are.

**Corollary 1** (Managing the more capable)**.** *Because $\rho$ does not depend on task skill, the with-AI value $U = \max\{u_{\text{Trust}}, u_{\text{Vet}}(e^*)\}$ and the verification channel do not depend on $a$, while $u_{\text{Solo}}$ is increasing in $a$ with slope $V + L > 0$. The with-AI value and the Solo value tie at a single threshold*

$$a^* = \frac{U + L + \lambda t_H}{V + L}. \tag{11}$$

*For an in-scope task with $a < a^*$ the engineer earns strictly positive returns to AI. Past the crossover $q^*$ from Proposition 2, direction is the leading skill of the two for any task skill below $a^*$.*

The threshold can fall above or below the tool's reliability $q$, so $a < q$ neither implies nor is implied by $a < a^*$.

The value of verification depends on both the frequency of errors and their stakes, $(1 - q)L$, and the stakes can overturn the decline. If a reliable tool's remaining errors are rare but catastrophic—so that $L$ rises as $(1 - q)$ falls—then $(1 - q)L$ may not disappear as $q \to 1$, since rising severity offsets falling frequency, and verification stays valuable however reliable the AI becomes (Figure 3)—provided detection does not itself collapse in the limit, as it does not in our family, where $\delta < 1$ keeps $\rho_{m_v}$ bounded below. Verification thus fades in the high-volume, low-stakes regime and persists in the rare-catastrophic one. Which regime a firm faces therefore depends on *which* errors persist as the tool becomes more reliable (i.e., how stakes rise as errors grow rare). That is a feature of the environment rather than a modeling choice.

### 3.3 Reallocation Toward Oversight

So far, we have considered an engineer's decision regarding a single task. We now aggregate over the engineer's whole portfolio to show where their time goes. As the AI tool's scope widens, the engineer does less Solo work and more directing and verifying. In fact, Johnston et al. (2026) document exactly this reallocation. As Codex use deepens, the most intensive users begin overseeing portfolios of concurrent agents rather than executing tasks themselves. As the tool becomes more reliable, delegation becomes more attractive, which moves work toward oversight. However, each delegated

[14] We assume in equation (3) that the worker can repair any error they identify at no cost. As a result, a caught error is worth the full swing $V\varphi + L$, from a shipped loss to a correct delivery. An alternative approach would require the worker to pay a repair cost $r \in (0, V\varphi + L)$ whenever they catch an error. The return to verification would then fall, and $q^*$ with it, making our $q^*$ an upper bound on where verification stops leading.

[15] If the scope of the work is endogenized, so that harder tasks are costlier to delegate rather than impossible, the crossover threshold will rise with the difficulty of the tasks still worth delegating, and there would be no crossover for the hardest reachable tasks.

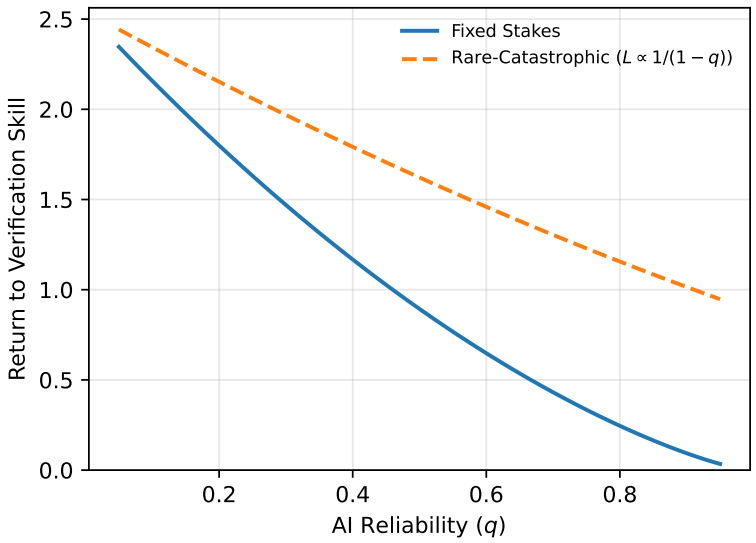



Figure 3: When verification fades

*Note:* At fixed stakes the return to verification falls toward zero as reliability rises. When stakes scale with the tail, so that $(1-q)L$ stays bounded, it persists.

task now needs less vetting, and so time spent verifying each one falls. Thus, whether a more reliable tool moves the engineer's time toward oversight or away from it is theoretically ambiguous.[16] The reallocation in scope is not, however. As scope widens, the share of the engineer's attention spent on oversight rather than on their own execution rises unconditionally (see Figure 4 and Proposition 8).

We can characterize when reliability helps, through two conditions. The first, the "vetted-value condition," says a more reliable tool raises what a vetted task is worth,

$$(1-\rho) + (1-q)\rho_q \geq 0 \quad \text{at the optimal effort } e^*. \tag{12}$$

The second, the "ship-correct condition," says it raises $q_{\text{sc}} = q + (1-q)\rho$, the probability a delegated task ships correctly,

$$\tfrac{d}{dq}\big[q + (1-q)\rho(m_v, e^*(q); q)\big] = (1-\rho) + (1-q)\big[\rho_q + \rho_e \tfrac{de^*}{dq}\big] \geq 0, \tag{13}$$

which is what makes the marginal return to direction nondecreasing in reliability (proof of Proposition 2, appendix A). Both hold under the parameterization we use for the figures presented, and both can fail for a strong verifier at low reliability, where the surviving errors get harder to catch faster than they get rare.

Ultimately, the introduction of AI has changed the shape of the engineer's day. They spend less of it executing tasks and more of it specifying work and checking what comes back. But of course,

[16] Xu et al. (2026) find the same tension one level up. In their two-layer hierarchy, the number of junior workers each expert supervises falls and then rises as AI improves. Their two forces affect the firm's staffing of workers and experts rather than the behavior of an individual worker, but the non-monotonicity has the same source as in our model.

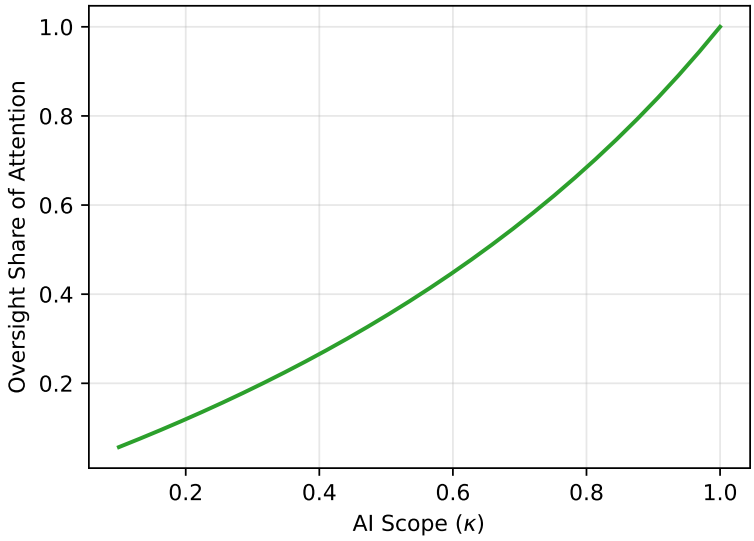



Figure 4: Reallocation toward oversight

*Note:* As the AI's scope $\kappa$ widens, the share of the engineer's attention spent on oversight (directing or vetting the AI) rather than on their own execution rises.

as the engineer increasingly relies on the new tool, they may become more skilled in its use. We next consider whether the repetition builds the managerial skills the tool now pays for, a question the static model cannot answer.

## 4 Building Managers: Skill Formation Through Use

Repeated use of an AI tool will likely make our engineer better at using it. For example, a few years into using the tool, they will be able to specify a subsystem faster and more completely than they used to, and catch more mistakes in the tool's output, even absent formal retraining. The work they are repeating, however, is managerial.[17] Thus far, we have held these two managerial skills at fixed levels. We now let them accumulate as stocks and consider the implications of this skill growth.

To do so, we take the static problem developed in Section 2 as a per-period *stage game*, and add a law of motion that characterizes how the mix of actions the worker engages in today affects their skill levels tomorrow. In this section, we show how using the AI builds managers in the ordinary course of the work, and then in Section 5 we ask whether it builds the *right* ones once they no longer bear the full cost of the errors they ship.

Let the engineer's skill state in period $t$ be $s_t = (a, m_{d,t}, m_{v,t})$. For this exercise, we hold task skill $a$ fixed to focus attention on the two managerial skills. Each period, the engineer solves the

[17] Shen and Tamkin (2026) document that the *mode* of AI use, and not merely its extent, is what governs what skill gets built. In a randomized trial, asking conceptual questions and working through the tool's output, preserved a worker's understanding, while pure delegation eroded it. We hold their outcome, task understanding, fixed.

stage game given their current skills. Skill formation is governed by two measures of those choices:

$$x_d(s;q,\kappa) = \Pr_{(d,L)}\{\text{delegate}\}, \qquad x_v(s;q,\kappa) = \mathbb{E}_{(d,L)}\big[e^*\,\mathbf{1}\{\text{vet}\}\big]. \tag{14}$$

$x_d$ is the share of the portfolio the engineer delegates; that is, the probability that the chosen action is Trust or Vet rather than Solo. $x_v$ is their average vetting effort. We measure the two differently because the engineer exercises the two skills differently. They direct every task they hand off, whether or not they verify what comes back, so $x_d$ measures how much directing the work requires. They exercise their verification skill only when checking output, and a detailed check exercises it more than a quick glance, so $x_v$ weights each vetted task by the effort spent on it. In what follows, we call $x_d$ and $x_v$ the engineer's *practice* in direction and in verification.

Importantly, these skills can depreciate with disuse. Letting $\eta \in (0,1)$ for the rate at which an unpracticed skill depreciates and $\psi_d, \psi_v > 0$ for what a unit of practice adds, the skills evolve as:

$$\begin{aligned} m_{d,t+1} &= \min\big\{1,\ (1-\eta)\,m_{d,t} + \psi_d\,x_d(s_t;q_t,\kappa_t)\big\},\\ m_{v,t+1} &= \min\big\{1,\ (1-\eta)\,m_{v,t} + \psi_v\,x_v(s_t;q_t,\kappa_t)\big\}, \end{aligned} \tag{15}$$

We assume throughout our analysis that our engineer is myopic. The engineer would learn that their verification ability has declined by observing (ex-post) the errors they fail to catch. As the tool grows more reliable there are fewer such errors, and the worker engages in less verification, and so little evidence of the skill decay reaches the engineer at all. They therefore solve each period's problem with no consideration for the future, and their skills accumulate and depreciate without their intending either. In Section 7, we discuss how things would change under a forward-looking engineer.[18]

Recall that in our baseline model, the engineer bears the full loss $L$ of any flaw they ship. This means that they delegate wherever delegation is worth doing and vet wherever vetting is, and each activity improves its own skill through equation (15). The setup cost $\lambda f$ shapes the vetting side from both ends. An engineer who vets at all will vet in earnest, since paying $\lambda f$ but then exerting no scrutiny will result in no caught errors, wasting the setup cost. Therefore, their effort is strictly positive wherever vetting is worth doing. An engineer whose verification skill is low enough will catch too few errors to justify paying $\lambda f$, so they will not vet at all. In that case, the skill will depreciate. Above a threshold in verification skill, the skills settle. Below it lies the deskilling trap we derive in Section 5.

Proposition 3 formalizes the skill growth we anticipated at the start of this section, specifying work better and catching more of the tool's mistakes.

**Proposition 3** (Skill formation as a byproduct)**.** *Consider myopic stage-optimal play under full accountability, $\theta = 1$, with practice $(x_d, x_v)$ and the law of motion* (15)*.* (i) Practice. *On any task on which Vet is stage-optimal the chosen effort is strictly positive, so $x_v > 0$ wherever vetting is*

[18] Siderius et al. (2026) instead model a forward-looking worker whose expertise atrophies without oversight effort, and their findings are broadly aligned with what we develop under the myopic approach here.

*stage-optimal on positive mass, and $x_d > 0$ wherever delegation is.* (ii) Convergence. *From any start with verification skill above a threshold $m_v^\dagger$, characterized in Section 5, both skills converge to a steady state $(m_d^*, m_v^*)$ with both coordinates strictly positive, and convergence is two-sided.* (iii) The floor. *Under the delegation floor of Assumption 4, direction practice holds $m_{d,t}$ above a positive bound from every start, whatever the path of verification. (See Proposition A2 for the full statement and its hypotheses.)*

For a new user, who starts below the steady state, that improvement now follows from stage-optimal play and the law of motion (15). That is, the engineer develops into a better director and a better verifier as a byproduct of completing work, without explicitly trying to become either. Of particular note is that the convergence result in Proposition 3 means that just as work can lift a novice up, it can also pull an experienced worker back down. That is, a skilled engineer placed in work that asks for little checking will lose verification ability at the depreciation rate $\eta$ until the two balance.

We next consider what happens to skill development as the AI tool improves. The reallocation we derived in Section 3 already moves the two practices in opposite directions: as the tool errs less, fewer tasks justify checking, and so verification practice $x_v$ decreases, and as it reaches harder tasks, more work repays delegating, and direction practice $x_d$ increases. If that reallocation continues as the tool improves, then the engineer's direction will settle higher and their verification will settle lower (see Proposition A2(iv), which carries the conditions in full). The incentives the worker faces mean the skill they *build* will be the skill the job *asks for*. Figure 5 traces this along one simulated career in which the tool improves each period and nothing about the engineer changes: direction climbs throughout, while verification holds as long as checking pays and then decays once it does not.

A notable asymmetry results from the dynamics in this section: direction ability has a floor, while verification does not. Even an engineer with no direction skill benefits from a correct design, so, under the delegation floor of Assumption 4, delegating pays from the first day and the practice never stops, keeping the stock above a positive bound (Proposition 3(iii)). For verification, however, at $m_v = 0$ detection is zero (Assumption 1), so a worker with no skill will not catch anything, and thus has no incentive to check. That makes $m_v = 0$ an absorbing state of equation (15), and no rise in reliability will make a skill-less checker worth deploying.

Thus, verification *can* collapse to zero. A worker who starts below the threshold in Proposition 3(ii) will never build the skill, and, along the path in Figure 5, a worker who starts well above it loses it anyway once the tool is reliable enough that checking stops paying. The next section characterizes that collapse and the critical reliability at which the skilled state disappears, and shows that a worker who bears less than the full cost of their errors is brought closer to both.

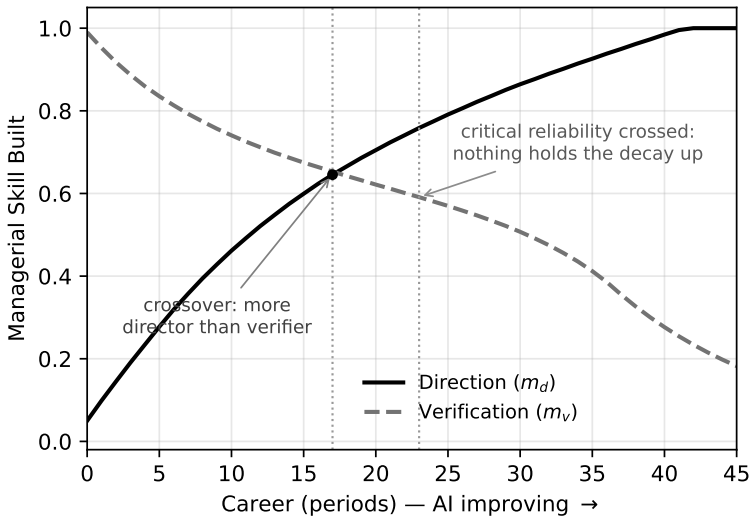



Figure 5: The skill the worker builds crosses over from verification to direction

*Note:* The managerial skills the worker builds as a byproduct of use, along one simulated career ($q$ and $\kappa$ rising with the period; one period is a year, decay $\eta = 0.08$ from Arthur et al. (1998), practice-to-decay ratio $\psi_v/\eta = 1.2$, stakes uniform on $(0, 3]$, scope $\kappa = 0.85$ in the other dynamic figures). Direction ($m_d$) climbs throughout. Verification ($m_v$), held while the tool still repays checking, then decays from disuse and collapses through the crossover. The drawn path is a property of the calibrated family: Proposition A2(iv) signs the steady-state response where its path conditions hold. The demand-side crossover of Figure 2 reappears here in supply.

# 5 The Accountability Wedge

So far, the engineer has paid in full: they internalize the entire loss $L$, so the delegation and vetting choices that maximize their private payoff are the ones the firm would choose itself. We now consider the implications of breaking that alignment by letting the engineer incur a fraction $\theta \in (0, 1]$ of the loss. We further assume that their vetting effort is unobservable, so the firm is unable to contract on it. In this setting, the engineer maximizes their private payoff, with $\theta L$ in place of $L$ throughout. We define the *accountability wedge* as the gap between these two valuations, the share $(1 - \theta)L$ of every shipped loss that the engineer does not internalize.

The accountability wedge has both a static consequence and a dynamic one. In a given period, it makes the engineer ship work the firm would have checked. Across multiple periods, it erodes the verification skill the engineer would otherwise have built by checking.

## 5.1 Over-Delegation and the Over-Reliance Band

The engineer's private vetting effort solves the same first-order condition as before, with $\theta L$ replacing $L$:

$$(1 - q)\big(V\varphi(m_d) + \theta L\big)\,\rho_e(m_v, e; q) = \lambda c'(e). \tag{16}$$

The distortion is visible in the condition itself. The worker still checks until the marginal value

of the errors they catch equals the marginal cost of their time, but the value they weigh is now $V\varphi + \theta L$ rather than $V\varphi + L$. The worker will therefore stop checking too early because their marginal benefit of effort falls short of the firm's by $(1-q)(1-\theta)L\,\rho_e$—the share of an uncaught error's cost they do not bear, weighted by the chance $(1-q)$ that there is an error to catch and the rate $\rho_e$ at which scrutiny turns into catches.

The engineer's private payoff from an action $\alpha$ on an in-scope task is

$$u_\theta(\alpha) \;=\; S(\alpha) \;+\; (1-\theta)\,D(\alpha), \tag{17}$$

where $S(\alpha)$ is its social value at the true $L$ and $D(\alpha)$ the expected loss it ships uncaught (Lemma 1). The engineer maximizes social value plus the share of the shipped loss that is not theirs, and every distortion that follows runs through that second term.

**Proposition 4** (Over-delegation)**.** *Fix an in-scope task and let $\theta < 1$; part* (iii) *imposes the Trust dominance of Assumption 3.* (i) Intensive margin. *The effort that maximizes the Vet branch, $e^w(\theta)$, is strictly increasing in $\theta$ wherever it is interior, and strictly below the first-best $e^* = e^w(1)$ whenever $m_v > 0$ and $L > 0$: the engineer under-vets whenever they vet.* (ii) Welfare. *Write $\alpha^\theta$ for an action the engineer chooses at accountability $\theta$. A smaller $\theta$ weakly raises both the expected shipped loss $D(\alpha^\theta)$ and the welfare loss* $\mathrm{WL}(\theta) = S(\alpha^1) - S(\alpha^\theta)$*, which is zero at $\theta = 1$ and at most the uninternalized share of the extra losses the wedge ships,*

$$\mathrm{WL}(\theta) \;\le\; (1-\theta)\big[D(\alpha^\theta) - D(\alpha^1)\big]. \tag{18}$$

(iii) The over-reliance band. *Under Assumption 3, because catching errors is worth less to the engineer ($\theta L < L$), the verification skill above which they vet, $m_v^w$, exceeds the firm's $m_v^{\mathrm{FB}}$ whenever the firm's lies in the unit interval. This partitions $m_v$ into three regimes: for $m_v < m_v^{\mathrm{FB}}$ both ship unchecked and* $\mathrm{WL} = 0$*; for $m_v \in [m_v^{\mathrm{FB}}, m_v^w)$ the firm would vet but they ship unchecked*—extensive-margin over-reliance*, where* WL *rises continuously from zero with $m_v$; and for $m_v \ge m_v^w$ both vet and only the intensive under-vetting of part* (i) *remains. The welfare loss drops discontinuously at $m_v^w$. (See Proposition A3 for the full statement and its hypotheses.)*

The static welfare loss concentrates on the middle band $[m_v^{\mathrm{FB}}, m_v^w]$: the competent-but-not-expert workers for whom the missing stake is enough to tip them into shipping the machine's output unchecked (Figure 6). Empirical work has shown that workers stop checking a tool they trust, and that holding them accountable for the outcome, in our case raising $\theta$, brings the checking back. Parasuraman and Manzey (2010) synthesize decades of studies on human oversight of automation and find both. Dell'Acqua et al. (2026) find the knowledge-worker version, in which those who ship the tool's output unchecked fall behind.

Over a portfolio whose stakes are unimodal in logs, the excess rate of unchecked shipping under the wedge is single-peaked in verification skill (Lemma 2, in appendix A).

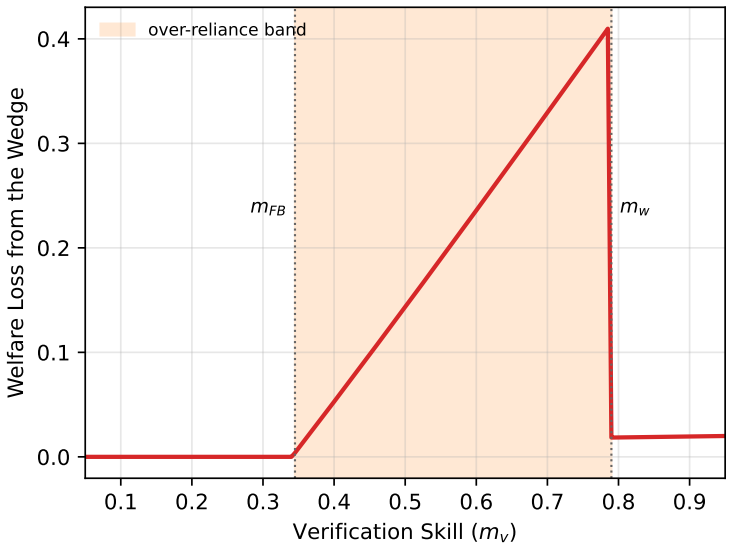



Figure 6: The over-reliance band

*Note:* Per-task welfare loss from the wedge against verification skill $m_v$. It is zero below $m_v^{\mathrm{FB}}$, where neither vets; rises across the band $[m_v^{\mathrm{FB}}, m_v^w]$, where the firm would vet but the worker ships unchecked, as the firm's detection (and so the gap) grows with skill; peaks approaching $m_v^w$; and drops to the small intensive-margin gap once the worker also vets. Parametric values $a = 0.5$, $q = 0.6$, $L = 3$, $\theta = 0.3$, setup cost $f = 1$.

## 5.2 Lopsided Formation and the Deskilling Trap

Importantly, this over-reliance of Section 5.1 is a single-period distortion. Carried into the skill formation of Section 4, the same missing stake bends the path of skill formation itself. Under full accountability, both skills build, and neither is discounted (Proposition 3(i)–(ii)). The wedge falls on only one of the two skills: direction raises the value $V\varphi$ they are paid for in full, while catching a flaw saves a loss of which they keep only $\theta$, so verification is under-built. That shortfall compounds. As the worker's verification ability declines, they find fewer flaws, which makes checking pay less, so they check even less, and the setup cost traps them at the bottom.

AI use therefore builds a *lopsided* manager, developed as a director but underdeveloped as a verifier. The decay of an unpracticed skill is well documented, and not unique to our setting. Arthur et al. (1998), in a meta-analysis of skill retention, find the loss sharpest for exactly the cognitive, accuracy-oriented skills that verification depends upon. Proposition 5 formalizes both halves, and adds the reliability at which the skilled state stops existing altogether.[19]

**Proposition 5** (Lopsided formation and the deskilling trap)**.** *Let $\theta < 1$.* (i) No direct wedge on direction. *Accountability discounts only the loss terms, so the wedge only affects direction indirectly by discouraging checking.* (ii) Verification under-built. *Under Assumption 3, the worker prices a*

[19]The free-repair assumption of equation (3) plays an important role here. A repair cost $r > 0$ charged on each caught error would lower the worker's vetting effort and, with it, their vetting practice $x_v$, wherever they vet and nowhere else. As a result, the trap's basin would widen, and the skilled state would vanish at a lower critical reliability. The deskilling results presented in this section should be viewed as conservative to the extent that workers face repair costs in practice.

*caught error at $V\varphi + \theta L$ while the firm prices it at $V\varphi + L$, so vetting effort, verification practice, and its steady-state stock all fall below their first-best levels, and are nondecreasing in $\theta$.* (iii) The trap. *For any setup cost $f > 0$, $m_v = 0$ is an absorbing state, with a trap floor $m_v^0 > 0$ below which verification always decays. In settings where practice builds skill slowly at low stocks and fast at high ones, verification has at most two resting points, an unstable threshold $m_v^\dagger$ and a stable skilled state $m_v^*$. In other settings, the model can produce additional stable steady states at intermediate skill levels. Under Assumption 3, lowering $\theta$ raises the threshold and lowers the skilled state.* (iv) The critical reliability. *If verification practice falls at every skill level as reliability rises, the skilled state and the threshold merge and vanish at a critical reliability $q_{\text{crit}}$, and past it no skilled steady state survives. (See Proposition A4 for the full statement and its hypotheses.)*

In Figure 7, we plot the steady state of the worker's verification skill, $m_v^*$, as a function of AI reliability. As the tool improves, verification drifts down, and then, at the "critical reliability" $q_{\text{crit}}$, the skilled steady state vanishes and verification quickly falls to zero. This collapse is irreversible, and the smaller $\theta$, the (weakly) sooner it arrives. Because the engineer's payoffs depend on accountability and stakes only through their product $\theta L$, a half-accountable worker on a double-stakes job has the same critical reliability as a fully accountable worker at base stakes. Importantly, direction has no critical reliability, since under the delegation floor of Assumption 4, the practice never stops (Proposition 3(iii)).[20]

### 5.3 When It Matters: The Rare-Catastrophic Regime

The skill collapse we derived in Proposition 5 matters in settings where the engineer may later need the lapsed skill to catch a particularly costly flaw. In practice, this may not be a significant concern. Most of the engineer's portfolio is routine, and if a small flaw ships, it would either be caught downstream or the cost would be negligible. Some tasks, however, are not routine or their cost is not negligible. Consider a brake controller, where a shipped flaw could be fatal, or an authentication path, where one could expose every user's data.

In the engineer's terms: as the AI model grows more reliable, checking routine tasks is no longer worth it, so they stop, and the verification skill built on that routine work decays into the trap (Proposition 9, in appendix B.1). When the brake controller arrives, they check it because the stakes clearly require it, but with their diminished verification ability, a flaw the AI tool introduced ships.[21] Ultimately, the failure is one of skill, not diligence.

This concern is not without precedent. Bainbridge (1983) identifies it as a primary "iron[y] of automation," in which the machine takes over the routine work and leaves the operator the

[20] We hold direction fixed here. If instead we allow it to vary, a better director delegates more, and so has more output worth checking, which sustains their verification practice. A worker with more direction skill therefore needs less verification skill of their own to stay clear of the trap. In our parameterized example the effect is small, and whether they escape remains almost entirely a matter of verification skill (Corollary 2).

[21] The static analysis of Section 5.1 understates the danger here. Its over-reliance band closes at catastrophic stakes, since even an engineer who bears a fraction of the loss still checks a brake controller (though perhaps not as hard as the firm would like). Within a single period, then, the high-stakes task is safe. Across periods it is not, because by the time it arrives their verification skill may have already decayed.

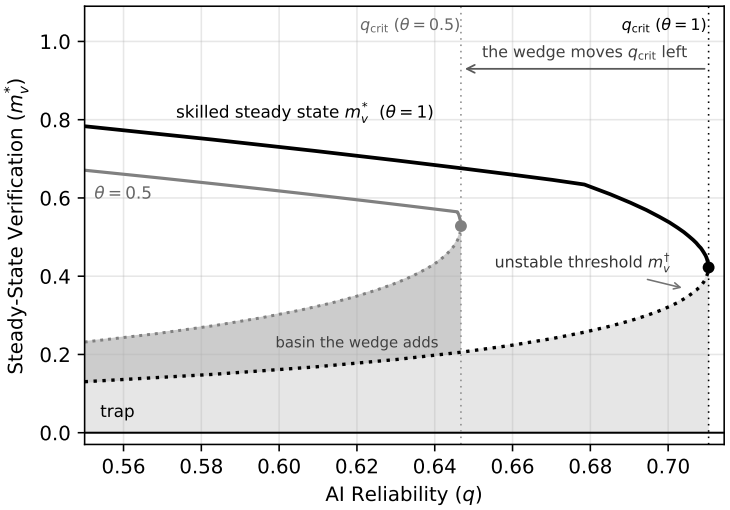



Figure 7: The verification trap: a point of no return

*Note:* Steady-state verification skill against AI reliability. The solid curve is the skilled steady state, the axis at zero is the trap, and the dotted curve between them is the unstable threshold separating their basins. A worker whose skill starts below the threshold decays to the trap, and one who starts anywhere above it converges to the skilled steady state, from below or from above. As reliability rises, the skilled steady state and the threshold meet at the critical reliability (vertical lines), beyond which no skilled state survives. The wedge ($\theta < 1$, in gray) lowers the skilled steady state, raises the threshold, and moves the critical reliability left. The darker region is the basin the wedge adds to the trap. Direction, under the delegation floor of Assumption 4, has no analogue. Calibration as in Figure 5.

rare abnormal case, while "physical skills deteriorate when they are not used." There is empirical support for this in high-stakes settings: Casner et al. (2014) find that airline pilots' higher-level flying skills (e.g., navigating without automation, diagnosing a failing instrument) atrophy under routine autopilot use, so that when a rare failure is staged, pilots are more likely to trust the faulty reading. Agarwal et al. (2023) find that AI assistance fails to raise radiologists' diagnostic accuracy because they neglect or misweight it. Our model provides a mechanism that connects this skill erosion to the tool's own improvement and to the share of the loss the worker does not internalize. This is not an abstract concern, as AI models rapidly improve with each iteration, and the pace of that improvement has itself roughly doubled since April 2024 (Edelman and Lee, 2025).

### 5.4 Hiring and Accountability

The firm has two ways of dealing with the accountability wedge. The first is hiring. In the static model, the firm hires a worker of a fixed type. In the dynamic model, it hires a worker whose skills will evolve with their AI use. Because the law of motion is bistable, where a hire ends up depends on their initial skills, and not on their output at the time of hire (Corollary 2, proved in appendix A.5).

**Corollary 2** (Hiring against the trap)**.** *Let the two-skill law of motion* (15) *be monotone, as the practice map of Proposition 3 provides, with skilled steady state* $m_v^* > 0$*.* (i) The hiring bar. *There is a hiring bar in verification,* $\sigma(m_d)$*, that weakly falls as direction skill rises, so that a hire above it*

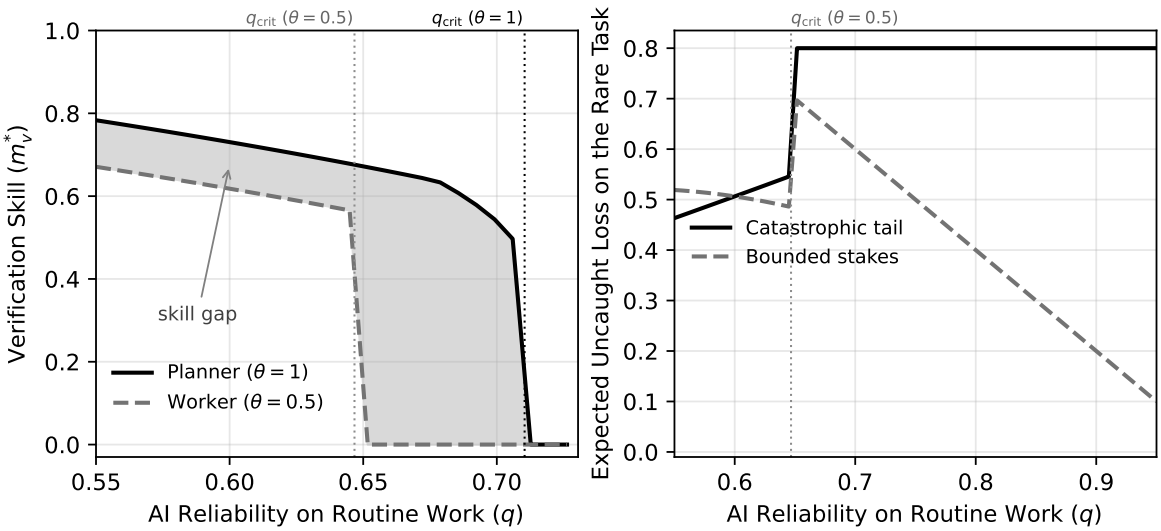



Figure 8: Verification skill atrophies on routine work; the rare catastrophe is what it leaves uncaught

*Note:* Left: steady-state verification skill against the AI's reliability on the routine bulk that builds it. As routine errors grow rare the skill atrophies into the trap; at $\theta = 0.5$ the worker's high branch ends at a lower reliability than the full-stakes benchmark's ($\theta = 1$), opening the shaded skill gap. Right: the engineer's expected uncaught loss on a rare task. For bounded stakes it vanishes as the AI improves; for a catastrophic tail (stakes scaling so that $(1 - q)L$ is fixed) it persists in full, because their lapsed skill can no longer catch the rare error. Calibration as in Figure 5.

*keeps a positive verification stock and a hire below it loses it.* (ii) How the bar moves. *The hiring bar always lies between the bar that applies to a worker with full direction skill and the bar that applies to a worker with none. Both bounds rise as accountability falls, and both rise with reliability wherever a steadier tool lowers verification practice. (See Proposition A5 for a more complete characterization of this result.)*

Those two levels are what a firm screens on when hiring. A worker who clears both levels keeps their verification skill whatever their direction skill, and one who clears neither loses it. In between, the outcome depends on both skills together. The result inverts the perhaps natural intuition that a better tool lets a firm lower its hiring bar. Instead, a more reliable AI means there are fewer errors to catch and thus workers engage in less verification practice, so, above a low-reliability trough, it *raises* the bar. Importantly, though, at the critical reliability threshold, the skilled steady state disappears, so even a worker hired with full verification skill deskills, and no initial skill is enough to prevent it. Selection therefore falls on one managerial skill far more than the other: verification the firm must bring in, because a worker below the bar loses it for good; direction it can leave to the job.

The second is to change the wedge itself. Thus far, we have assumed that $\theta$ is a fixed feature of the environment. But in reality, the firm chooses an *accountability intensity* $\theta$ itself, through sign-off requirements, review gates, and the share of a shipped loss charged back to the engineer. Such policies, however, are costly, as in any principal-agent setting (Holmström, 1979), in both the

time they take and the apparatus they require. A higher $\theta$ lifts the worker's vetting effort and their practice at verification, moving both toward the first best of Proposition 4.

How much accountability a firm buys depends on what it values. A *myopic* firm will only consider the current period: it will tighten accountability until the cost of tightening further would exceed the shipped-error loss it saves, taking the worker's verification skill as given. A forward-looking firm would also consider the skills built under the level of accountability it chooses. We compare the two in our parameterized example. Away from the critical reliability threshold, the forward-looking firm's accountability weakly exceeds the myopic firm's. Near the threshold the forward-looking firm's objective splits in two. It can set accountability high enough to maintain the worker's verification skill, or it can let that skill collapse and stop paying for accountability at all, since once the worker completely loses their verification skill, they'll never vet and so there is no benefit to further tightening accountability. The surplus the skill earns shrinks as the tool improves while at the same time the accountability needed to sustain it rises. As a result, past a reliability threshold of its own the firm will rationally let the skill collapse.[22] Of course, a firm that sets $\theta$ low will pay for it later, while today's output will be unaffected since it reflects the verification skill the worker already has. The worker's skill will slowly erode over the periods that follow, and the firm will ultimately bear the cost when a rare and costly error goes through uncaught.[23]

# 6 The Market for Managers: Sorting and Wages

In the last section, we analyzed how a single firm decides whom to hire in the presence of the deskilling trap. But most firms are not monopsonists, and so they hire from a shared labor market. We now ask how that market sorts workers across jobs, and what it pays them.

A firm filling a position ranks its candidates by an index $W$ of what each would produce across the job's portfolio of tasks. We define this as

$$W(a, m_d, m_v; q, \kappa) = \mathbb{E}_{(d,L)}\big[\max\{u_{\text{Solo}},\, u_{\text{Trust}},\, u_{\text{Vet}}(e^*)\}\big] = \mathbb{E}_{(d,L)}[u_{\text{Solo}}] + \mathbb{E}_{(d,L)}[R], \tag{19}$$

their Solo value plus the expected return $\mathbb{E}[R]$ AI adds. We define the firm's "screening gradient" to be the three returns $(\partial W/\partial a,\ \partial W/\partial m_d,\ \partial W/\partial m_v)$, capturing how much a candidate is worth with a little more of each skill. Without AI it is almost all task skill, but with AI, the firm has a second thing to screen for. As the tool improves, the return to task skill falls while the "trusted floor" under direction's, its return when no delegated work is checked, does not, so the whole vector turns away from task skill and toward managerial ability. That turn is what we mean by a "rotation"

[22] This calculation also assumes the firm expects to keep the worker indefinitely. Verification is a general skill—an engineer who can spot a wing defect can do so at Boeing or at Airbus—and so a firm that expects to lose a worker has even less of a reason to pay for it (Becker, 1964). Note, however, that Acemoglu and Pischke (1998) show wage compression and search frictions can restore the incentive.

[23] Bastani and Cachon (2026) reach a similar place by a different route. Their firm pays a wage to motivate inspection, and under limited liability that wage rises without bound as the tool grows more reliable, so it often gives up on human oversight altogether. Ours buys too little accountability not because accountability is expensive but because the return to it arrives after the decision is made.

of the screen, and we establish it in the following proposition.

**Proposition 6** (Screening rotation)**.** *Rank candidates for one position by $W$ in equation* (19)*, with $(d, L)$ drawn from the portfolio distribution of Section 2.* (i) Task skill. *The return to task skill is the stakes-weighted mass of tasks the worker still executes Solo: strictly positive whenever some tasks lie beyond the tool's reach ($\kappa < 1$), and weakly falling as scope widens, at any joint distribution.* (ii) Managerial ability. *Both managerial returns are weakly positive—direction earning on the delegated mass through the ship-correct probability $q_{\text{sc}}$, verification on the vetted mass through the errors it catches—and, at interior reliability, each is strictly positive exactly when its mass is.* (iii) Rotation. *Along any path on which reliability and scope rise together and the vetted-value condition* (12) *holds task by task, the task-skill return is nonincreasing while the trusted floor of the direction return is nondecreasing: the screen turns from task skill toward managerial ability, in the bounded sense Proposition A6 makes exact, with its verification component following the crossover of Proposition 2.*

We now consider the workers of Section 3 in the context of a competitive labor market. There is a continuum of them, each with an innate managerial type $m = (m_d, m_v)$, and a continuum of jobs to fill. We hold task skill $a$ common across workers and treat it as a feature of the job rather than the type, so that workers sort on managerial ability alone.[24] We define a job as a pair $(\kappa, \ell)$: how much of its work the tool can complete, and a stakes scale $\ell > 0$ that multiplies a fixed base distribution of stakes, $L = \ell\, L_0$ with $L_0 \sim F_0$. A high-$\ell$ job is a high-stakes one (medicine, aviation, etc.) and vice versa. For this exercise, we assume every job uses the same AI tool, so reliability $q$ is common across jobs (though scope is not). Utility is transferable and the assignment is competitive, so the surplus a type-$m$ worker creates in job $(\kappa, \ell)$ is the hiring value (19), now the output of a match rather than one firm's ranking criterion.

### 6.1 What the Market Pays for Each Skill

In a competitive assignment, the outcome is the surplus-maximizing matching, supported by a wage $w(m)$ for the worker and a residual value $r(\kappa, \ell)$ for the firm that sum to the surplus on every realized match (Becker, 1973; Sattinger, 1993). A worker takes the best terms available, so their wage solves $w(m) = \max_{(\kappa,\ell)}\{W(m; \kappa, \ell) - r(\kappa, \ell)\}$. Where a worker is matched to a single job, that wage is differentiable, and its gradient is the screening gradient of Proposition 6, evaluated at the job they hold:

$$\nabla w(m) = \nabla_m W\big(m; \kappa(m), \ell(m)\big). \tag{20}$$

The market therefore pays for each skill exactly what that skill adds to the value of a worker's match. Because task skill is common across the workers in this market, equation (20) prices the two managerial skills and nothing else, and the rotation in Proposition 6 is a statement about how a firm ranks candidates, not about what anyone is paid.

[24] Restoring dispersion in $a$ would add a third sorting dimension, which we do not consider here. Recall that neither managerial return depends on $a$, although the Solo value does.

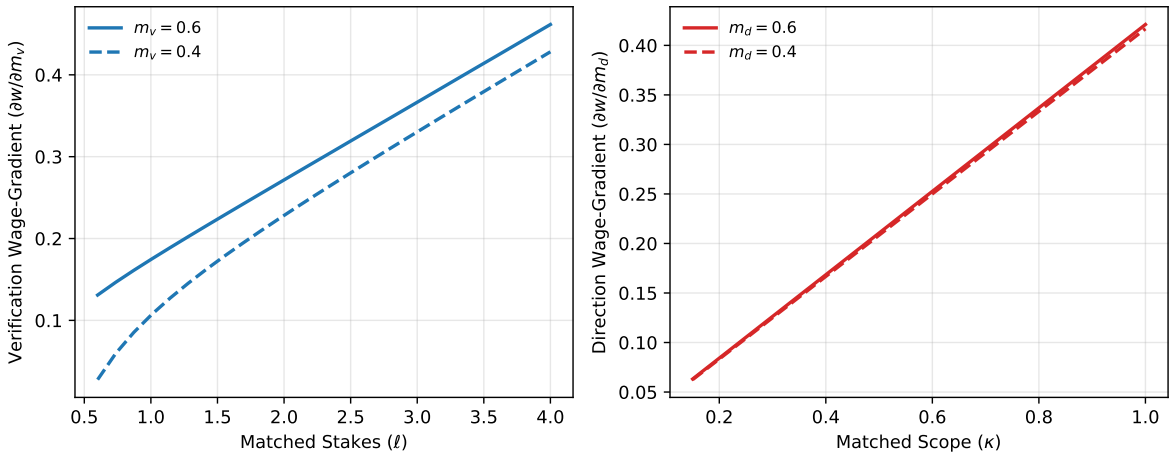



Figure 9: The two sorting premia

*Note:* By equation (20) the equilibrium wage gradient equals the screening gradient $\partial W/\partial m$ at the matched job. Left: the verification premium $\partial w/\partial m_v$ rises with the job's stakes scale $\ell$. Right: the direction premium $\partial w/\partial m_d$ rises with the job's AI scope $\kappa$. Reliability held at $q = 0.6$, common task skill $a = 0.5$.

Verification pays off in proportion to the stakes (Proposition 1(ii)), so pay rises faster with verification skill in higher-stakes jobs. Direction pays off on the tasks the worker delegates (Proposition 2), so pay rises faster with direction skill in higher-scope jobs (Figure 9). We call these two prices the "verification premium" and the "direction premium."

## 6.2 Sorting Workers to Jobs

If workers differed in one skill only, the assignment would be trivial: rank the workers, rank the jobs, match them in order. Our workers differ along two dimensions and there is no clear way to order a worker who directs better and verifies worse than one who does the reverse. Which job a worker is best suited for depends on how the job's stakes and scope interact with each skill. There are four effects to consider. Two of them are the premia of Figure 9: scope raises the value of direction, and stakes raise the value of verification. The other two are why the match is not obvious. Stakes also raise the value of direction, and scope also raises the value of verification, and if either does so strongly enough, the tidy pairing of each skill to its own attribute breaks.

We impose Assumption 3 on the portfolio of every job in the market, so that the worker never does an in-scope task Solo. The match surplus then separates:

$$W(m;\kappa,\ell) = \mathbb{E}_L\big[u_{\text{Solo}}\big] + F_D(\kappa)\, g(m,\ell), \qquad g(m,\ell) = \mathbb{E}_{L_0}\big[R(m;\ell L_0)\big], \tag{21}$$

so the worker's type enters only through $g$, their return on a single in-scope task. All four effects are then nonnegative: each skill gains from both of a job's attributes. The following proposition works out who ends up in which job.

**Proposition 7** (Sorting)**.** *Impose Assumption 1, Assumption 2, and Assumption 3 on the portfolio*

*of every job in the market.* (i) Jobs that differ in one way. *Across jobs sharing a scope, the strongest verifiers among workers of equal direction skill take the highest-stakes jobs; across jobs sharing a stakes scale, the strongest directors among workers of equal verification skill take the widest scope. This holds in some surplus-maximizing assignment always, and in every one wherever the relevant complementarity is strictly positive between the matches.* (ii) Jobs that differ in both ways. *Take two workers, one better at both skills than the other. In any surplus-maximizing assignment, the better worker's job cannot be lower in both scope and stakes, ties aside.* (iii) The end of vetting. *Each match has a reliability threshold $\bar{q}(m, \ell)$ which is nondecreasing in the job's stakes and the worker's verification skill. Past that threshold, the worker does not vet and so verification is not valued by the firm. Once the AI tool's reliability passes the largest such threshold in the market, the assignment problem collapses to a one-dimensional matching of direction against scope. (See Proposition A7 for the full statement and its hypotheses.)*

A wider scope means the worker has more tasks to delegate, so it raises the return to both skills at once, not favoring one over the other. Higher stakes raise the value of every caught error, affecting the return to verification directly, and direction to a lesser degree. For example, consider two workers, one who is the better director and another who is the better verifier, along with two jobs, one with a wider scope and the other with higher stakes. Whether each worker gets the job that matches their strength depends on how large each worker's advantage is and how "far apart" the two jobs are (38).

While the return to direction overtakes that of verification at the point of the Proposition 2 crossover, the worker does not stop checking at that point. Between the crossover and the end of vetting, directing earns the worker more than checking, but checking still provides some benefit, so the market keeps paying for verification. In our parameterized example the crossover sits at 0.69, while vetting ends at 0.83 at the central calibration and at 0.945 for the highest-stakes job held by the strongest verifier.

### 6.3 The Wedge's Cost Falls on the High-Stakes Jobs

In equilibrium, the market sends stronger verifiers, at a given direction skill, to higher-stakes jobs, which is where Proposition 9 places the catastrophic uncaught loss and where the accountability wedge of Section 5 does the most damage. Re-introducing $\theta < 1$, the sort survives: in our parameterized example the pairing of verification to stakes weakens but does not reverse, and the cost of the wedge, though not monotone in stakes, concentrates in the highest-stakes jobs (Corollary 3, in appendix B.2).

Since we consider only a spot market, each match is priced correctly given how the worker behaves in the moment, but no one pays for the practice a job would provide. If the firm could retain a worker long enough, they would be willing to pay for the vetting a high-stakes job requires, since that vetting would keep verification above the trap threshold ($m_v^\dagger$ of Proposition 5). Pricing that would require accounting for match duration, worker mobility, and a firm horizon, which we do not consider here.

This sorting result is partly self-correcting. A higher-stakes job will induce more vetting, so the skill lasts longer there. At the parameter values used for Figure 7, the critical reliability under the wedge is 0.797 in a job at four times the base stakes, against 0.647 at the base. The market thus naturally concentrates practice in the higher-stakes jobs. But as the tool improves, routine checking will eventually stop being worth its cost in every job, the high-stakes ones last, and so when the tool passes their critical reliability the collapse will end up where a missed error costs the most.

# 7 Concluding Discussion

When a knowledge worker is given an AI tool, they suddenly take on a managerial role. We model the worker's managerial ability along two dimensions, direction and verification, and characterize the tool by its reliability and its scope. The worker's two abilities are paid on different margins, direction on the work the tool can reach and verification on the cost of the work it gets wrong, and as the tool grows more reliable the higher-return ability crosses over from verification to direction (Proposition 2).

Skills, though, are not fixed, and to the extent that exercising the skills improves the worker's ability, checking will build verification and delegation will build direction. Using the tool therefore builds the manager it requires naturally and in the ordinary course of the work rather than through deliberate training. However, if the worker is insulated from the consequences of the mistakes they ship, AI use will build a lopsided manager. Indeed, once the tool passes a critical reliability the checking stops being worth its cost and the verification skill will atrophy (Proposition 5).

Whether that decay is consequential is a property of the setting. Across most of a firm's portfolio it is likely to be benign, since the work is routine and a flaw that ships costs little. The danger is concentrated where errors are rare and catastrophic, because the stakes there rise as the errors thin out.

Finally, we have shown that the market will tolerate this skill collapse, pricing it rather than prohibiting it. Verification will earn its premium where the stakes are high and direction where the tool's reach is wide. The strongest verifiers will end up in the jobs in which a missed error costs the most (Proposition 7). While a high-stakes job induces more checking, holding the skill longer than a low-stakes one would, this only defers the eventual collapse.

Much of our analysis rests on the assumption that catching the tool's mistakes is independent of the worker's task skill, so that a worker can recognize an answer is wrong even in cases where they cannot produce the correct response themselves. There is empirical support for this notion. For example, Riedl and Weidmann (2025) estimate an AI-collaboration ability that is only weakly correlated with solo task skill. Of course, this separation is not necessarily reasonable in every setting. Where checking is essentially re-derivation, verification decays along with the task skill it borrows from and the collapse occurs sooner than our model says. We additionally assume that attention is priced rather than budgeted, so that tasks are separable (see appendix B.3 for the coupled version), and that managerial ability is fully a worker-specific characteristic, which ignores

the important role an organization's own practices play in how productive workers are and what roles they sort into (Bloom et al., 2017; Garicano and Rossi-Hansberg, 2006).

We also assume throughout our analysis that workers are myopic, when in practice they may be forward-looking—anticipating how their work today will improve their human capital tomorrow. The deskilling trap we derive is robust to forward-looking workers because, in our parameterized example, the collapse arrives well before checking becomes worthless. As AI tooling improves, what runs out is not the workers' desire to check but the number of tasks worth checking, so while foresight may move the point at which the skill collapses, it does not remove it.

Our analysis has clear relevance for hiring firms. Firms have long screened entry-level knowledge workers on task skill: i.e., can this candidate write the code, read the scan, draft the memo, etc. Our results suggest that this is now an incomplete screen, since a worker handed an AI tool is doing managerial work from day one, and that firms should also screen for managerial skills. While screening on direction can provide some initial benefit, direction can be at least partially left to the job, since its practice never stops and an initially weak director will improve over time. In the presence of AI, screening for verification, however, becomes an essential part of the hiring process because a hire below the bar loses the skill for good and no on-the-job exposure rebuilds it (Corollary 2). This finding contradicts the typical intuition that as AI improves, firms will be able to relax their hiring standards. Instead, a more reliable AI will result in fewer errors to catch and therefore less practice, raising the bar rather than lowering it.

Our work additionally motivates a number of empirical questions. The most immediate is how much a period of practice adds to each managerial ability, and how fast an unpracticed one decays. A more ambitious question is to what extent the two skill premia are present in the cross-section of wages, as our model predicts steeper pay gradients in verification in high-stakes jobs and in direction in high-scope ones. Our model also predicts how these premia should change over time: the verification premium should fall as the tool becomes more reliable, while the direction premium should rise as its scope widens. Moreover, as AI adoption increases within and across industries, both premia should appear in a growing share of jobs while the gradient in task skill, which firms have traditionally screened new hires on, should flatten. We leave it to future work to test these predictions using wage data spanning multiple generations of AI models. Finally, since verification skill is built on routine work, and the tool is taking over much of that work, where the next cohort of verifiers will come from is an open question.

From a policy perspective, liability rules, professional sign-off requirements, and internal review gates all raise the share of an error's cost that lands on the worker who shipped it, and each moves the critical reliability at which the checking skill collapses. While such policies can help delay the collapse, they cannot fully prevent it, since higher accountability moves the critical reliability threshold but does not eliminate it. The more useful thing the model offers a regulator is a way to identify which settings are most at risk. The cost this collapse imposes on a firm is the product of how deskilled its workers are and what its errors cost. While the first piece is unobservable, it is likely to be common across firms facing the same incentives and using the same AI tooling. The

second piece can be calculated from how often the tool is wrong and what a wrong answer costs.

We opened this paper with a passage from Drucker's 1954 treatise on management. His claim was that the new technology of his day would not make managers superfluous but would instead spur even greater demand for them, requiring people counted until then as rank-and-file to think managerially. This is what is playing out now, as AI becomes an increasingly pervasive tool within firms. It has not, however, followed the route he described: the demand arrived not because firms created managerial roles and filled them, but because this new tool has turned ordinary work into delegation. Largely without anyone deciding it, everyone has become a manager in Drucker's sense, and almost no one was hired as one.

## References


**Acemoglu, Daron and Jörn-Steffen Pischke**, "Why Do Firms Train? Theory and Evidence," *The Quarterly Journal of Economics*, 1998, *113* (1), 79–119.

**_ , Dingwen Kong, and Asuman Ozdaglar**, "Artificial Intelligence, Human Cognition, and Knowledge Collapse," NBER Working Paper 34910, National Bureau of Economic Research 2026.

**Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz**, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," NBER Working Paper 31422, National Bureau of Economic Research 2023.

**Aghion, Philippe and Jean Tirole**, "Formal and Real Authority in Organizations," *Journal of Political Economy*, 1997, *105* (1), 1–29.

**Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb**, "Prediction, Judgment, and Complexity: A Theory of Decision-Making and Artificial Intelligence," NBER Working Paper 24243, National Bureau of Economic Research 2018.

**Alekseeva, Liudmila, José Azar, Mireia Giné, and Sampsa Samila**, "Artificial Intelligence Adoption and the Demand for Managerial Expertise," *Strategic Management Journal*, 2026.

**Arthur, Winfred, Winston Bennett, Pamela L. Stanush, and Theresa L. McNelly**, "Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis," *Human Performance*, 1998, *11* (1), 57–101.

**Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans**, "The Allocation of Decision Authority to Human and Artificial Intelligence," *AEA Papers and Proceedings*, 2020, *110*, 80–84.

**Babina, Tania, Alex X. He, and Renhao Jiang**, "Canaries in the Gold Mine: Early Productivity Gains from Artificial Intelligence Creating Organization Capital," Working Paper 35684, National Bureau of Economic Research 2026.

**Bainbridge, Lisanne**, "Ironies of Automation," *Automatica*, 1983, *19* (6), 775–779.

**Bastani, Hamsa and Gerard P. Cachon**, "The Human-AI Contracting Paradox," SSRN Working Paper 5962739 2026.

**Becker, Gary S.**, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, New York: National Bureau of Economic Research, 1964.

_ , "A Theory of Marriage: Part I," *Journal of Political Economy*, 1973, *81* (4), 813–846.

**Bick, Alexander, Adam Blandin, David J. Deming, Nicola Fuchs-Schündeln, and Jonas Jessen**, "Mind the Gap: AI Adoption in Europe and the U.S.," Working Paper 34995, National Bureau of Economic Research March 2026. Prepared for the Brookings Papers on Economic Activity, Spring 2026.

**Bloom, Nicholas, Raffaella Sadun, and John Van Reenen**, "Management as a Technology?," NBER Working Paper 22327, National Bureau of Economic Research 2017.

**Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond**, "Generative AI at Work," *Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

**Caplin, Andrew, David J. Deming, Shangwen Li, Daniel J. Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye**, "The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration," NBER Working Paper 33021, National Bureau of Economic Research 2024.

**Casner, Stephen M., Richard W. Geven, Matthias P. Recker, and Jonathan W. Schooler**, "The Retention of Manual Flying Skills in the Automated Cockpit," *Human Factors*, 2014, *56* (8), 1506–1516.

**Catalini, Christian, Xiang Hui, and Jane Wu**, "Some Simple Economics of AGI," arXiv:2602.20946 2026.

**Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine Kellogg et al.**, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Organization Science*, 2026.

**Dessein, Wouter**, "Authority and Communication in Organizations," *Review of Economic Studies*, 2002, *69* (4), 811–838.

**Drucker, Peter F.**, *The Practice of Management*, New York: Harper & Row, 1954.

**Edelman, Yafah and Jaeho Lee**, "AI Capabilities Progress Has Sped Up," Epoch AI, Data Insights 2025. `https://epoch.ai/data-insights/ai-capabilities-progress-has-sped-up`.

**Ellis, Keaton and Wanying Huang**, "Should I State or Should I Show? Aligning AI with Human Preferences," Working paper, Monash University 2026.

**Garicano, Luis**, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 2000, *108* (5), 874–904.

_ **and Esteban Rossi-Hansberg**, "Organization and Inequality in a Knowledge Economy," *Quarterly Journal of Economics*, 2006, *121* (4), 1383–1435.

_ **, Jin Li, and Yanhui Wu**, "Weak Bundle, Strong Bundle: How AI Redraws Job Boundaries," CEPR Discussion Paper 21453, Centre for Economic Policy Research 2026.

**Holmström, Bengt**, "Moral Hazard and Observability," *The Bell Journal of Economics*, 1979, *10* (1), 74–91.

_ **and Paul Milgrom**, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 1991, *7*, 24–52.

**Huang, Lingxiao, Wenyang Xiao, and Nisheeth K. Vishnoi**, "Delegation and Verification under AI," arXiv:2603.02961; ICML 2026 2026.

**Ide, Enrique and Eduard Talamàs**, "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy*, 2025, *133* (12), 3762–3800.

**Johnston, Drew, David Holtz, Alex Martin Richmond, Christopher Ong, Prasanna Tambe, and Aaron Chatterji**, "The Shift to Agentic AI: Evidence from Codex," Working Paper, OpenAI 2026.

**Milgrom, Paul R.**, "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 1981, *12* (2), 380–391.

**Noy, Shakked and Whitney Zhang**, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

**Otis, Nicholas G., Rowan Philip Clarke, Solene Delecourt, David Holtz, and Rembrand Koning**, "The Uneven Impact of Generative AI on Entrepreneurial Performance," Technical Report, Working paper 2024.

**Parasuraman, Raja and Dietrich H. Manzey**, "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors*, 2010, *52* (3), 381–410.

**Quispe, Alexander and Kevin Xu**, "Agentic Delegation and the Language Frontier of Software Developers: A Model and Evidence from Claude Code on GitHub," arXiv:2605.25438 2026.

**Riedl, Christoph and Ben Weidmann**, "Quantifying Human-AI Synergy," Technical Report, Working paper 2025.

**Rosen, Sherwin**, "Authority, Control, and the Distribution of Earnings," *Bell Journal of Economics*, 1982, *13* (2), 311–323.

**Sattinger, Michael**, "Assignment Models of the Distribution of Earnings," *Journal of Economic Literature*, 1993, *31* (2), 831–880.

**Scur, Daniela, Raffaella Sadun, John Van Reenen, Renata Lemos, and Nicholas Bloom**, "The World Management Survey at 18: Lessons and the Way Forward," *Oxford Review of Economic Policy*, 2021, *37* (2), 231–258.

**Shen, Judy Hanwen and Alex Tamkin**, "How AI Impacts Skill Formation," 2026.

**Siderius, James, Robert A. Shumsky, and Alva H. Taylor**, "Use it or Slowly Lose it: Expertise Atrophy with Organizational AI Usage," SSRN Working Paper 6398398, Tuck School of Business, Dartmouth 2026.

**Weidmann, Ben, Joseph Vecci, Farah Said, David J. Deming, and Sonia R. Bhalotra**, "How Do You Find a Good Manager?," NBER Working Paper 32699, National Bureau of Economic Research 2024.

**_ , Yixian Xu, and David J. Deming**, "Measuring Human Leadership Skills with AI Agents," NBER Working Paper 33662, National Bureau of Economic Research 2025.

**Xu, Fasheng, Jing Hou, Wei Chen, and Karen Xie**, "Generative AI and Organizational Structure in the Knowledge Economy," arXiv:2506.00532 (rev. March 2026); SSRN 5242296 2026.

# Appendices

## A Assumptions and Proofs of the Results in the Text

### A.1 Assumptions

We state here the conditions Section 2 describes in words. Assumption 1 is maintained throughout. The remaining three are invoked by name where they are used.

**Assumption 1** (Detection technology)**.** $\rho(m_v, e; q)$ takes values in $[0, 1]$, is twice continuously differentiable, nondecreasing in $m_v$ and in effort $e$, with strictly positive partials $\rho_{m_v} > 0$ and $\rho_e > 0$ where $m_v > 0$ and $e > 0$, concave in $e$, and nonincreasing in reliability $q$, with cross-partials $\rho_{m_v q} \leq 0$, $\rho_{eq} \leq 0$, and $\rho_{m_v e} \geq 0$, and with $\rho(m_v, 0; q) = 0$ and $\rho(0, e; q) = 0$.

**Assumption 2** (Independent difficulty and stakes)**.** Across the portfolio, difficulty $d$ and stakes $L$ are independently distributed.

**Assumption 3** (Trust dominance)**.** On the tasks in question,

$$qV\varphi(m_d) + \lambda(t_H - t_A) \;\geq\; aV \qquad \text{and} \qquad u_{\text{Trust}} \;\geq\; u_{\text{Solo}} \text{ at } L = L_{\max}. \tag{22}$$

The condition loosens as direction skill rises, since $\varphi' > 0$ while the $L$-slope of the gap carries no $m_d$, and as reliability rises, since both inequalities in equation (22) are increasing in $q$. Imposing it at the lowest $m_d$ and the lowest $q$ a result visits therefore covers the rest, which is how Proposition 2 invokes it along a path in $q$.

**Assumption 4** (Delegation floor)**.** Write $\varphi_0 \equiv \varphi(0) > 0$. At $m_d = 0$, delegation still beats Solo on a set of tasks of positive mass:

$$V\big(q\varphi_0 - a\big) + L\big(q - a\big) + \lambda(t_H - t_A) \;>\; 0 \quad \text{on positive stakes mass.} \tag{23}$$

For the floor to generate direction practice, the tasks satisfying equation (23) must also lie in scope, and under Assumption 2 positive stakes mass is enough. The condition is sufficient rather than necessary, since a strong verifier can carry delegation at the floor through Vet even where Trust fails.

Throughout we maintain $\varphi > 0$, $\varphi' > 0$, $c' > 0$, $c'' > 0$, and a detection technology that satisfies Assumption 1 and does not depend on $a$. The setup cost $f > 0$ shifts $u_{\text{Vet}}$ by a constant, so it leaves the skill marginals below unchanged, though it does help locate the Trust–Vet boundary $q_T$ in the proof of Proposition 2.

We also maintain $c(0) = 0$. Because $c'$ is positive and increasing, $c$ is unbounded, and because $\rho \leq 1$, vetting effort never exceeds the $\bar{e}$ solving $\lambda c(\bar{e}) = V\varphi(1) + L_{\max}$: any larger effort is dominated by zero effort. So every maximization over effort is over $[0, \bar{e}]$, the maximizer is unique by strict concavity in effort, and by the maximum theorem it and the value function are continuous in their parameters. We use this without further comment. Where a condition is verified only in the parametric family the statement says so.

## A.2 The Baseline Results

Proofs for Section 3, with the full statement of the crossover and the reallocation proposition.

*Proof of Proposition 1.* The reduction (5) and the first-order condition (6) are algebraic identities. Since $u_{\text{Solo}} = aV - (1-a)L - \lambda t_H$ and the with-AI value carries no $a$ wherever an AI action is chosen, $\partial R/\partial a = -\partial u_{\text{Solo}}/\partial a = -(V+L) < 0$. Differentiating equation (3) in $m_d$ gives $(q+(1-q)\rho)V\varphi' > 0$, and differentiating in $m_v$ gives $(1-q)(V\varphi + L)\rho_{m_v}$. The indirect effect through $e^*$ vanishes by the envelope theorem, since $\partial u_{\text{Vet}}/\partial e = 0$ at the optimum, so that partial is the total derivative.

Part (i) is the first of those signs. Because $\partial^2 u_{\text{Solo}}/\partial a \partial L = 1 > 0$, the Solo option's task-skill advantage grows with stakes, which deepens $\partial R/\partial a < 0$. For part (ii), implicit differentiation of equation (6) gives the effort response to stakes,

$$\frac{de^*}{dL} = \frac{(1-q)\rho_e}{\lambda c'' - (1-q)(V\varphi + L)\rho_{ee}} > 0,$$

with a positive numerator because $\rho_e > 0$ and a positive denominator because it is the second-order

condition. The verification channel then moves with stakes as

$$\frac{\partial^2 R}{\partial m_v \, \partial L} = (1-q)\rho_{m_v} + (1-q)(V\varphi + L)\,\rho_{m_v e}\,\frac{de^*}{dL} > 0,$$

where the first term is the direct effect and the second the effort response, nonnegative because $\rho_{m_v e} \geq 0$ (Assumption 1). Without that condition the second term could flip the sign. The direction channel rises in $L$ through the same response, $\partial^2 R/\partial m_d \, \partial L = V\varphi'(1-q)\rho_e \, de^*/dL > 0$. □

**Proposition A1** (The crossover: full statement of Proposition 2)**.** *Parts* (i)–(iii) *are as stated in Proposition 2 in the main text.*

(iv) (Existence and order.) *Impose Assumption 3 at $q = 0$. Its two inequalities loosen as reliability rises, so imposing them there covers the whole path. Suppose also that verification leads at low reliability, $\Delta^*(q) > 0$ for every $q$ in some right-neighborhood of 0. Then*

$$q^* \;\equiv\; \sup\{q \in [0,1] : \Delta^*(q) \geq 0\} \tag{24}$$

*lies in $(0,1)$, attains its supremum, $\Delta^*(q^*) \geq 0$, and direction leads strictly at every $q > q^*$. Under Assumption 2 both marginal returns carry the common factor $F_D(\kappa)$, so $q^*$ is scope-free.*

(v) (Two regimes.) *Under the hypotheses of (iv), let $[0, q_T]$ be the set of reliabilities at which the engineer vets and $\Delta^{\text{Vet}}$ the gap on the vetted branch. If $\Delta^{\text{Vet}}(q_T) < 0$, the crossing is* interior*: $q^* < q_T$ is the largest zero of $\Delta^{\text{Vet}}$ on $[0, q_T]$, and $\text{MV}_{m_v} = \text{MV}_{m_d}$ there. If $\Delta^{\text{Vet}}(q_T) \geq 0$, the crossover is an* exit*: $q^* = q_T$, the engineer stops vetting while verification still carries the weakly higher marginal return, and the gap falls at $q_T$ to $-q_T V\varphi' < 0$. In either regime the return to direction past $q^*$ is at least $V\varphi' q$ on every delegated task. Under the ship-correct condition* (13) *holding on $[0, q_T]$, $\Delta^{\text{Vet}}$ is strictly decreasing there, so the sign change is unique.*

*Proof of Proposition 2.* (i) Implicit differentiation of equation (6) gives

$$\frac{de^*}{dq} = \frac{(V\varphi + L)\big[-\rho_e + (1-q)\rho_{eq}\big]}{\lambda c'' - (1-q)(V\varphi + L)\rho_{ee}} < 0.$$

The numerator is strictly negative because $\rho_e > 0$ and $\rho_{eq} \leq 0$ (Assumption 1), and the denominator is the second-order condition, strictly positive because $c'' > 0$ and $\rho_{ee} \leq 0$. (ii) The cross-partial in Proposition 2(ii) is the total derivative of $\text{MV}_{m_v} = (1-q)(V\varphi + L)\rho_{m_v}$ along $e^*(q)$. Its first term is strictly negative because $\rho_{m_v} > 0$. Its bracket is nonpositive because $\rho_{m_v q} \leq 0$ and $\rho_{m_v e} \geq 0$ (Assumption 1) while $de^*/dq < 0$ by (i), so $\text{MV}_{m_v}$ strictly decreases in $q$. (iii) The chosen action on an in-scope task does not depend on $d$, so with difficulty and stakes independent the portfolio value is $F_D(\kappa)\,\mathbb{E}_L\big[\max\{u_{\text{Solo}}, u_{\text{Trust}}, u_{\text{Vet}}\}\big] + (1 - F_D(\kappa))\,\mathbb{E}_L[u_{\text{Solo}}]$ and $\partial(\text{value})/\partial\kappa = f_D(\kappa)\,\mathbb{E}_L[R]$, the density at the frontier times the expected return on the marginal tasks there. Differentiating again in $m_d$ gives $\partial^2(\text{value})/\partial\kappa\,\partial m_d = f_D(\kappa)\,\mathbb{E}_L[\partial R/\partial m_d]$, with $f_D(\kappa) > 0$ by the standing assumption on the difficulty marginal. By Proposition 1 the integrand is $V\varphi'(q + (1-q)\rho)$ on a vetted task, $qV\varphi'$ on a trusted one, and zero on a Solo one. It is nonnegative, and strictly positive when the

marginal task is delegated with positive probability.

(iv)–(v) Fix a task at stakes $L$ and suppress it. We proceed in four steps.

*Step 1: the vetted branch is an interval.* Let $G(q) = u_{\text{Vet}}(e^*) - u_{\text{Trust}} = \max_e\{(1-q)\rho(m_v, e; q)(V\varphi + L) - \lambda c(e)\} - \lambda f$ be the value of catching errors, net of the setup cost, from equation (5). By the envelope theorem at the maximizer $e^*$,

$$G'(q) = (V\varphi + L)\big[-\rho(m_v, e^*; q) + (1-q)\rho_q(m_v, e^*; q)\big] \le 0,$$

both terms nonpositive by Assumption 1, and $G' = 0$ where $e^* = 0$ because $\rho(m_v, 0; q) = 0$ for every $q$. So $G$ is nonincreasing on all of $[0,1]$, which rules out re-entry: $\{G \ge 0\}$ is the interval $[0, q_T]$, empty if and only if $G(0) < 0$. On that set the decline is strict, since $G \ge 0$ forces $\rho > 0$ and so $G' < 0$. And $G(1) = -\lambda f < 0$ puts $q_T < 1$. Under Assumption 3 Solo is dominated on the task, so the chosen action is Vet on $[0, q_T]$ and Trust on $(q_T, 1]$. Without Trust dominance the vetted set can fail to be an interval: in our parameterized example a task can be vetted at low reliability, revert to Solo at moderate reliability, and be vetted again above that, and then parts (iv) and (v) fail with this step.

*Step 2: the gap $\Delta^*$.* Both $u_{\text{Trust}}$ and $u_{\text{Vet}}(e^*)$ are continuously differentiable in $(m_d, m_v)$, the latter by the envelope theorem, the maximizer being unique and continuous as noted in the preamble. The maximum of two such functions has one-sided derivatives everywhere, and at a tie the derivative from above in each coordinate is the larger of the two branch derivatives. That is the vetted one in both coordinates, since $\partial u_{\text{Vet}}/\partial m_v = (1-q)(V\varphi + L)\rho_{m_v} \ge 0 = \partial u_{\text{Trust}}/\partial m_v$ and $\partial u_{\text{Vet}}/\partial m_d = V\varphi'[q + (1-q)\rho] \ge qV\varphi' = \partial u_{\text{Trust}}/\partial m_d$. Hence

$$\Delta^*(q) = \begin{cases} \Delta^{\text{Vet}}(q) \equiv \text{MV}_{m_v} - V\varphi'\big[q + (1-q)\rho\big], & q \in [0, q_T], \\ -qV\varphi', & q \in (q_T, 1], \end{cases}$$

continuous on each piece.

*Step 3: existence and the two regimes.* On $(q_T, 1]$, $\Delta^* = -qV\varphi' < 0$, so $q^* \le q_T < 1$. The premise puts $\Delta^* > 0$ on a right-neighborhood of 0, which forces $q_T > 0$ and $q^* > 0$. On the closed interval $[0, q_T]$ the set $\{\Delta^{\text{Vet}} \ge 0\}$ is closed and contains that neighborhood, so the supremum in equation (24) is attained. If $\Delta^{\text{Vet}}(q_T) < 0$, then $\Delta^{\text{Vet}}$ has a largest zero $z \in (0, q_T)$, is negative on $(z, q_T]$, and $\Delta^* < 0$ beyond, so $q^* = z$ with $\Delta^*(z) = 0$. If $\Delta^{\text{Vet}}(q_T) \ge 0$, then $\Delta^*(q_T) \ge 0$ and $\Delta^* < 0$ on $(q_T, 1]$, so $q^* = q_T$, and the gap falls at $q_T$ from $\Delta^{\text{Vet}}(q_T) \ge 0$ to $-q_T V\varphi' < 0$. The bound past $q^*$ holds because the direction channel returns at least $V\varphi' q$ on every delegated task.

*Step 4: uniqueness under ship-correct.* Suppose equation (13) holds on $[0, q_T]$. There, $d\,\text{MV}_{m_d}/dq = V\varphi'\big[(1-\rho) + (1-q)\big(\rho_q + \rho_e\, de^*/dq\big)\big]$ is $V\varphi'$ times the left-hand side of the ship-correct condition (13), so it is nonnegative wherever that condition holds. With (ii), $\Delta^{\text{Vet}}$ is then strictly decreasing and its zero, if any, is unique.

*Scope.* Under Assumption 2, aggregating over the portfolio multiplies both marginal values by $F_D(\kappa) > 0$, which cancels from the comparison, so the sign of $\Delta^*$, and with it $q^*$, is scope-free.

Without independence the conditional stakes distribution on the reachable set moves with $\kappa$, and the switch point can move with it. Aggregating sums per-task gaps, each with its own jump at that task's $q_T(L)$, so the aggregate sign change need not be unique: in our parameterized example, with Solo active on part of the portfolio, the aggregated gap crosses zero twice and stays negative past the second crossing. □

*Proof of Corollary 1.* The with-AI payoffs (2) and (3) contain no $a$, so $U$ and the verification channel of Proposition 1 are $a$-free, while $\partial u_{\text{Solo}}/\partial a = V + L > 0$. Hence $u_{\text{Solo}} - U$ strictly increases in $a$ and crosses zero exactly once, at the $a^*$ of equation (11), which is equation (1) solved for $a$. The delegation region is $a < a^*$, empty if $a^* \leq 0$ and the whole unit interval if $a^* \geq 1$. On it $R = U - u_{\text{Solo}} > 0$. The return there is earned through $\varphi$ and $\rho$ alone, and past $q^*$ direction is the leading one of the two by Proposition A1(iv). $U$ aggregates reliability, the direction multiplier, the recovered errors, and the time saved, so $a^*$ can lie on either side of $q$, and $a > a^*$ is a live case at any reliability.

The threshold also moves with the price of attention, rising wherever handing the task off, checking time included, is faster than doing it Solo. $\lambda$ enters equation (11) through $t_H$ directly and through $U$. Where Trust is optimal, $dU/d\lambda = -t_A$, so

$$\frac{da^*}{d\lambda} = \frac{t_H - t_A}{V + L} > 0 \tag{25}$$

unconditionally, since delegation is the faster option ($t_A < t_H$). Where Vet is optimal, the envelope theorem at $e^*$ gives $dU/d\lambda = -(t_A + f + c(e^*))$, so the oversight time nets out,

$$\frac{da^*}{d\lambda} = \frac{t_H - t_A - f - c(e^*)}{V + L}, \tag{26}$$

which is positive only when Solo is slower than vetting, and negative where attention is cheap enough that the engineer buys more scrutiny than doing the task themselves would have cost. □

**Proposition 8** (Reallocation)**.** *Aggregate over the task portfolio of Section 2. The share of the engineer's tasks handled by oversight, meaning direction or vetting rather than their own execution, rises in the AI's scope $\kappa$ unconditionally. It rises in reliability $q$ wherever the vetted-value condition* (12) *holds across the portfolio, as more tasks clear the bar for delegation. The attention each delegated task demands falls as vetting effort drops, so the share of attention spent on oversight rises in scope but need not be monotone in reliability.*

*Proof of Proposition 8.* On an in-scope task the chosen action does not depend on $d$, since difficulty enters payoffs only through the gate, so the oversight task share is $F_D(\kappa)\,\Pr_L(\text{delegate})$ with the probability free of $\kappa$. It rises in $\kappa$ unconditionally, $F_D$ being strictly increasing. For reliability, $u_{\text{Solo}}$ carries no $q$, $\partial u_{\text{Trust}}/\partial q = V\varphi + L > 0$, and by the envelope theorem $\partial u_{\text{Vet}}/\partial q = (V\varphi + L)\big[(1 - \rho) + (1 - q)\rho_q\big]$, which is nonnegative wherever the vetted-value condition (12) holds. Where it holds across the portfolio, the with-AI value $\max\{u_{\text{Trust}}, u_{\text{Vet}}\}$ is nondecreasing in $q$ while $u_{\text{Solo}}$ is flat, so

the delegate set expands and the oversight share rises. Where it fails the share can fall.

For attention, each delegated task's oversight time $\lambda(f + c(e^*))$ falls in $q$ by Proposition 2(i), so the attention share multiplies a rising task share by a falling per-task demand and need not be monotone in $q$. In $\kappa$ it is monotone. Widening scope moves the marginal tasks from Solo execution time to delegation, adding their oversight time and removing their execution time while touching no other task, so the oversight share of attention rises. □

### A.3 Skill Formation as a Byproduct

The full statement of Proposition 3, with the stability and path conditions its parts require and a fourth part on the crossover in the steady-state mix along an AI-improvement path.

**Proposition A2** (Skill formation as a byproduct: full statement of Proposition 3)**.** *Consider myopic stage-optimal play with $\theta = 1$. Write $x = (x_d, x_v)$ for the practice map of Section 4 and, at a skill state $m = (m_d, m_v)$,*

$$A(m) = \eta^{-1}\mathrm{diag}(\psi_d, \psi_v)\left.\frac{\partial(x_d, x_v)}{\partial(m_d, m_v)}\right|_m, \qquad a_{ij} = A_{ij}(m),$$

*for the matrix of practice feedback, entrywise nonnegative by the monotonicity the proof of part* (ii) *establishes. We take $x$ to be continuously differentiable in $(m, q)$, as it is in the parametric family, where the stakes distribution has a density and the action boundaries move transversally in the stakes. Call a set $B \subseteq [0,1]^2$* stable *if there are positive weights $w \in \mathbb{R}^2_{++}$ and $\zeta < 1$ with $A(m)w \leq \zeta w$ at every $m \in B$. With $w = (1,1)$ this says each row of $A(m)$ sums to at most $\zeta$, so practice feedback dies out rather than compounds. For a nonnegative matrix such weights exist exactly when its spectral radius lies below one, so this is the usual stability condition, imposed uniformly on $B$, and a point is stable if its singleton is.* (i) *On any task on which Vet is stage-optimal the chosen effort is strictly positive, $e^* > 0$ (which in particular requires $m_v > 0$), so $x_v > 0$ wherever vetting is stage-optimal on positive mass, and $x_d > 0$ wherever delegation is.* (ii) *As stated in Proposition 3, under two hypotheses. Write $m_v^\dagger$ and $m_v^\circ$ for the threshold and the skilled state of Proposition 5(iii) at its conservative $m_d = 0$ slice, the two crossings of the slice map $T_0(m_v) = \min\{1, (1-\eta)m_v + \psi_v x_v(0, m_v)\}$ with the 45° line. First, the sigmoid condition* (32) *holds at that slice. Second, the order interval $[(0, m_v^\circ), (1,1)]$ is stable.* (iii) *Under Assumption 4 and the in-scope caveat that follows it, direction practice carries a floor: $x_d(m) \geq \underline{x}_d \equiv x_d(0,0) > 0$ at every skill state $m \in [0,1]^2$, and along any path of myopic play, whatever the verification stock does, $\liminf_{t\to\infty} m_{d,t} \geq \min\{1, \psi_d \underline{x}_d/\eta\} > 0$.* (iv) *Take an AI-improvement path, parameterized by $q$ with $\kappa$ rising along it, and a stable interior steady state $m^*$ of equation* (15)*. Let $v_d \equiv \eta^{-1}\psi_d\, dx_d/dq$ and $v_v \equiv \eta^{-1}\psi_v\, dx_v/dq$, both evaluated at fixed $m = m^*$, be the direct effects of the improvement on the two practices, collected as $v = (v_d, v_v)$. Impose two conditions at $m^*$.* Saturation (S)*: $\partial x_d/\partial m_v = 0$, so that $a_{12} = 0$. A marginal unit of verification skill then moves no task across the delegation boundary, either because the marginal Solo tasks are stakes-limited rather than detection-limited or because no in-scope task is kept Solo at all.* Net drain (N)*: $v_d > 0$ and $a_{21}v_d + (1 - a_{11})v_v < 0$.*

*The improvement adds to direction practice directly, and the indirect gain to verification practice, through the direction skill that practice builds, is smaller than the direct drain. The second inequality forces $v_v < 0$, since $a_{21} \geq 0$, $v_d > 0$, and $a_{11} < 1$: there is no net drain to condition on unless the improvement drains verification practice. Then the steady-state mix crosses over,*

$$\frac{dm_d^*}{dq} = \frac{v_d}{1 - a_{11}} > 0, \qquad \frac{dm_v^*}{dq} = \frac{a_{21}v_d + (1 - a_{11})v_v}{(1 - a_{11})(1 - a_{22})} < 0.$$

*We prove this pointwise: at any $q$ at which the steady state is interior and stable and* (S) *and* (N) *hold, the implicit function theorem returns a locally unique $C^1$ branch $q \mapsto m^*(q)$ along which both signs hold.*

*Proof of Proposition 3, part (i).* We argue by revealed preference. Fix a task on which Vet is stage-optimal, so that $u_{\text{Vet}}(e^*) \geq u_{\text{Trust}}$. Trust is available because Vet is an option only on a task the tool can attempt, and on such a task all three actions are available (Section 2). At zero effort the vetted payoff (3) collapses onto the trusted one, since $\rho(m_v, 0; q) = 0$ (Assumption 1) and $c(0) = 0$ give $u_{\text{Vet}}(0) = qV\varphi - (1 - q)L - \lambda(t_A + f) = u_{\text{Trust}} - \lambda f$. Since $\lambda, f > 0$, $u_{\text{Vet}}(e^*) \geq u_{\text{Trust}} > u_{\text{Vet}}(0)$, so $e^* \neq 0$ and hence $e^* > 0$. The same collapse rules out vetting at $m_v = 0$, where $\rho \equiv 0$ (Assumption 1) leaves $u_{\text{Vet}}(e) = u_{\text{Trust}} - \lambda(f + c(e)) < u_{\text{Trust}}$ at every effort, so Trust strictly dominates Vet and Vet is never stage-optimal there. A task on which Vet *is* stage-optimal therefore has $m_v > 0$, a condition implied rather than imposed. Hence $x_v = \mathbb{E}[e^* \mathbf{1}\{\text{vet}\}] > 0$ wherever Vet is chosen on positive mass, and $x_d = \Pr\{\text{delegate}\} > 0$ wherever delegation is. □

*Proof of Proposition 3, part (ii).* Write equation (15) as $m_{t+1} = T(m_t)$ with $T(m) = \min\{1, (1 - \eta)m + \psi \circ x(m)\}$ componentwise on $[0, 1]^2$. We first show that the practice map is componentwise nondecreasing in $(m_d, m_v)$. On an in-scope task $u_{\text{Solo}}$ carries neither managerial skill. Raising $m_d$ raises $u_{\text{Trust}}$ by $qV\varphi'$ and $u_{\text{Vet}}$ by $(q+(1-q)\rho)V\varphi'$ (envelope), so the delegate set $\{\max\{u_{\text{Trust}}, u_{\text{Vet}}\} \geq u_{\text{Solo}}\}$ weakly expands task by task. Raising $m_v$ raises $u_{\text{Vet}}$ alone, by $(1 - q)(V\varphi + L)\rho_{m_v} \geq 0$ (envelope), and expands the same set. Hence $\partial x_d/\partial m_d,\ \partial x_d/\partial m_v \geq 0$.

For $x_v = \mathbb{E}[e^* \mathbf{1}\{\text{vet}\}]$ the Vet region and the effort on it move together. The Vet-minus-Trust gap (5) is maximized over $e$ at $e^*$, so its derivatives in $m_d$ and $m_v$ are $(1 - q)\rho V\varphi' \geq 0$ and $(1 - q)(V\varphi + L)\rho_{m_v} \geq 0$, and the Vet-minus-Solo gap rises by the corresponding envelope terms, so the Vet region weakly expands against both margins. On it, differentiating equation (6) gives $de^*/dm_d = (1 - q)V\varphi'\rho_e/D \geq 0$ and $de^*/dm_v = (1 - q)(V\varphi + L)\rho_{m_v e}/D \geq 0$, where $D = \lambda c'' - (1 - q)(V\varphi + L)\rho_{ee} > 0$ is the second-order condition and the last sign uses $\rho_{m_v e} \geq 0$ (Assumption 1). Hence $\partial x_v/\partial m_d,\ \partial x_v/\partial m_v \geq 0$, and $T$ is monotone on the box $[0, 1]^2$. With atomless $(d, L)$ the practice measures $x_d$ and $x_v$ are continuous in $m$, payoff ties sitting on null sets, so limits of monotone iterate sequences are fixed points of $T$.

The iterates $T^n(0, 0)$ increase to the least fixed point and $T^n(1, 1)$ decrease to the greatest, and from any sub-solution, an $m$ with $T(m) \geq m$, the iterates increase to a fixed point above it. Where the vetting practice curve of Proposition 5, drawn at the direction floor $\varphi_0$, clears the 45° line just above the threshold, the point $s = (0, m_{v,s})$ with $m_{v,s} \in (m_v^\dagger, m_v^\circ)$ and $m_{v,s} \leq m_{v,0}$

is such a sub-solution, since its first coordinate cannot fall. That slice is the conservative one, since $x_v$ is nondecreasing in $m_d$. The start $m_0$ is then sandwiched between $s$ and $(1,1)$, with $T^n(s) \le T^n(m_0) \le T^n(1,1)$ and both bounds converging to fixed points $\underline{p} \le \bar{p}$ in the order interval $[s,(1,1)]$. Both lie in the stable box $[(0,m_v^\circ),(1,1)]$: monotonicity of $x_v$ in both arguments gives $T^n(s)_v \ge T_0^n(m_{v,s})$, and under the sigmoid condition $T_0^n(m_{v,s})$ rises to $m_v^\circ$ by Lemma 5, so $\underline{p}_v \ge m_v^\circ$.

Stability of the box gives uniqueness. Let $w$ and $\zeta$ be the weights and bound in the hypothesis, write $\|z\|_w = \max_i |z_i|/w_i$, and write $\Sigma(m) = (1-\eta)m + \psi \circ x(m)$ for the uncapped map. The segment between $\underline{p}$ and $\bar{p}$ lies in the box, and the mean-value form gives $\Sigma(\bar{p}) - \Sigma(\underline{p}) = \big[(1-\eta)I + \eta M\big](\bar{p} - \underline{p})$ with $M = \int_0^1 A\big(\underline{p} + t(\bar{p} - \underline{p})\big)\, dt$, so that $M \ge 0$ and $Mw \le \zeta w$. Set $z = \bar{p} - \underline{p} \ge 0$. Since $\Sigma$ is monotone and the cap $\min\{1,\cdot\}$ is 1-Lipschitz, $z = T(\bar{p}) - T(\underline{p}) \le \Sigma(\bar{p}) - \Sigma(\underline{p})$, and since $z \le \|z\|_w\, w$ entrywise and $M \ge 0$,

$$z \;\le\; \|z\|_w \big[(1-\eta)w + \eta M w\big] \;\le\; \big[(1-\eta) + \eta\zeta\big]\, \|z\|_w\, w.$$

Taking the weighted norm of both sides gives $\|z\|_w \le [(1-\eta) + \eta\zeta]\|z\|_w$ with $(1-\eta) + \eta\zeta < 1$, so $z = 0$: the two limits coincide, and every admitted start converges to the same steady state, from below and from above.

Both coordinates of the common limit $m^*$ are strictly positive. It dominates the sub-solution, $m^* \ge s = (0, m_{v,s})$, so $m_v^* \ge m_{v,s} > 0$ at once. For direction, positive vetting practice at $s$ means positive vetted mass there, and a vetted task is a delegated one, so $x_d(s) > 0$. Monotonicity carries that to $x_d(m^*) > 0$, and $\eta m_d^* = \psi_d x_d(m^*) > 0$ wherever the cap is slack, with $m_d^* = 1$ where it binds. □

*Proof of Proposition 3, part (iii). The floor on practice.* By equation (23), at $m_d = 0$ the gap $u_{\text{Trust}} - u_{\text{Solo}} = V(q\varphi_0 - a) + L(q - a) + \lambda(t_H - t_A)$ is strictly positive on a set of stakes of positive mass. With difficulty independent of stakes and $F_D(\kappa) > 0$, that set carries positive mass of *in-scope* tasks, which is the caveat following Assumption 4. ($F_D(\kappa) > 0$ at any $\kappa > 0$ by the standing assumption that the difficulty density is strictly positive on $[0,1]$.) On each of those tasks the chosen action is Trust or Vet, since $\max\{u_{\text{Trust}}, u_{\text{Vet}}\} \ge u_{\text{Trust}} > u_{\text{Solo}}$, and neither $u_{\text{Trust}}$ nor $u_{\text{Solo}}$ carries $m_v$, so the comparison holds at every verification stock, and so does the delegation. Hence $\underline{x}_d = x_d(0,0) > 0$, and by the componentwise monotonicity of the practice map established in the proof of part (ii), $x_d(m) \ge x_d(0,0) = \underline{x}_d$ at every $m \in [0,1]^2$.

*The bound.* Write $g(y) = \min\{1, (1-\eta)y + \psi_d \underline{x}_d\}$ and $\underline{m}_d = \min\{1, \psi_d \underline{x}_d/\eta\}$, so that $\eta \underline{m}_d \le \psi_d \underline{x}_d$. By equation (15) and the floor, $m_{d,t+1} \ge g(m_{d,t})$ whatever $m_{v,t}$ is doing. The map $g$ is nondecreasing on $[0,1]$, so induction gives $m_{d,t} \ge g^t(m_{d,0})$. And $\underline{m}_d$ is the unique fixed point of $g$, with $g(y) > y$ below it and $g(y) < y$ above. A nondecreasing map with that structure carries every $y \in [0,1]$ monotonically to $\underline{m}_d$, so $g^t(m_{d,0}) \to \underline{m}_d$ and $\liminf_t m_{d,t} \ge \underline{m}_d > 0$. Neither step restricts the verification path. Because the floor binds at every $(m_d, m_v)$, a collapse in verification cannot starve direction practice. An improving tool does not undo the floor either, since the left-hand side of equation (23) rises in $q$, at rate $V\varphi_0 + L > 0$, and the in-scope mass $F_D(\kappa)$ rises in $\kappa$, so a floor

that holds at the start of an AI-improvement path holds along it. □

*Proof of Proposition 3, part (iv).* At an interior steady state the cap in equation (15) is slack and the law of motion reads $\eta m^* = \psi \circ x(m^*)$. Differentiating implicitly along the path gives $(I - A)\, dm^*/dq = v$, with $A$ and $v$ as in the statement and $A$ entrywise nonnegative by the four monotonicity signs in the proof of part (ii). That step is legitimate because $x$ is $C^1$ and $I - A$ is invertible at a stable state, so the implicit function theorem returns a locally unique $C^1$ branch $m^*(q)$. Under (S) we have $a_{12} = 0$, so $A$ is lower triangular, its eigenvalues are its diagonal entries $a_{11}$ and $a_{22}$, and stability amounts to $a_{11} < 1$ and $a_{22} < 1$. Its first row is $(1 - a_{11})\, dm_d^*/dq = v_d$, so $dm_d^*/dq = v_d/(1 - a_{11})$, strictly positive under (N). Substituting that into the second row, $-a_{21}\, dm_d^*/dq + (1 - a_{22})\, dm_v^*/dq = v_v$, gives

$$\frac{dm_v^*}{dq} = \frac{a_{21} v_d + (1 - a_{11}) v_v}{(1 - a_{11})(1 - a_{22})},$$

whose denominator is positive by stability and whose numerator is negative by (N). Without saturation the two rows are coupled, and signing $dm_d^*/dq$ takes the further condition $(1 - a_{22})v_d + a_{12} v_v > 0$, which stability alone does not deliver. □

**The two conditions in the parametric family.** The direct effects have the assumed signs on the paths we consider. Direction practice gains through reach unconditionally, since widening $\kappa$ moves marginal tasks from Solo execution into delegation (Proposition 8), and it gains through reliability wherever the vetted-value condition (12) holds. Verification practice drains in $q$ outside the low-reliability region of the footnote to Proposition A5. Reach, however, feeds *both* practices, so the net drain in (N) restricts the relative speed of the two improvements rather than following from Proposition 8, and it enters as a hypothesis on the path. In our parameterized example, along an AI-improvement path on which scope widens slowly relative to reliability, both conditions hold at almost every stable interior steady state, saturation because at those states every in-scope task is delegated, so $x_d = F_D(\kappa)$ and both of its skill derivatives vanish. Wherever the two hold, the crossover signs of part (iv) follow.

## A.4 The Accountability Wedge

Proofs for Section 5.

**Lemma 1** (Wedge decomposition and monotone selection)**.** *Fix an in-scope task, holding the worker's skills, the tool, the stakes, and the remaining primitives fixed. Let* $\mathcal{A} = \{\text{Solo}, \text{Trust}\} \cup \{(\text{Vet}, e) : e \geq 0\}$*, let* $S(\alpha)$ *be the social value of* $\alpha \in \mathcal{A}$ *and* $D(\alpha)$ *its expected shipped loss,*

$$D(\text{Solo}) = (1 - a)L, \qquad D(\text{Trust}) = (1 - q)L, \qquad D(\text{Vet}, e) = (1 - q)\big(1 - \rho(m_v, e; q)\big)L. \quad (27)$$

*Then, under Assumption 1 alone:* (a) $u_\theta(\alpha) = S(\alpha) + (1 - \theta)D(\alpha)$ *for every* $\alpha$ *and every* $\theta$*, and* $\arg\max_{\alpha \in \mathcal{A}} u_\theta(\alpha)$ *is nonempty for every* $\theta \in [0, 1]$*.* (b) (Monotone selection.) *For* $\theta < \theta'$ *and*

any *selections* $\alpha^\theta$, $\alpha^{\theta'}$ *from the respective argmax sets,* $D(\alpha^\theta) \geq D(\alpha^{\theta'})$ *and* $S(\alpha^\theta) \leq S(\alpha^{\theta'})$. (c) $\mathrm{WL}(\theta) \equiv S(\alpha^1) - S(\alpha^\theta)$ *is nonnegative, nonincreasing in* $\theta$ *along any selection, and zero at* $\theta = 1$. (d) (Exact split.) *For every* $\theta$ *and every selection,*

$$\underbrace{\mathrm{WL}(\theta)}_{\text{firm's loss}} \;+\; \underbrace{\big[u_\theta(\alpha^\theta) - u_\theta(\alpha^1)\big]}_{\text{worker's private gain}} \;=\; (1-\theta)\big[D(\alpha^\theta) - D(\alpha^1)\big],$$

*with all three quantities nonnegative. Dropping the worker's gain leaves* $\mathrm{WL}(\theta) \leq (1-\theta)[D(\alpha^\theta) - D(\alpha^1)]$, *which is equation* (18).

*Proof of Lemma 1.* (a) Substituting equation (27) into equations (1) to (3) gives the identity term by term: on Solo, $aV - (1-a)\theta L - \lambda t_H = [aV - (1-a)L - \lambda t_H] + (1-\theta)(1-a)L$, and likewise on Trust and on Vet with $\rho$ evaluated at the same $e$. Existence follows from the standing bound on effort.

(b) Write $t = 1-\theta > t' = 1-\theta'$. Optimality of each choice at its own $\theta$ gives

$$S(\alpha^\theta) + t\,D(\alpha^\theta) \;\geq\; S(\alpha^{\theta'}) + t\,D(\alpha^{\theta'}), \qquad S(\alpha^{\theta'}) + t'D(\alpha^{\theta'}) \;\geq\; S(\alpha^\theta) + t'D(\alpha^\theta).$$

Adding and cancelling the $S$ terms leaves $(t-t')\big[D(\alpha^\theta) - D(\alpha^{\theta'})\big] \geq 0$, and $t > t'$ gives the first claim. Substituting it into the second displayed inequality gives $S(\alpha^{\theta'}) - S(\alpha^\theta) \geq t'\big[D(\alpha^\theta) - D(\alpha^{\theta'})\big] \geq 0$, which is the second. Both inequalities use only the optimality of each selection at its own $\theta$, so they cover Solo switches too.

(c) $\alpha^1$ maximizes $u_1 = S$ over $\mathcal{A}$ and $\alpha^\theta \in \mathcal{A}$, so $\mathrm{WL} \geq 0$, with $\mathrm{WL}(1) = 0$ because $S(\alpha^1)$ is the common maximized value. Monotonicity is the second claim of (b) applied with $\theta' > \theta$.

(d) Expand $u_\theta(\alpha^\theta) - u_\theta(\alpha^1)$ using (a):

$$u_\theta(\alpha^\theta) - u_\theta(\alpha^1) = \big[S(\alpha^\theta) - S(\alpha^1)\big] + (1-\theta)\big[D(\alpha^\theta) - D(\alpha^1)\big] = -\mathrm{WL}(\theta) + (1-\theta)\big[D(\alpha^\theta) - D(\alpha^1)\big],$$

which rearranges to the display. The worker's gain is nonnegative because $\alpha^\theta$ is optimal at $\theta$, $\mathrm{WL} \geq 0$ by (c), and the right-hand bracket is nonnegative by (b) with $\theta' = 1$. The bound follows by dropping the worker's gain. □

**Proposition A3** (Over-delegation: full statement of Proposition 4)**.** *Fix an in-scope task as in Lemma 1 and let* $\theta < 1$. *Write* $\alpha^\theta$ *for any selection from* $\arg\max_{\alpha \in \mathcal{A}} u_\theta(\alpha)$. (i) Intensive margin. *The Vet branch is strictly concave in* $e$ *on its interior, so* $e^w(\theta)$ *is its unique maximizer. Wherever that maximizer is interior,* $e^w$ *is strictly increasing in* $\theta$, *and* $e^w(\theta) < e^* = e^w(1)$ *whenever* $m_v > 0$ *and* $L > 0$. (ii) Welfare. *Apply Lemma 1*(b)–(d) *with* $D$ *given by equation* (27). *Along* any *selections,* $D(\alpha^\theta)$ *is nonincreasing in* $\theta$, *the welfare loss* $\mathrm{WL}(\theta) = S(\alpha^1) - S(\alpha^\theta)$ *is nonnegative and nonincreasing in* $\theta$ *with* $\mathrm{WL}(1) = 0$, *and equation* (18) *holds.* (iii) The over-reliance band. *Let* $L > 0$ *and impose Assumption 3, under which Trust weakly dominates Solo for the engineer and for the firm's benchmark alike, and take an optimal selection that never plays Solo. Let* $m_v^{\mathrm{FB}}$ *and* $m_v^w$ *be the least verification skills at which vetting covers its setup cost at stakes* $L$ *and at*

*internalized stakes $\theta L$, under the convention that the infimum of the empty set is $+\infty$. In that case nobody vets at any skill, the band is empty and* WL $\equiv 0$. *Then $0 < m_v^{\mathrm{FB}} \le m_v^w$, strictly whenever $m_v^{\mathrm{FB}} \le 1$, and the three regimes are those of Proposition 4*(iii). *At $m_v^w$ the welfare loss drops by exactly $(1-q)(1-\theta)L\,\rho(m_v^w, e^w; q)$, so the band's top is the maximum over the two lower regimes. It is the global maximum under condition* (30) *below.*

*Proof of Proposition 4.* Throughout write

$$\Gamma(m_v, x) \;\equiv\; \max_{e \ge 0} \Big\{ (1-q)\,\rho(m_v, e; q)\big(V\varphi(m_d) + x\big) - \lambda c(e) \Big\} - \lambda f \tag{28}$$

for the value of vetting over trusting when the decision-maker internalizes stakes $x$, and $e^*(x; m_v)$ for the maximizer. By equation (5) the engineer vets rather than trusts iff $\Gamma(m_v, \theta L) \ge 0$ and the firm's benchmark iff $\Gamma(m_v, L) \ge 0$. The maximand in equation (28) is continuous and nondecreasing in $m_v$ and in $x$ pointwise in $e$, so by the standing bound on effort $\Gamma$ is continuous and nondecreasing in each argument, with envelope derivatives $\Gamma_x = (1-q)\rho(m_v, e^*) \ge 0$ and $\Gamma_{m_v} = (1-q)(V\varphi + x)\rho_{m_v}(m_v, e^*) \ge 0$. Because $\rho(0, e; q) = 0$ and $c(0) = 0$ (Assumption 1), $\Gamma(0, x) = -\lambda f < 0$ for every $x$, so nobody with no verification skill vets.

(i) The Vet branch is strictly concave in $e$ on its interior ($\lambda c'' > 0$ and $\rho_{ee} \le 0$), so $e^w(\theta)$ is the unique root of $(1-q)(V\varphi + \theta L)\rho_e = \lambda c'(e)$. Implicit differentiation gives

$$\frac{de^w}{d\theta} = \frac{(1-q)L\,\rho_e}{\lambda c'' - (1-q)(V\varphi + \theta L)\rho_{ee}} \;>\; 0$$

whenever $m_v > 0$, $L > 0$ and $e^w > 0$, the denominator being the second-order condition. Integrating from $\theta$ to 1 gives $e^w(\theta) < e^*$.

(ii) This is Lemma 1(b)–(d) verbatim, with $D$ given by equation (27), and it uses nothing beyond Assumption 1.

(iii) *Trust dominance removes Solo.* The gap $u_{\mathrm{Trust}} - u_{\mathrm{Solo}}$ at internalized stakes $x$ is $V\big(q\varphi(m_d) - a\big) + x\big(q - a\big) + \lambda(t_H - t_A)$, affine in $x$. Assumption 3 states that it is nonnegative at $x = 0$ (first inequality) and at $x = L_{\max}$ (second), so it is nonnegative on all of $[0, L_{\max}]$. Because $\theta L \in [0, L_{\max}]$ for every $\theta \in [0, 1]$ and every task in the portfolio, Trust weakly dominates Solo for the engineer at every accountability level and for the firm's benchmark ($\theta = 1$) alike. Some optimal selection therefore never plays Solo, and any selection that does is payoff-equivalent to it. We take that selection throughout.

*Thresholds.* $\Gamma(\cdot, x)$ is continuous and nondecreasing, so $\{m_v : \Gamma(m_v, x) \ge 0\}$ is a closed upper interval. Define $m_v^{\mathrm{FB}} = \inf\{m_v \in [0,1] : \Gamma(m_v, L) \ge 0\}$ and $m_v^w = \inf\{m_v \in [0,1] : \Gamma(m_v, \theta L) \ge 0\}$, with the convention $\inf \emptyset = +\infty$ (nobody vets at any skill, the band is empty and WL $\equiv 0$). $\Gamma$ is nondecreasing in $x$ and $\theta L \le L$, so $m_v^w \ge m_v^{\mathrm{FB}}$, and $m_v^{\mathrm{FB}} > 0$ because $\Gamma(0, \cdot) < 0$. The inequality is strict when $m_v^{\mathrm{FB}} \le 1$. At $m_v^{\mathrm{FB}}$ we have $\Gamma(m_v^{\mathrm{FB}}, L) = 0$ by continuity. Suppose $\Gamma(m_v^{\mathrm{FB}}, \theta L) \ge 0$, and write $e^w = e^*(\theta L; m_v^{\mathrm{FB}})$. Then the engineer's bracket at $e^w$ covers $\lambda f > 0$, which forces

$\rho(m_v^{\mathrm{FB}}, e^w) > 0$, and the firm's bracket at that same effort gives

$$\Gamma(m_v^{\mathrm{FB}}, L) \;\geq\; \Gamma(m_v^{\mathrm{FB}}, \theta L) + (1-q)(1-\theta)L\,\rho(m_v^{\mathrm{FB}}, e^w) \;>\; 0,$$

a contradiction, the added term being positive because $L > 0$, $q < 1$, and $\theta < 1$. Hence $\Gamma(m_v^{\mathrm{FB}}, \theta L) < 0$ and $m_v^w > m_v^{\mathrm{FB}}$.

*The three regimes.* Below $m_v^{\mathrm{FB}}$ neither vets, both play Trust, the actions coincide and $\mathrm{WL} = 0$. This step needs Trust dominance: without it the fallback can be Solo and the wedge distorts that switch too. On $[m_v^{\mathrm{FB}}, m_v^w)$ the firm vets at $e^{\mathrm{FB}} = e^*(L; m_v)$ and the engineer trusts, so, using $S(\mathrm{Vet}, e) = S(\mathrm{Trust}) + (1-q)\rho(e)(V\varphi + L) - \lambda c(e) - \lambda f$,

$$\mathrm{WL}(m_v) = \Gamma(m_v, L), \tag{29}$$

which is continuous, equals zero at $m_v^{\mathrm{FB}}$, and is nondecreasing in $m_v$ with slope $(1-q)(V\varphi + L)\rho_{m_v}(m_v, e^{\mathrm{FB}}) \geq 0$. Above $m_v^w$ both vet and $\mathrm{WL}(m_v) = S(\mathrm{Vet}, e^{\mathrm{FB}}) - S(\mathrm{Vet}, e^w) \geq 0$ by Lemma 1(c). At $m_v^w$ the engineer is indifferent between vetting and trusting, $u_\theta(\mathrm{Vet}, e^w) = u_\theta(\mathrm{Trust})$, so Lemma 1(a) gives $S(\mathrm{Vet}, e^w) - S(\mathrm{Trust}) = (1-\theta)\big[D(\mathrm{Trust}) - D(\mathrm{Vet}, e^w)\big] = (1-q)(1-\theta)L\,\rho(m_v^w, e^w)$. That is the difference between WL just below $m_v^w$ and at it, so WL drops at $m_v^w$ by exactly that amount, which is positive because $\Gamma(m_v^w, \theta L) = 0$ forces $\rho(m_v^w, e^w) > 0$ as in the threshold step. Because WL is zero below the band and nondecreasing on it, the band's top is the maximum over the two lower regimes.

*When the band's peak is global.* A sufficient condition for the peak at $m_v^w$ to be global is

$$\rho\big(m_v, e^*(L; m_v)\big) - \rho\big(m_v, e^*(\theta L; m_v)\big) \;\leq\; \rho\big(m_v^w, e^*(\theta L; m_v^w)\big) \qquad \text{for all } m_v \geq m_v^w, \tag{30}$$

that the effort-induced detection gap never exceeds the detection already achieved at the band's top. To see that it suffices, fix $m_v \geq m_v^w$ and write $\rho^{\mathrm{FB}}$ and $\rho^w$ for detection at the two efforts. By Lemma 1(a), $\mathrm{WL}(m_v) = \big[u_\theta(e^{\mathrm{FB}}) - u_\theta(e^w)\big] + (1-q)(1-\theta)L\big(\rho^{\mathrm{FB}} - \rho^w\big)$, and the first bracket is nonpositive because $e^w$ maximizes $u_\theta$. So $\mathrm{WL}(m_v) \leq (1-q)(1-\theta)L(\rho^{\mathrm{FB}} - \rho^w)$, which under equation (30) is at most the drop at $m_v^w$, and the drop is at most the value at the band's top. □

Proposition 4(iii) describes one task. The prediction in Section 5.1 is about a portfolio, and the step from one to the other rests on an exact scaling. The engineer's problem depends on accountability and stakes only through the internalized stakes $\theta L$, so the whole band translates rigidly in log stakes as verification skill varies.

**Lemma 2** (Sliding window and the aggregate band)**.** *Impose Assumptions 1 and 3. Let* $\Gamma$ *be as in equation* (28) *and define the* internalized-stakes vetting threshold $\bar{x}(m_v) = \inf\{x \geq 0 : \Gamma(m_v, x) \geq 0\}$ *and* $\widetilde{m_v} = \inf\{m_v : \Gamma(m_v, 0) \geq 0\}$ *(with* $\inf \emptyset = +\infty$*). Then:* (a) (Exact scaling.) *For every* $m_v$*, the firm's and the engineer's stakes thresholds are* $\bar{L}^{\mathrm{FB}}(m_v) = \bar{x}(m_v)$ *and* $\bar{L}^w(m_v) = \bar{x}(m_v)/\theta$*, so* $\bar{L}^w(m_v) = \bar{L}^{\mathrm{FB}}(m_v)/\theta$ *identically in* $m_v$*.* (b) (Fixed-width sliding window.) *On* $(0, \widetilde{m_v}) \cap (0, 1]$*,* $\bar{x}$ *is finite, strictly positive, continuous and strictly decreasing, with* $\bar{x}(m_v) \to \infty$ *as* $m_v \downarrow 0$ *and*

$\bar{x}(m_v) \to 0$ *as* $m_v \uparrow \widetilde{m_v}$. *Write* $s(m_v) = \ln \bar{x}(m_v)$. *The over-reliance band in log stakes is then the interval* $[\, s(m_v),\, s(m_v) + \omega \,)$ *of fixed width* $\omega = \ln(1/\theta)$, *and* $s$ *is continuously differentiable and strictly decreasing.* (c) (Aggregation.) *Let log stakes* $u = \ln L$ *have density* $h$ *on* $\mathbb{R}$ *(zero off the support) and let difficulty be independent of stakes, so that in-scope tasks are a share* $F_D(\kappa)$ *at every stakes level. The excess unchecked-shipping rate at verification skill* $m_v$ *(the portfolio share on which the firm's benchmark vets and the engineer does not) is*

$$E(m_v) = F_D(\kappa)\, M\big(s(m_v)\big), \qquad M(s) = \int_s^{s+\omega} h(u)\, du.$$

(d) (Single peak.) *If* $h$ *is unimodal on* $\mathbb{R}$ *(nondecreasing then nonincreasing), then* $M$ *is single-peaked in* $s$, *and hence* $E$ *is single-peaked in* $m_v$, *with* $E(m_v) \to 0$ *as* $m_v \downarrow 0$ *and* $E(m_v) = 0$ *for* $m_v \geq \widetilde{m_v}$. *If* $\widetilde{m_v} > 1$ *the conclusion holds on* $(0, 1]$, *where the peak may be attained at* $m_v = 1$.

*Proof of Lemma 2.* (a) Under Assumption 3 the vetting decision is Vet-versus-Trust for both parties (proof of Proposition 4(iii)), so the engineer vets iff $\Gamma(m_v, \theta L) \geq 0$ and the firm's benchmark iff $\Gamma(m_v, L) \geq 0$. Both conditions are the same condition on the *internalized* stakes, and $\Gamma(m_v, \cdot)$ is nondecreasing, so each vet set is an upper interval in its own argument, namely $L \geq \bar{x}(m_v)$ for the firm and $\theta L \geq \bar{x}(m_v)$, i.e. $L \geq \bar{x}(m_v)/\theta$, for the engineer.

(b) At any $m_v$ with $\bar{x}(m_v) > 0$ we have $\Gamma(m_v, \bar{x}(m_v)) = 0$ by continuity, and there $\Gamma \geq 0$ forces $\rho > 0$, so $\Gamma_x = (1-q)\rho(m_v, e^*) > 0$ and $\Gamma_{m_v} = (1-q)(V\varphi + \bar{x})\rho_{m_v}(m_v, e^*) > 0$, using $\rho_{m_v} > 0$ at $e^* > 0$ (Assumption 1). These envelope partials are continuous because $e^*$ is, so $\Gamma$ is continuously differentiable near its zero and the implicit function theorem gives $\bar{x}$ continuously differentiable with $d\bar{x}/dm_v = -\Gamma_{m_v}/\Gamma_x < 0$. By definition of $\widetilde{m_v}$, $\bar{x}(m_v) > 0$ exactly on $(0, \widetilde{m_v})$ and $\bar{x} \to 0$ at its right endpoint. Taking logs, the firm's benchmark vets at $u \geq s(m_v)$ and the engineer at $u \geq s(m_v) + \ln(1/\theta)$, so the band is $[s(m_v), s(m_v) + \omega)$ with $\omega = \ln(1/\theta)$ independent of $m_v$.

(c) Out-of-scope tasks are Solo for both parties and contribute nothing. With difficulty independent of stakes, the in-scope share is $F_D(\kappa)$ at every $L$, and the remaining share of tasks on which only the firm's benchmark vets is the stakes mass in the band, which in log stakes is $M(s(m_v))$.

(d) $M$ is differentiable a.e. with $M'(s) = h(s+\omega) - h(s)$. Let $u^*$ be a mode of $h$. If $s + \omega \leq u^*$ then $s < s + \omega \leq u^*$ and $h$ is nondecreasing there, so $M' \geq 0$. If $s \geq u^*$ then $h$ is nonincreasing on $[s, s+\omega]$, so $M' \leq 0$. On the remaining interval $s \in (u^* - \omega, u^*)$, $s \mapsto h(s+\omega)$ is nonincreasing (its argument lies above $u^*$) and $s \mapsto h(s)$ is nondecreasing (its argument lies below $u^*$), so $M'(s) = h(s+\omega) - h(s)$ is nonincreasing and changes sign at most once, from $+$ to $-$. Hence $M$ is nondecreasing then nonincreasing on $\mathbb{R}$, so it is single-peaked. Then $E = F_D(\kappa)\, M \circ s$ with $s$ continuous and strictly decreasing, and a single-peaked function composed with a continuous strictly monotone reparameterization is single-peaked. The endpoints follow from (b). As $m_v \downarrow 0$ the window escapes to $+\infty$ and $M \to 0$, and for $m_v \geq \widetilde{m_v}$ the two thresholds coincide at zero and the band is empty. Unimodality is needed: without it the excess rate can have one hump per stakes cluster, and in our parameterized example two well-separated clusters produce two humps. $\square$

## A.5 The Deskilling Trap

We call the critical reliability of Proposition 5(iv) the *fold*, the reliability at which the skilled state and the threshold merge and vanish.

**Proposition A4** (Lopsided formation and the deskilling trap: full statement of Proposition 5)**.** *Parts* (i)–(ii) *let* $\theta < 1$*, with part* (ii) *under the Trust dominance of Assumption 3. Parts* (iii)–(iv) *hold at any* $\theta \in (0, 1]$*, and their accountability comparative statics inherit part* (ii)*'s hypothesis.* (i) No direct wedge on direction. *Accountability multiplies only the loss terms of equations* (1) *to* (3)*, and no loss term contains* $m_d$*. Conditional on the action and on the vetting effort, the marginal return to direction therefore carries no* $\theta$ *and coincides with the social one. One channel survives that conditioning: the optimized Vet-region return to direction is nondecreasing in* $\theta$*, so the wedge weakly lowers even direction's return, and only through the checking it discourages. The delegation share* $x_d$ *moves with* $\theta$*, but not in one direction. Where the worker is better than the tool (*$a > q$*) and the marginal delegated task is trusted, the wedge expands delegation and* $m_d^*$ *weakly exceeds its first-best level. Where a strong verifier delegates a vetted high-stakes tail, it pulls that tail back toward Solo.* (ii) The discounted bracket. *The worker values a caught error at* $V\varphi(m_d) + \theta L$ *where the planner values it at* $V\varphi(m_d) + L$*. Only the avoided-loss component is discounted, so the private return to checking is the first-best return scaled by* $\big[V\varphi(m_d) + \theta L\big] / \big[V\varphi(m_d) + L\big] \in (\theta, 1)$ *rather than by* $\theta$*. The wedge nonetheless bites on both margins. Vetting effort, the practice* $x_v$*, and hence* $m_v^*$ *all fall below the first-best, with the shortfall nondecreasing in* $(1 - \theta)$*.* (iii) The trap, and a floor under it. *For any setup cost* $f > 0$*,* $m_v = 0$ *is a fixed point of the verification law of motion* (15)*, and its basin has a floor. There is* $m_v^0 > 0$*, independent of* $m_d$*, below which no task in the portfolio is vetted, so verification decays to zero from every start below it. If the practice curve* $(\psi_v/\eta)\, x_v(m_v)$ *never reaches the* 45° *line the trap is global. If it clears the line, the least crossing* $m_v^\dagger$ *is an unstable threshold and the greatest is the stable skilled state* $m_v^*$*, with monotone convergence outside them and a start at* $m_v^\dagger$ *itself staying put. Under a sigmoid shape condition, meaning the curve is convex up to an inflection and concave above it, these are the* only *two crossings. Without it the crossings can alternate, and the model admits additional stable steady states at intermediate skill levels (Remark 1). Lowering* $\theta$ *shifts the curve down by part* (ii)*, raising the threshold and lowering the skilled state, and improving the AI does the same wherever* $\partial x_v/\partial q < 0$ *(Lemmas 3 to 5).* (iv) The fold. *The path is continuous, so skill drifts along the high branch and collapses to the trap only once reliability crosses the* fold $q_{\text{crit}}$*, at which the skilled steady state and the threshold merge and vanish, and past which neither survives (Lemma 6). Lowering* $\theta$ *raises* $m_v^\dagger$ *and enlarges the trap's basin. Improving the AI past the low-reliability trough of Proposition A5 enlarges it too. Direction, under the delegation floor of Assumption 4, has no analogue (Proposition 3(iii)). The floor itself survives the wedge on an interval of accountability levels containing* $\theta = 1$*, the internalized-stakes floor gap being affine in* $\theta$*: the interval is all of* $(0, 1]$ *unless* $q > a$ *and the zero-stakes gap* $V(q\varphi_0 - a) + \lambda(t_H - t_A)$ *is negative, and in our parameterized example it reaches below every accountability level the analysis visits.*

*Proof of Proposition 5, parts (i)–(ii).* (i) Under the wedge every shipped loss is scaled: the stage

payoffs carry $-(1-a)\theta L$ on Solo, $-(1-q)\theta L$ on Trust, and $-(1-q)(1-\rho)\theta L$ on Vet. None of these terms contains $m_d$, so at a fixed action and a fixed vetting effort the private and social payoffs differ by a term free of $m_d$, and their $m_d$-derivatives coincide: 0, $qV\varphi'$, and $V\varphi'[q+(1-q)\rho]$ on the three actions.

One channel survives the conditioning. *The effort channel.* Differentiating the private first-order condition $(1-q)(V\varphi+\theta L)\rho_e = \lambda c'(e)$ gives

$$\frac{de^w}{d\theta} = \frac{(1-q)L\,\rho_e}{D} \;\geq\; 0, \qquad \frac{de^w}{dm_d} = \frac{(1-q)V\varphi'\rho_e}{D} \;\geq\; 0, \qquad D \equiv \lambda c'' - (1-q)(V\varphi+\theta L)\rho_{ee} > 0, \tag{31}$$

the denominator being the second-order condition. Since $\rho$ is nondecreasing in $e$ (Assumption 1), the optimized Vet-region direction return $V\varphi'[q+(1-q)\rho(m_v, e^w(\theta); q)]$ is nondecreasing in $\theta$, strictly increasing wherever $\rho_e > 0$. The wedge therefore weakly lowers the return to direction as well, though only inside the Vet region and only through the scrutiny it discourages.

*The delegation share.* $x_d$ is an action share, and $\theta$ enters the payoff levels that set its boundaries. Because the private problem depends on $(\theta, L)$ only through $\theta L$, every boundary scales as $\bar{L}(\theta) = \bar{L}(1)/\theta$. At a trusted Solo/Trust margin, $\partial(u_{\text{Trust}} - u_{\text{Solo}})/\partial\theta = L(q-a)$, so where $a > q$, lowering $\theta$ widens the gap, and when the delegate set is a lower interval in the stakes the wedge stretches it by $1/\theta$, so $x_d$, hence $m_d^*$, weakly exceeds the first-best level. At a vetted Solo/Vet margin the envelope derivative is $\partial(u_{\text{Vet}} - u_{\text{Solo}})/\partial\theta = L\big[(1-a)-(1-q)(1-\rho(e^w))\big]$, positive wherever marginal detection is strong, $\rho(e^w) > (a-q)/(1-q)$. There the discount mutes the loss-avoidance motive that put the high-stakes task into vetted delegation, the wedge pulls the task back toward Solo, and the share can fall.

(ii) We compare brackets. The worker's value of catching an error on a task is $(1-q)\rho\,(V\varphi(m_d)+\theta L)$ against the planner's $(1-q)\rho\,(V\varphi(m_d)+L)$, so the private return is the social one scaled by

$$\frac{V\varphi(m_d)+\theta L}{V\varphi(m_d)+L} \;\in\; (\theta, 1) \qquad \text{for } L>0,\ \theta<1,$$

strictly above $\theta$ because the directed value $V\varphi$ is undiscounted, and strictly below 1 because the avoided loss is not.

*Intensive.* $e^*$ is strictly increasing in the bracket (Proposition 4(i)), so $e^w(\theta) < e^*$ for $L > 0$, with $de^w/d\theta > 0$ by equation (31). *Extensive.* The Vet-minus-Trust gap (5) at the private optimum, $G(\theta) = (1-q)\rho(e^w)(V\varphi+\theta L) - \lambda f - \lambda c(e^w)$, satisfies $dG/d\theta = (1-q)L\,\rho(e^w) \geq 0$ by the envelope theorem, so the set of tasks on which Vet beats Trust shrinks as $\theta$ falls. Under Assumption 3 the fallback wherever checking is not worth its cost is Trust, never Solo, so no offsetting Solo margin can enlarge the vetted set. Without the assumption the Vet/Solo boundary moves the other way whenever $\rho(e^w) < (a-q)/(1-q)$. Both margins therefore lower $x_v = \mathbb{E}[e^*\mathbf{1}\{\text{vet}\}]$ as $\theta$ falls, strictly wherever positive mass vets at $\rho_e > 0$. The practice curve shifts down, and by Lemma 4(d) below the threshold rises and the skilled state falls. □

**Lemma 3** (A floor under the trap)**.** *Fix the environment* $(q, \kappa, \theta)$ *and let the portfolio's stakes be*

*bounded, $L \le L_{\max}$. Set*

$$\bar{U} \equiv (1-q)\big(V\varphi(1) + \theta L_{\max}\big), \qquad \bar{e} \text{ by } \lambda c(\bar{e}) = \bar{U}, \qquad \bar{\rho}(m_v) \equiv \max_{e \in [0,\bar{e}]} \rho(m_v, e; q).$$

*Then $\bar{\rho}$ is continuous and nondecreasing with $\bar{\rho}(0) = 0$, and*

$$m_v^0 \;\equiv\; \sup\big\{m \in [0,1] :\; \bar{\rho}(m) < \lambda f/\bar{U}\big\} \;>\; 0$$

*has the property that at any verification stock $m_v < m_v^0$ no task in the portfolio is vetted, at any $m_d \in [0,1]$. Consequently $x_v \equiv 0$ on $[0, m_v^0)$, the map (15) reduces to $m_{v,t+1} = (1-\eta)m_{v,t}$ there, and every start with $m_{v,0} < m_v^0$ has $m_{v,t} \downarrow 0$ whatever the direction stock does. In particular $m_v = 0$ is a fixed point of equation (15) and the strip $[0,1] \times [0, m_v^0)$ lies in its basin.*

*Proof of Lemma 3. Effort is bounded.* As in the preamble, $\bar{e}$ exists, and any chosen effort lies in $[0, \bar{e}]$: the gross value of vetting at effort $e$ is $(1-q)\rho(m_v, e; q)(V\varphi(m_d) + \theta L) \le \bar{U}$, and Vet is chosen only if that covers $\lambda(f + c(e))$, which forces $c(e) < c(\bar{e})$.

*The floor.* $\bar{\rho}$ is continuous as the maximum of a continuous function over a compact interval. It is nondecreasing because $\rho$ is nondecreasing in $m_v$, and $\bar{\rho}(0) = 0$ because $\rho(0, e; q) = 0$ at every $e$ (Assumption 1). Since $\lambda f/\bar{U} > 0$, continuity at 0 gives $m_v^0 > 0$. For $m_v < m_v^0$ and any task in the portfolio, the Vet-minus-Trust gap (5) obeys

$$(1-q)\rho(m_v, e; q)\big(V\varphi(m_d) + \theta L\big) - \lambda c(e) - \lambda f \;\le\; \bar{U}\,\bar{\rho}(m_v) - \lambda f \;<\; 0$$

at every admissible $e$, so Trust strictly dominates Vet task by task and $x_v(m_d, m_v) = 0$. The reduced law of motion follows from equation (15), and $(1-\eta)^t m_{v,0} \downarrow 0$. The bound $m_v^0$ uses $\varphi(1)$ and $L_{\max}$ only, so it is uniform in $m_d$ and across the portfolio. □

**Lemma 4** (Crossings and monotone convergence)**.** *Hold $m_d$ fixed and write $F(m_v) = (\psi_v/\eta)x_v(m_v)$ and $T(m_v) = \min\{1, (1-\eta)m_v + \psi_v x_v(m_v)\}$. Suppose $x_v$ is continuous and nondecreasing in $m_v$ (the monotonicity and continuity established in the proof of Proposition 3(ii), which applies verbatim at $\theta < 1$ because the private problem depends on $(\theta, L)$ only through $\theta L$). Then:* (a) *$T$ is continuous and nondecreasing, $T(m_v) = (1-\eta)m_v + \eta F(m_v)$ wherever the cap is slack, and for $m_v < 1$, $\operatorname{sign}\big(T(m_v) - m_v\big) = \operatorname{sign}\big(F(m_v) - m_v\big)$. The cap $m_v = 1$ is a fixed point of $T$ if and only if $F(1) \ge 1$.* (b) *From any $m_{v,0}$ with $T(m_{v,0}) \le m_{v,0}$ the iterates fall monotonically to the greatest fixed point of $T$ below $m_{v,0}$. From any $m_{v,0}$ with $T(m_{v,0}) \ge m_{v,0}$ they rise monotonically to the least fixed point above it.* (c) *If $F(m_v) < m_v$ for every $m_v \in (0,1]$ the trap is global. Otherwise let $m_v^\dagger = \inf\{m_v : F(m_v) > m_v\}$ and $m_{v-}^* = \sup\{m_v : F(m_v) > m_v\}$. Both are fixed points of $T$, and $F \le \mathrm{id}$ on $[0, m_v^\dagger)$ and on $(m_{v-}^*, 1)$. Iterates started below $m_v^\dagger$ fall monotonically to the trap, and iterates started above $m_{v-}^*$ fall monotonically to the greatest fixed point of $T$ below the start. A start at $m_v^\dagger$ itself remains there for all $t$.* (d) (Comparative statics.) *If $F_1 \le F_2$ pointwise on $[0,1]$, both continuous, then $m_{v1}^\dagger \ge m_{v2}^\dagger$ and $m_{v-,1}^* \le m_{v-,2}^*$, with the trap global when the crossing set is*

*empty. In particular, lowering $\theta$ or, wherever $\partial x_v/\partial q < 0$, raising $q$ shifts the practice curve down, raising the threshold and lowering the skilled state.*

*Proof of Lemma 4.* (a) $T$ is the minimum of 1 and a sum of nondecreasing continuous functions. For $m_v < 1$, $T(m_v) \geq m_v$ iff $(1-\eta)m_v + \psi_v x_v \geq m_v$ (the cap binds only above $m_v$) iff $\eta m_v \leq \psi_v x_v = \eta F(m_v)$, and the reverse inequality is symmetric. At $m_v = 1$, $T(1) = 1$ iff $(1-\eta) + \psi_v x_v(1) \geq 1$ iff $F(1) \geq 1$. (b) If $T(m_0) \leq m_0$ then by monotonicity $T^2(m_0) \leq T(m_0)$ and inductively the iterates are nonincreasing. They are bounded below by 0, so they converge, and by continuity the limit is a fixed point. It is the greatest fixed point below $m_0$, since any fixed point $p \leq m_0$ satisfies $p = T^t(p) \leq T^t(m_0)$ for all $t$. The increasing case is symmetric. (c) If $F < \mathrm{id}$ on $(0,1]$ then $T(m_v) < m_v$ everywhere on $(0,1]$ by (a), the iterates fall, and the only fixed point they can reach is 0. Otherwise the set $\{F > \mathrm{id}\}$ is nonempty and open in $(0,1]$, and it excludes a neighborhood of 0 because $F \equiv 0$ on $[0, m_v^0]$ (Lemma 3), so $m_v^\dagger \geq m_v^0 > 0$. Continuity of $F - \mathrm{id}$ gives $F(m_v^\dagger) = m_v^\dagger$ and $F(m_{v-}^*) = m_{v-}^*$, and $F \leq \mathrm{id}$ off the open set, which is the stated sign pattern. By (a), both points are then fixed points of $T$. The convergence statements are (a) plus (b). (d) $\{F_1 > \mathrm{id}\} \subseteq \{F_2 > \mathrm{id}\}$, so the infimum of the first set is at least that of the second and its supremum at most. The pointwise shifts are Proposition 5(ii) ($\partial x_v/\partial\theta \geq 0$) and the maintained sign $\partial x_v/\partial q < 0$. □

**Lemma 5** (Two crossings under a sigmoid practice curve)**.** *Say the practice curve is* sigmoid *if*

$$\exists\, \hat{m} \in [0,1]: \quad F \text{ is convex on } [0,\hat{m}] \text{ and concave on } [\hat{m},1]. \tag{32}$$

*Suppose $F$ is continuous and nondecreasing with $F(0) = 0$, $F \equiv 0$ on $[0, m_v^0]$ for some $m_v^0 > 0$ (Lemma 3), equation* (32) *holds, and $F$ does not coincide with the 45° line on any interval. Then $F(m_v) = m_v$ has at most two solutions in $(0,1]$. When it has exactly two, $m_v^\dagger < m_v^*$, the lesser is an upcrossing and the greater a downcrossing, and the dynamics of equation* (15) *are the two-state geometry of Figure 7. Iterates from $[0, m_v^\dagger)$ fall to the trap, iterates from $(m_v^\dagger, 1]$ converge to the greatest fixed point of $T$, and $m_v^\dagger$ itself is stationary.*

*Proof of Lemma 5.* Write $g = F - \mathrm{id}$. On $[0,\hat{m}]$ the function $g$ is convex, so it is nonincreasing and then nondecreasing there, and on $[\hat{m},1]$ it is concave, so nondecreasing and then nonincreasing. Concatenated, $g$ is nonincreasing on $[0,a]$, nondecreasing on $[a,b]$, and nonincreasing on $[b,1]$ for some $a \leq \hat{m} \leq b$. Because $g(m_v) = -m_v$ on $[0, m_v^0]$ and $g$ is nonincreasing on $[0,a]$, $g < 0$ on $(0,a]$. On a monotone piece the zero set of $g$ is an interval on which $F = \mathrm{id}$, which the hypothesis excludes unless it is a single point. So $g$ has at most one zero $z_1$ on $[a,b]$, an upcrossing after which $g > 0$ on the rest of that piece, and at most one zero on $[b,1]$, a downcrossing: if $z_1 < b$ then $g(b) > 0$ and the nonincreasing piece crosses zero at most once, if $z_1 = b$ then $b$ is the only zero on $[b,1]$, and if no $z_1$ exists then $g < 0$ on all of $(0,1]$. Hence $g$ has at most two zeros in $(0,1]$, and with exactly two, $z_1 < z_2$, we have $g < 0$ on $(0,z_1)$, $g > 0$ on $(z_1,z_2)$, and $g < 0$ on $(z_2,1)$, making $z_1 = m_v^\dagger$ an upcrossing and $z_2 = m_v^*$ a downcrossing.

*Dynamics.* By Lemma 4(a), $T(m_v) < m_v$ on $(0, m_v^\dagger)$, $T(m_v) > m_v$ on $(m_v^\dagger, m_v^*)$, and $T(m_v) < m_v$ on $(m_v^*, 1)$. Lemma 4(b) then gives monotone convergence to 0 from below the threshold and to the

greatest fixed point of $T$ from above it. That fixed point is $m_v^*$ when $F(1) < 1$, and when $F(1) \geq 1$ concavity forces $z_2 = 1$ and the greatest fixed point is the cap. Stationarity at $m_v^\dagger$ is immediate. □

*Remark* 1 (Intermediate steady states). The shape condition (32) delivers the two crossings, and Assumption 1 does not imply it. A continuous nondecreasing practice curve may meet the 45° line any number of times, and the crossings then alternate: unstable, stable, unstable, stable, and so on up from the trap. Vetting activates mode by mode in the stakes, since each cluster of tasks carries its own verification skill at which checking begins to cover $\lambda f$, so a portfolio whose stakes are multi-modal has a practice curve that climbs in steps, and a step can straddle the diagonal. The intermediate stable states are portfolios on which checking is worth its cost over the high-stakes part alone. A worker who loses skill from the top state descends one rung, to the state at which they still check the tail and no longer check the routine bulk, and reaches the trap only from the final rung. Everything in Lemmas 3 and 4 survives without equation (32): the trap and its floor $m_v^0$, the least-unstable and greatest-stable structure, monotone convergence outside the outermost crossings, and the direction in which the curve shifts with $\theta$ and $q$. Uniqueness of the threshold is lost, and Figure 7 no longer reads as a complete phase diagram. In our parameterized example the condition holds under the uniform stakes the paper uses throughout, and two separated stakes clusters produce such intermediate states.

*Proof of Proposition 5, part (iii).* At $m_v = 0$, $\rho \equiv 0$ (Assumption 1) and $u_{\text{Vet}}(e) = u_{\text{Trust}} - \lambda(f + c(e)) < u_{\text{Trust}}$ at every effort, so Vet is never chosen and $x_v(0) = 0$. Lemma 3 strengthens this from a point to a strip. For any $f > 0$ there is $m_v^0 > 0$, uniform in $m_d$, below which no task in the portfolio is vetted, so the trap's basin contains $[0, 1] \times [0, m_v^0)$. The practice curve $F(m_v) = (\psi_v/\eta)x_v(m_v)$ is continuous and nondecreasing (proof of Proposition 3(ii), which applies unchanged at $\theta < 1$) and vanishes on $[0, m_v^0]$. Lemma 4 supplies the existence dichotomy, identifies the least crossing as the unstable threshold and the greatest as the stable skilled state, and gives monotone convergence outside them and the stationarity of a start at the threshold. That these are the *only* crossings is Lemma 5, under the sigmoid condition (32) and not otherwise. Remark 1 records what the model permits when equation (32) fails. Lowering $\theta$ shifts the curve down by part (ii), raising $m_v^\dagger$ and lowering $m_v^*$, and improving the AI does the same wherever $\partial x_v/\partial q < 0$. Those comparative statics are Lemma 4(d). The path is continuous in $m_{v,t}$ by equation (15). The one-dimensional argument holds $m_d$ fixed, and Corollary 2 replaces the fixed-$m_d$ reduction with the two-skill statement. Part (iv) is Lemma 6. □

**Lemma 6** (The fold). *Write $F(m_v; q) = (\psi_v/\eta)\, x_v(m_v; q)$ for the practice curve along a reliability path, with $m_d$ held fixed or evaluated at its own steady state, and $g(m_v; q) = F(m_v; q) - m_v$. Suppose $F$ is continuous in $(m_v, q)$, nonincreasing in $q$ and strictly decreasing in $q$ wherever $F > 0$, and at each $q$ satisfies the hypotheses of Lemma 5. Let $Q = \{q : F(m_v; q) > m_v \text{ for some } m_v\}$, assumed nonempty, and let $q_{\text{crit}} = \sup Q$, assumed to lie inside the reliability range (in the model $F(\cdot; 1) \equiv 0$, so $q_{\text{crit}} < 1$). Then:* (i) *for $q < q_{\text{crit}}$ the threshold $m_v^\dagger(q)$ is the unique upcrossing of Lemma 5, and the skilled state $m_v^*(q)$, the greatest fixed point of $T$, is its downcrossing, or the cap $m_v = 1$ when*

$F(1;q) \geq 1$; (ii) *for* $q > q_{\text{crit}}$, $F(\cdot;q) < \text{id}$ *on* $(0,1]$ *and every start falls to the trap;* (iii) *as* $q \uparrow q_{\text{crit}}$, $m_v^\dagger(q)$ *rises and* $m_v^*(q)$ *falls to a common limit* $m_c$, *at which* $g(\cdot;q_{\text{crit}}) \leq 0$ *with equality at* $m_c$. *If* $F$ *is differentiable in* $m_v$ *there and* $m_c < 1$, *the curve is tangent to the* $45^\circ$ *line at the merge,*

$$F(m_c;q_{\text{crit}}) = m_c, \qquad F_{m_v}(m_c;q_{\text{crit}}) = 1. \tag{33}$$

*Proof of Lemma 6.* Because $F$ is nonincreasing in $q$, $Q$ contains every reliability below any of its members, so $Q$ is an interval with right endpoint $q_{\text{crit}}$. (i) For $q < q_{\text{crit}}$, $q \in Q$, and Lemma 5 with Lemma 4(c) gives the crossings and the dynamics. (ii) For $q > q_{\text{crit}}$ pick $q' \in (q_{\text{crit}}, q)$. Then $q' \notin Q$, so $F(\cdot;q') \leq \text{id}$, and for $m_v > 0$ either $F(m_v;q) = 0 < m_v$ or $F(m_v;q) < F(m_v;q') \leq m_v$ by strict decrease where $F > 0$. Lemma 4(c) finishes. (iii) The set $\{F(\cdot;q) > \text{id}\}$ shrinks as $q$ rises, so $m_v^\dagger(q)$ rises and $m_v^*(q)$ falls by Lemma 4(d), and the monotone limits $\underline{m} \leq \bar{m}$ exist. Each crossing satisfies $F(m_v;q) = m_v$, so by continuity of $F$ in $(m_v, q)$ both $\underline{m}$ and $\bar{m}$ are zeros of $g(\cdot;q_{\text{crit}})$, with one adjustment: where the skilled state is the cap, $g(1;q) \geq 0$ on a left-neighborhood of $q_{\text{crit}}$, so continuity gives $g(1;q_{\text{crit}}) \geq 0$ and the nonpositivity shown next makes $\bar{m} = 1$ a zero. That function is nonpositive: if $g(m_v;q_{\text{crit}}) > 0$ at some $m_v$, continuity in $q$ would keep it positive slightly above $q_{\text{crit}}$, contradicting the definition of $q_{\text{crit}}$. Under equation (32) a nonpositive $g$ has at most one zero in $(0,1]$, since two zeros force $g > 0$ between them by the sign pattern of Lemma 5. Hence $\underline{m} = \bar{m} = m_c$. If $F$ is differentiable in $m_v$ and $m_c < 1$, then $m_c$ is an interior maximizer of $g(\cdot;q_{\text{crit}})$, so $g_{m_v}(m_c;q_{\text{crit}}) = 0$, which is equation (33). □

**Proposition A5** (Hiring against the trap: full statement of Corollary 2)**.** *Let* $T(m_d, m_v)$ *be the two-skill map* (15)*, which is monotone on* $[0,1]^2$ *by the componentwise monotonicity of the practice map (proof of Proposition 3(ii)), and let* $m^* = \lim_n T^n(1,1)$ *be its greatest fixed point, the skilled steady state, assumed to have* $m_v^* > 0$. (i) The skilled basin is an upper set. *Call a start* skilled *if its verification stock stays away from the trap,* $\liminf_t m_{v,t} > 0$, *and let* $B$ *be the set of skilled starts.* $B$ *is an upper set. If* $(m_{d,0}, m_{v,0}) \in B$ *then so is every* $(\tilde{m}_{d,0}, \tilde{m}_{v,0}) \geq (m_{d,0}, m_{v,0})$. *The trap side is therefore a lower set, and the boundary between them is the graph of a* nonincreasing *function* $\sigma : [0,1] \to [0,1]$. *Every state with* $m_{v,0} > \sigma(m_{d,0})$ *keeps a positive verification stock, and every state with* $m_{v,0} < \sigma(m_{d,0})$ *loses it. At* $m_{v,0} = \sigma(m_{d,0})$ *the state may remain on the separatrix, as it does at the threshold fixed point itself.* (ii) The two bounding levels. *Let* $\tau(m_d)$ *denote the one-dimensional threshold of Proposition 5(iii) computed with direction frozen at* $m_d$, *read as* 1 *where the frozen practice curve never clears the* $45^\circ$ *line, in which case the bound it enters is vacuous. Then*

$$m_v^0 \;\leq\; \tau(1) \;\leq\; \sigma(1) \;\leq\; \sigma(m_d) \;\leq\; \sigma(0) \;\leq\; \tau(0) \qquad \textit{for every } m_d \in [0,1],$$

*with* $m_v^0$ *the uniform dead-zone floor of Lemma 3. The hiring rule follows from these bounds. A hire with* $m_{v,0} > \tau(0)$ *keeps a positive verification stock whatever their direction skill, a hire with* $m_{v,0} < \tau(1)$ *loses it whatever their direction skill, and between the two levels the outcome depends on the joint state through the nonincreasing* $\sigma$. *In our parameterized example the two levels bracket the bar narrowly, so whether a hire escapes the trap turns almost entirely on verification skill.* (iii)

Comparative statics. *Both bounding levels move with the environment in the direction Lemma 4(d) gives. $\tau(0)$ and $\tau(1)$ rise as $\theta$ falls and, wherever $\partial x_v/\partial q < 0$, as $q$ rises, up to the fold of Lemma 6, beyond which no hire survives.*[25]

*Proof of Corollary 2.* (i) $T$ is monotone and continuous, and $T(1,1) \le (1,1)$, so $T^n(1,1)$ is nonincreasing and converges to a fixed point $m^*$, which dominates every fixed point. If $\tilde{m}_0 \ge m_0$ then monotonicity gives $T^t(\tilde{m}_0) \ge T^t(m_0)$ componentwise for every $t$, hence $\liminf_t \tilde{m}_{v,t} \ge \liminf_t m_{v,t}$. Membership of $B$ passes upward, so $B$ is an upper set and its complement $B^c$ a lower set. Define $\sigma(m_d) = \sup\{m_v : (m_d, m_v) \in B^c\}$, with $\sup \emptyset = 0$. If $m_v < \sigma(m_d)$ then $(m_d, m_v) \le (m_d, m'_v)$ for some $(m_d, m'_v) \in B^c$, and $B^c$ is a lower set, so $(m_d, m_v) \in B^c$. If $m_v > \sigma(m_d)$ then $(m_d, m_v) \notin B^c$, so it is in $B$. If $m_d < m'_d$ and $(m'_d, m_v) \in B^c$ then $(m_d, m_v) \le (m'_d, m_v)$ lies in $B^c$ too, so $\sigma(m_d) \ge \sigma(m'_d)$ and $\sigma$ is nonincreasing.

(ii) *The upper bound.* Since $x_v$ is nondecreasing in $m_d$, the verification component of the two-skill map dominates the map $T_0(m_v) = \min\{1, (1-\eta)m_v + \psi_v x_v(0, m_v)\}$ obtained by freezing direction at zero, and $T_0$ is monotone, so $m_{v,t} \ge T_0^t(m_{v,0})$ by induction. If $\tau(0) = 1$ the upper bound is vacuous, so suppose the frozen curve clears the line. The threshold $\tau(0)$ is then a crossing of that curve with the identity, so $T_0(\tau(0)) = \tau(0)$, and $s = (0, \tau(0))$ satisfies $T(s) \ge s$. The iterates from $s$ rise monotonically to a fixed point $p \ge s$, whose verification coordinate is at least $\tau(0) > 0$. Any start $m_0 \ge s$ then has $m_t \ge T^t(s)$ and so $\liminf_t m_{v,t} \ge p_v > 0$. So $s \in B$, and since $B$ is an upper set every $(m_{d,0}, m_{v,0})$ with $m_{v,0} > \tau(0)$ is in $B$, giving $\sigma(m_d) \le \tau(0)$ for all $m_d$. *The lower bound.* Symmetrically, $x_v(m_d, \cdot) \le x_v(1, \cdot)$ gives $m_{v,t} \le T_1^t(m_{v,0})$ with $T_1$ the direction-frozen-at-one map. For $m_{v,0} < \tau(1)$ the frozen iterates fall to zero (Lemma 4), so $m_{v,t} \to 0$, $\liminf_t m_{v,t} = 0$, and the state is not in $B$, giving $\sigma(m_d) \ge \tau(1)$. The floor $m_v^0 \le \tau(1)$ is Lemma 3, which puts the whole strip below it in the trap's basin.

(iii) $\tau(0)$ and $\tau(1)$ are least upcrossings of one-dimensional practice curves at fixed $m_d$, so Lemma 4(d) applies to each. □

## A.6 Sorting and Wages

Proofs for Section 6.

**Proposition A6** (Screening rotation: full statement of Proposition 6)**.** *Rank candidates for one position by $W$ in equation* (19)*, with $(d, L)$ drawn from the portfolio distribution of Section 2.*

(i) (Task skill.) *The firm's return to task skill is the stakes-weighted mass of tasks the worker still executes Solo,*

$$\frac{\partial W}{\partial a} = \mathbb{E}_{(d,L)}\big[(V+L)\,\mathbf{1}\{\textit{Solo}\}\big] \;\ge\; \big(1 - F_D(\kappa)\big)\big(V + \mathbb{E}[L \mid d > \kappa]\big) \;>\; 0 \qquad \textit{for } \kappa < 1. \tag{34}$$

*With difficulty independent of stakes the bound reads $(V + \mathbb{E}[L])(1 - F_D(\kappa))$. Widening scope weakly*

[25] The condition can fail at low reliability, where the bar is U-shaped in $q$. A more reliable tool first pulls tasks *into* the vet set faster than it drains checking from them, so in our parameterized example the bar falls to a trough at low reliability before it rises. The hiring-bar claims concern the range above the trough.

*lowers the return at any joint distribution, $\partial^2 W/\partial a\,\partial\kappa \le 0$, because the frontier tasks lose their a-slope as they are handed off. The return reaches zero only as $\kappa \to 1$ with the tool preferred on every task it can reach, as under Assumption 3.*

(ii) (Managerial ability.) *Both managerial returns are weakly positive at every parameter,*

$$\frac{\partial W}{\partial m_d} = V\varphi'(m_d)\,\mathbb{E}\big[q_{\text{sc}}\,\mathbf{1}\{\textit{delegated}\}\big] \ge 0, \qquad \frac{\partial W}{\partial m_v} = \mathbb{E}\big[(1-q)(V\varphi + L)\rho_{m_v}\,\mathbf{1}\{\text{Vet}\}\big] \ge 0, \quad (35)$$

*where the ship-correct probability $q_{\text{sc}}$ of equation (13) equals $q$ on a trusted task and $q + (1-q)\rho$ on a vetted one. The direction return is strictly positive if and only if the delegated mass is positive and delegated work ships correctly with positive probability, which holds whenever $q > 0$ and at $q = 0$ requires positive vetted mass. The verification return is strictly positive if and only if the vetted mass is positive and $q < 1$. A portfolio whose stakes are low enough that no task is worth checking has $\partial W/\partial m_v = 0$.*

(iii) (Rotation.) *Take any path along which $q$ and $\kappa$ are nondecreasing and the vetted-value condition (12) holds task by task. Then $\partial W/\partial a$ is nonincreasing along the path, and the direction return splits as*

$$\frac{\partial W}{\partial m_d} = \underbrace{V\varphi'\,q\,\mu_{\text{del}}}_{\textit{trusted floor}} + \underbrace{V\varphi'\,\mathbb{E}\big[(1-q)\rho\,\mathbf{1}\{\text{Vet}\}\big]}_{\textit{vetting bonus}}. \quad (36)$$

*The floor is the return to direction when no delegated work is checked. The bonus is the value checking adds, and it fades with the errors. Along the path the floor is nondecreasing, since $q$ rises and the delegated mass $\mu_{\text{del}}$ weakly grows, while the bonus vanishes as the tool improves. The pairwise ratio therefore satisfies*

$$\frac{\partial W/\partial m_d}{\partial W/\partial a} \;\ge\; \frac{V\varphi'\,q\,\mu_{\text{del}}}{\partial W/\partial a},$$

*a lower bound that is nondecreasing along the path. Under Assumption 3 the ratio diverges as $\kappa \to 1$. Within the managerial pair the mix follows Proposition 2, so which of the two the screen favors depends on reliability rather than scope.*

*Proof of Proposition 6.* Write $W = \mathbb{E}_{(d,L)}[u_{\text{Solo}} + R]$, an identity since the chosen value equals $u_{\text{Solo}} + R$ on every task by equation (4) ($R = 0$ where only Solo is available).

(i) $\partial u_{\text{Solo}}/\partial a = V + L$ on every task, while $\partial R/\partial a = -(V + L)$ wherever an AI action is chosen and 0 elsewhere (Proposition 1 and Corollary 1). The two cancel on every handed-off task, leaving the expression in equation (34), with $L$ inside the expectation because it varies across tasks. That identity uses no property of the joint distribution. For the bound, $\{d > \kappa\}$ is contained in the Solo set, so $\partial W/\partial a \ge \mathbb{E}[(V + L)\mathbf{1}\{d > \kappa\}] = (1 - F_D(\kappa))(V + \mathbb{E}[L \mid d > \kappa])$, and $1 - F_D(\kappa) > 0$ for $\kappa < 1$ because the difficulty density is strictly positive on $[0, 1]$. Independence lets $\mathbb{E}[L \mid d > \kappa]$ be replaced by $\mathbb{E}[L]$. Under negative dependence between difficulty and stakes the unconditional version is not a bound. For the cross-partial, differentiating under the integral in $\kappa$ moves only the tasks at the frontier, at density $f_D(\kappa)$, and those handed off lose their $a$-slope, giving $\partial^2 W/\partial a\,\partial\kappa = -f_D(\kappa)\mathbb{E}_L[(V + L)\mathbf{1}\{\text{delegated}\} \mid d = \kappa] \le 0$, strict when the frontier task is

delegated with positive probability. The return reaches zero only when the Solo set is null, which for $\kappa < 1$ it cannot be. As $\kappa \to 1$ it does so if and only if no in-scope task is kept Solo, which Assumption 3 delivers.

(ii) Both expressions in equation (35) integrate the channels of Proposition 1 over the sets on which they are live, the $m_v$ channel being the total effect by the envelope theorem. Each integrand is nonnegative, so each return is. For strictness on direction, $q_{\text{sc}} = q > 0$ on a trusted task whenever $q > 0$, and $q_{\text{sc}} = q + (1-q)\rho > 0$ on a vetted task because $\rho > 0$ there (effort is strictly positive wherever vetting is chosen, by Assumption 1 and $f > 0$). Hence $\partial W/\partial m_d > 0$ if and only if the delegated set has positive measure and $q_{\text{sc}} > 0$ on a positive-measure subset of it, which at $q = 0$ requires positive vetted mass. For strictness on verification, at $m_v > 0$ we have $\rho_{m_v} > 0$ on the vetted set (Assumption 1, effort being strictly positive there), so $\partial W/\partial m_v > 0$ if and only if the vetted mass is positive and $q < 1$. At $m_v = 0$, Assumption 1 signs $\rho_{m_v}$ only weakly and the return is nonnegative without being strict. In our parameterized example a portfolio of low-stakes tasks facing a reliable tool has no task worth checking, the vetted mass is zero, and $\partial W/\partial m_v = 0$.

(iii) *Task skill.* $\partial W/\partial a$ weights the Solo set, so it suffices that the Solo set weakly shrinks along the path. Raising $\kappa$ adds tasks to the reachable set and removes none, and a task's Solo status can only be lost. Raising $q$ leaves $u_{\text{Solo}}$ untouched, while $\partial u_{\text{Trust}}/\partial q = V\varphi + L > 0$ and $\partial u_{\text{Vet}}/\partial q = (V\varphi + L)[(1-\rho) + (1-q)\rho_q] \geq 0$, which is the vetted-value condition (12). The with-AI value is then nondecreasing in $q$ against a flat outside option and the Solo set shrinks. Hence $\partial W/\partial a$ is nonincreasing.

*Direction.* Split $q_{\text{sc}} = q + (1-q)\rho\,\mathbf{1}\{\text{Vet}\}$ to get equation (36). The floor $V\varphi' q\,\mu_{\text{del}}$ rises because $q$ rises and $\mu_{\text{del}}$ weakly grows, by the same two arguments as for the Solo set. The bonus is nonnegative and vanishes as the tool improves, and the argument needs nothing beyond equation (5). Since $\rho \leq 1$, the bonus is at most $V\varphi'(1-q)$ times the vetted mass, and with stakes bounded by $L_{\max}$ that mass is zero once $q > 1 - \lambda f/(V\varphi + L_{\max})$, since $G(q) \leq (1-q)(V\varphi + L_{\max}) - \lambda f < 0$ on every task there. The ratio bound follows by dividing a nondecreasing numerator by a nonincreasing positive denominator. For the limit, under Assumption 3 no in-scope task is Solo, so $\partial W/\partial a = (1 - F_D(\kappa))(V + \mathbb{E}[L \mid d > \kappa]) \to 0$ as $\kappa \to 1$, while $\partial W/\partial m_d \geq V\varphi' q F_D(\kappa)$.

*The full direction return.* $\partial W/\partial m_d$ can fall along such a path even where equations (12) and (13) hold. Those two conditions sign the two continuous channels, the mass entering the delegated set and the ship-correct probability on tasks that stay vetted. A third channel is discrete: where a task leaves the vetted set for Trust its payoffs agree, but its ship-correct probability drops by $(1-q)\rho$, and where a positive mass of tasks exits together the direction return drops with it, whatever the two conditions say. In our parameterized example both conditions hold with slack at a point where a narrow-stakes portfolio is about to leave the vetted region, and the direction return falls through the exit. The floor in equation (36) is monotone regardless, and the rotation claim uses the floor. □

**Proposition A7** (Sorting: full statement of Proposition 7)**.** *Impose Assumptions 1 and 2, and impose Assumption 3 on the portfolio of every job in the market, so that no in-scope task is done*

*Solo. Write the match surplus as*

$$W(m;\kappa,\ell) = \mathbb{E}_L\big[u_{\text{Solo}}\big] + F_D(\kappa)\, g(m,\ell), \qquad g(m,\ell) = \mathbb{E}_{L_0}\big[R(m;\ell L_0)\big],$$

*which is equation* (21)*, so that the type enters only through the second term. Write* $r_i(m,\ell) = \partial g/\partial m_i$ *for the per-in-scope-task return to skill* $i \in \{m_d, m_v\}$*, and collect the four cross-partials in the* sorting matrix

$$M \;=\; \begin{pmatrix} \partial^2 W/\partial m_d\,\partial\kappa & \partial^2 W/\partial m_d\,\partial\ell \\ \partial^2 W/\partial m_v\,\partial\kappa & \partial^2 W/\partial m_v\,\partial\ell \end{pmatrix} \;=\; \begin{pmatrix} f_D(\kappa)\, r_{m_d} & F_D(\kappa)\,\partial r_{m_d}/\partial\ell \\ f_D(\kappa)\, r_{m_v} & F_D(\kappa)\,\partial r_{m_v}/\partial\ell \end{pmatrix},$$

*with entries* $M_{d\kappa}, M_{d\ell}, M_{v\kappa}, M_{v\ell}$*.* (i) The four signs. *Every entry of* $M$ *is nonnegative.* $M_{d\kappa} \geq 0$ *(Proposition 2(iii)), strictly positive wherever delegated mass is positive and* $q > 0$*.* $M_{v\ell} > 0$ *wherever vetted mass is positive and* $q < 1$*, the portfolio aggregate of the stakes-weighting of Proposition 1(ii). The scope–verification cross effect* $M_{v\kappa} = f_D(\kappa)\, r_{m_v} \geq 0$ *is the per-task verification return itself. The stakes–direction cross effect* $M_{d\ell} \geq 0$ *runs through the extra scrutiny a higher-stakes task calls forth and the tasks it tips from Trust into Vet, and Assumption 3 signs it.* (ii) Monotone assignment on the rays. *Take either ray, jobs varying in* $\ell$ *at common* $\kappa$*, or in* $\kappa$ *at common* $\ell$*. The matched attribute is nondecreasing in the componentwise order on* $m = (m_d, m_v)$*, in some surplus-maximizing assignment always, and in every one wherever the relevant column of* $M$ *is strictly positive between the two matches. Where that column vanishes, assignments tie. Across workers who share direction skill the strongest verifiers take the highest-stakes jobs, and across workers who share verification skill the strongest directors take the widest-scope jobs.* (iii) Componentwise no-reversal in two dimensions. *In* any *surplus-maximizing assignment, no matched pair is reversed in both coordinates at once. If* $m \geq m'$ *componentwise with* $m \neq m'$*, their matched jobs cannot satisfy* $\kappa(m) < \kappa(m')$ *and* $\ell(m) < \ell(m')$*, the only exceptions being ties in which the two workers' surplus functions coincide on the jobs in question.* (iv) Vetting death and the collapse of the cross effects. *Let* $\Gamma(m_v, x; q)$ *denote the gap* (5) *maximized over scrutiny at internalized stakes* $x$*, the value of vetting over trusting net of the setup cost* $\lambda f$*, and let* $\bar{L}_0$ *be the top of the base-stakes support. Then*

$$\bar{q}(m,\ell) \;\equiv\; \sup\big\{q \in [0,1] : \Gamma\big(m_v, \ell\bar{L}_0; q\big) \geq 0\big\} \qquad (\sup\emptyset = 0) \tag{37}$$

*is well defined and nondecreasing in* $\ell$ *and in* $m_v$*. Below it the vetted set is an upper interval in the stakes that retreats toward the top of the support as* $q$ *rises. At* $\bar{q}(m,\ell)$ *it empties, and for every* $q > \bar{q}(m,\ell)$ *the match's sorting matrix collapses to the single surviving entry* $M_{d\kappa} = f_D(\kappa)\, qV\varphi' > 0$*. Because* $\bar{q}$ *is nondecreasing in both arguments, it is largest at the market's highest-stakes job held by its strongest verifier. Above that value the collapse holds at every match, verification is unpriced, and the assignment reduces to a one-dimensional positive assortative matching of direction against scope. Verification still earns a positive return just above the crossover* $q^*$ *of Proposition 2, so* $q^* \leq \bar{q}$*.* (v) The crossed pairing and elasticity dominance. *Take incomparable workers* $m, m'$ *with* $\Delta m_d > 0 > \Delta m_v$ *and jobs* $y, y'$ *with* $\Delta\kappa > 0 > \Delta\ell$*, and write* $\xi = \Delta m_d/|\Delta m_v|$ *and* $\beta = \Delta\kappa/|\Delta\ell|$ *for the type and job dispersion ratios. The "crossed" pairing sends the stronger director to the wider*

*scope and the stronger verifier to the higher stakes. With $M$ averaged over the rectangle between them, that pairing beats its swap if and only if the dispersion-weighted condition*

$$\xi\beta\, M_{d\kappa} + M_{v\ell} \;>\; \xi\, M_{d\ell} + \beta\, M_{v\kappa} \tag{38}$$

*holds. When $M_{d\kappa} > 0$ and $M_{d\ell} > 0$, equation* (38) *holds at every job dispersion $\beta > 0$ if and only if $\xi$ lies in the closed band $\left[M_{v\kappa}/M_{d\kappa},\, M_{v\ell}/M_{d\ell}\right]$ when $\det M > 0$, and for no $\xi$ otherwise. Such pairs therefore exist if and only if the verification return is the more stakes-elastic of the two,*

$$\frac{\partial}{\partial \ell}\left(\frac{r_{m_v}}{r_{m_d}}\right) > 0, \tag{39}$$

*the elasticity-dominance condition. When $M_{d\ell} = 0$ the condition holds for every $\xi > M_{v\kappa}/M_{d\kappa}$, and past $\bar{q}$ it holds at every dispersion. Lemma 7 derives equation* (39) *from primitive conditions on the detection technology and the stakes distribution.*

*Proof of Proposition 7.* Throughout, difficulty enters payoffs only through the gate $d \le \kappa$ and is independent of stakes, so $W = \mathbb{E}_L[u_{\text{Solo}}] + F_D(\kappa)\,\mathbb{E}_{L_0}[R(m;\ell L_0)]$, which is equation (21). With $d \sim U[0,1]$, $F_D(\kappa) = \kappa$ and $f_D \equiv 1$. Only the second term carries the type, which gives the form of $M$. The $\kappa$-column of $M$ is $f_D(\kappa)$ times a per-in-scope-task return, and the $\ell$-column is $F_D(\kappa)$ times its stakes-derivative. The stakes marginal is atomless with a density, so $\partial W/\partial m_i$ is continuous and, by the envelope theorem applied task by task, equals the expectation of $\partial R/\partial m_i$ at the chosen action, the boundary contributing nothing to the level because the two actions are payoff-equivalent there.

(i) *The $\kappa$-column.* $r_{m_d} = \mathbb{E}_{L_0}[\partial R/\partial m_d]$ with integrand $V\varphi'(q + (1-q)\rho)$ on a vetted task and $qV\varphi'$ on a trusted one (Proposition 1, Proposition 2(iii)). Both are nonnegative, and with delegation on positive mass and $q > 0$ the second is strictly positive, so $r_{m_d} > 0$. $r_{m_v} = \mathbb{E}_{L_0}[(1-q)(V\varphi + L)\rho_{m_v}\mathbf{1}\{\text{Vet}\}] \ge 0$, strictly positive where vetted mass is positive and $q < 1$.

*The $\ell$-column: verification.* Differentiate $r_{m_v}$ in $\ell$ at $L = \ell L_0$. On a task that stays vetted the integrand's $\ell$-derivative is $L_0\,\partial^2 R/\partial m_v\,\partial L > 0$ by Proposition 1(ii). At the boundary, $\frac{d}{dL}(u_{\text{Vet}}(e^*) - u_{\text{Trust}}) = (1-q)\rho > 0$ by the envelope theorem, so the vetted set only grows with $\ell$, and on the entering tasks $\partial R/\partial m_v$ jumps from 0 to a strictly positive value. Both contributions are nonnegative and the interior one is strictly positive wherever vetting is alive, so $\partial r_{m_v}/\partial \ell > 0$.

*The $\ell$-column: direction.* $\partial R/\partial m_d$ carries no explicit $L$ on either branch (proof of Proposition 1), so on the trusted set its $\ell$-derivative is zero, and on the vetted set it is

$$L_0\,\frac{\partial^2 R}{\partial m_d\,\partial L} = L_0\,V\varphi'(1-q)\,\rho_e\,\frac{de^*}{dL} \;\ge\; 0,$$

using $\rho_e \ge 0$ (Assumption 1) and the effort response $de^*/dL > 0$ derived in the proof of Proposition 1(ii). At the Trust-to-Vet boundary $\partial R/\partial m_d$ steps *up*, by $V\varphi'(1-q)\rho \ge 0$, and the vetted set grows with $\ell$, so that contribution is nonnegative too. There is no third boundary, because Assumption 3 removes Solo from the in-scope choice set at every task in the portfolio. Hence

$\partial r_{m_d}/\partial \ell \geq 0$, and it is zero where no task is vetted. Without Assumption 3, a job whose stakes rise past the point at which $a > q$ makes Solo attractive sends tasks from Trust to Solo, where $\partial R/\partial m_d$ drops from $qV\varphi'$ to zero, and the entry turns negative. In our parameterized example it does so at a calibration on the boundary of Assumption 3, while on cells that satisfy the assumption with slack all four entries stay nonnegative.

(ii)–(iii) Both parts are exchange arguments on the increasing differences that (i) delivers. Let $y = (\kappa, \ell)$ and $y' = (\kappa', \ell')$ with $y \leq y'$ componentwise, and $m \geq m'$ componentwise. Write $H(y) = W(m, y) - W(m', y)$ and integrate along axis paths, first in the job and then in the type,

$$\Delta \;\equiv\; H(y') - H(y) = \sum_{j \in \{\kappa, \ell\}} \int_{y_j}^{y'_j} \sum_{i \in \{d, v\}} \int_{m'_i}^{m_i} M_{ij}\, dm_i\, dy_j \;\geq\; 0,$$

since every $M_{ij} \geq 0$ by (i). The increasing differences also hold without smoothness of $F_0$: under Assumption 3, $R = [u_{\text{Trust}} - u_{\text{Solo}}] + \max\{0,\, u_{\text{Vet}}(e^*) - u_{\text{Trust}}\}$, the first bracket carries no $(m_i, L)$ interaction, and the second is the positive part of a gap that is nondecreasing in each argument and has the nonnegative cross-partials above. The positive part preserves increasing differences of such a gap, and expectation over any $F_0$ preserves them too.

For (ii), fix one job attribute and vary the other, so that $\Delta$ integrates the relevant column of $M$ alone. In a surplus-maximizing assignment no exchange improves surplus, so if $m \geq m'$ were matched below $m'$ on the ray, swapping the two jobs would change total surplus by $\Delta \geq 0$, a weakly improving exchange. Some optimum is therefore monotone (Becker, 1973), and every optimum is monotone wherever the column is strictly positive on a nondegenerate rectangle between the two matches. Where the column vanishes identically, as the $\ell$-column does past $\bar{q}$, assignments tie. Setting $m_d = m'_d$ gives the slice statement, in which the componentwise order on $m$ is the order on $m_v$ alone, and the scope ray is symmetric.

For (iii), suppose $m \geq m'$, $m \neq m'$, are matched to $y < y'$ in both coordinates. Optimality rules out an improving exchange, so $\Delta \leq 0$, hence $\Delta = 0$ and every term in the display vanishes. If $m_d > m'_d$, the $(m_d, \kappa)$ term integrates $M_{d\kappa} = f_D(\kappa) r_{m_d} > 0$ over a rectangle of positive area, since delegation on positive mass and $q > 0$ give $r_{m_d} > 0$ and $f_D > 0$ by Section 2, a contradiction. If $m_d = m'_d$ then $m_v > m'_v$, and the $(m_v, \ell)$ term integrates $M_{v\ell}$, strictly positive wherever vetted mass is positive, the same contradiction. Where vetted mass is zero throughout the rectangle, $\partial W/\partial m_v \equiv 0$ there, the two workers' surplus functions coincide on the jobs in question, and the exchange is a relabeling rather than a reversal. The vetted set is monotone in $(m_v, \ell)$, so the zero-mass hypothesis is checkable at the top corner alone.

(iv) $\Gamma(m_v, x; q)$ of equation (28) is a maximum over $e$ of $(1-q)\rho(m_v, e; q)(V\varphi + x) - \lambda c(e)$, less $\lambda f$. Pointwise in $e$ the maximand is nonincreasing in $q$, since both $(1-q)$ and, by Assumption 1, $\rho$ are, so $\Gamma$ is nonincreasing in $q$, and it is continuous in $q$ by the standing bound on effort. It is nondecreasing in $x$ with $\Gamma_x = (1-q)\rho \geq 0$, and at $q = 1$ the maximand is $-\lambda c(e) \leq 0$, so $\Gamma(m_v, x; 1) = -\lambda f < 0$. The set $\{q : \Gamma(m_v, \ell \bar{L}_0; q) \geq 0\}$ is therefore a closed lower interval and equation (37) is its right endpoint. Monotonicity of $\Gamma$ in $x$ and in $m_v$ makes $\bar{q}$ nondecreasing in $\ell$ and

in $m_v$. The vetted set is the upper interval $\{L \geq \bar{x}(m_v; q)\}$ with $\bar{x}(m_v; q) = \inf\{x : \Gamma(m_v, x; q) \geq 0\}$ nondecreasing in $q$, again because $\Gamma$ is nonincreasing in $q$, so the vetted set shrinks toward the top of the support as the tool improves. It is empty, up to the single top-of-support point at the boundary case, if and only if $\bar{x} \geq \ell \bar{L}_0$, that is when $q > \bar{q}(m, \ell)$. On that range every in-scope task is Trusted under Assumption 3, so $R = qV\varphi(m_d) - (1-q)L - \lambda t_A - u_{\text{Solo}}$, whose $m_d$-derivative $qV\varphi'$ carries no $L$ and whose $m_v$-derivative is zero. Hence $r_{m_v} \equiv 0$, $\partial r_{m_v}/\partial \ell \equiv 0$ and $\partial r_{m_d}/\partial \ell \equiv 0$, leaving $M_{d\kappa} = f_D(\kappa) qV\varphi' > 0$ as the only surviving entry. The collapse is per match, and it covers the whole market once $q$ clears the supremum of $\bar{q}$, attained at the highest-stakes job held by the strongest verifier. For the ordering $q^* \leq \bar{q}$, past $\bar{q}$ the verification return is identically zero while direction's is $qV\varphi' > 0$, so the comparison $\Delta^*$ of equation (9) is strictly negative there and its crossing must sit weakly below $\bar{q}$. It sits strictly below wherever the vetted set vanishes continuously, the verification return then being positive just under $\bar{q}$. In our parameterized example the scope–verification cross effect vanishes at $\bar{q}$, as equation (37) predicts, and not at the crossover, which sits well below it.

(v) With $u = m - m'$ and $v = y - y'$, the surplus difference between the crossed pairing and its swap is $\iint u^\top M v$. At the sign pattern $u_{m_d} > 0 > u_{m_v}$, $v_\kappa > 0 > v_\ell$ the integrand is $|u_{m_v}||v_\ell|\,(\xi\beta M_{d\kappa} + M_{v\ell} - \xi M_{d\ell} - \beta M_{v\kappa})$ with $\xi = |u_{m_d}|/|u_{m_v}|$ and $\beta = |v_\kappa|/|v_\ell|$, which is equation (38). Written as $\beta(\xi M_{d\kappa} - M_{v\kappa}) + (M_{v\ell} - \xi M_{d\ell}) > 0$, it holds for every $\beta > 0$ if and only if both brackets are nonnegative with at most one of them zero. For $M_{d\kappa}, M_{d\ell} > 0$ that is the closed band $\xi \in [M_{v\kappa}/M_{d\kappa}, M_{v\ell}/M_{d\ell}]$ when its endpoints are strictly ordered, that is, when $\det M > 0$, since at each endpoint the other bracket equals $\det M$ divided by a positive entry. When $\det M \leq 0$ no $\xi$ qualifies, the interval at $\det M = 0$ degenerating to the single point at which both brackets vanish. The determinant $\det M = f_D F_D\big(r_{m_d}\,\partial r_{m_v}/\partial \ell - r_{m_v}\,\partial r_{m_d}/\partial \ell\big)$ shares the sign of $\partial(r_{m_v}/r_{m_d})/\partial \ell$, which is equation (39). Past $\bar{q}$, $M_{d\ell} = M_{v\kappa} = M_{v\ell} = 0$ and the bracket reduces to $\xi\beta M_{d\kappa} > 0$ at every dispersion. In our parameterized example the stakes–direction cross effect is a small fraction of the verification–stakes complementarity, the comparison the text quantifies, so the band is wide, and $\det M > 0$ wherever verification is alive across a grid of types and jobs. Pairing higher skill with higher stakes on each coordinate separately is not optimal there at moderate reliability and is at high reliability. □

The next lemma derives the elasticity-dominance condition (39) from two primitive conditions, one on the detection technology and one on the stakes distribution. The second cannot be dropped: in our parameterized example two well-separated stakes clusters defeat the condition, as they defeat the single peak of Lemma 2(d).

**Lemma 7** (Elasticity dominance under separable detection and log-concave stakes)**.** *Impose Assumptions 1 and 2, impose Assumption 3 on the job's portfolio, and set $\theta = 1$. Fix a reliability $q \in (0,1)$ and a worker $(m_d, m_v)$, with base stakes $L_0 \sim F_0$ and $L = \ell L_0$ as in Proposition A7. Add two conditions.* (Sep) Separable detection. *$\rho(m_v, e; q) = a(m_v; q)\, b(e; q)$, with $a(0; q) = 0$ and $a' > 0$ on $(0, 1]$.* (L) Log-concave log stakes. *The density of $\ln L_0$ is log-concave. When $f_0$ is differentiable this says the density elasticity $uf_0'(u)/f_0(u)$ is nonincreasing. Then $r_{m_v}/r_{m_d}$ is nondecreasing in*

*$\ell$ at every job, and at every job with positive vetted mass the elasticity-dominance condition* (39) *holds and* $\det M > 0$. *The parametric family of Section 2 satisfies* (Sep) *with* $a = m_v(1-\delta q)$ *and* $b = 1 - e^{-\gamma e}$. *Uniform, power, and log-normal base stakes satisfy* (L), *as do gamma and Weibull stakes truncated to the portfolio's bounded support, and* $\mathrm{Beta}(\alpha, \beta)$ *with* $\beta \geq 1$.

*Proof.* Under Assumption 3 every in-scope task is Trusted or Vetted (proof of Proposition 4(iii)), and at $\theta = 1$ the vetted set is $\{L \geq \bar{x}(m_v)\}$ with the threshold $\bar{x}$ of Lemma 2 free of $\ell$. Write $x = \ell L_0$ for a task's stakes, $\rho^*(x) = \rho(m_v, e^*(x); q)$ for detection at the optimal scrutiny, and

$$J_0(\ell) = \mathbb{E}_{L_0}\big[\rho^*(x)\,\mathbf{1}\{x \geq \bar{x}\}\big], \qquad J_1(\ell) = \mathbb{E}_{L_0}\big[x\,\rho^*(x)\,\mathbf{1}\{x \geq \bar{x}\}\big]$$

for the detection mass and the detection-weighted stakes on the vetted set. $\rho^*$ is nondecreasing, with $d\rho^*/dx = \rho_e\, de^*/dL \geq 0$ by the effort response in the proof of Proposition 1(ii), and $\rho^* > 0$ on the vetted set because effort is strictly positive wherever vetting is chosen. A worker who vets on positive mass has $m_v > 0$, since at $m_v = 0$ vetting recovers nothing at any scrutiny, so $a(m_v) > 0$ there.

*Step 1: the two returns.* From the proof of Proposition A7(i), $r_{m_d} = V\varphi'\big(q + (1-q)J_0\big)$, every in-scope task being delegated. Under (Sep), $\rho_{m_v} = (a'/a)\,\rho$ pointwise in $e$, so the verification return uses the same two aggregates,

$$r_{m_v} = (1-q)\,\mathbb{E}_{L_0}\big[(V\varphi + x)\,\rho_{m_v}\,\mathbf{1}\{x \geq \bar{x}\}\big] = \frac{a'(m_v)}{a(m_v)}\,(1-q)\,\big(V\varphi\,J_0 + J_1\big).$$

*Step 2: the ratio is monotone.* Dividing and factoring,

$$\frac{r_{m_v}}{r_{m_d}} = \frac{(a'/a)(1-q)}{V\varphi'}\left(V\varphi + \frac{J_1}{J_0}\right)\left(\frac{q}{J_0} + 1 - q\right)^{-1} \qquad \text{wherever } J_0 > 0.$$

$J_0$ is nondecreasing in $\ell$, because raising the scale grows the vetted set $\{L_0 \geq \bar{x}/\ell\}$ and raises $\rho^*(\ell L_0)$ pointwise. For $J_1/J_0$, note it is the mean of $x$ under the measure with density proportional to $\rho^*(x)\,f_0(x/\ell)/\ell$ on $\{x \geq \bar{x}\}$. Compare scales $\ell' > \ell$. The likelihood ratio of the two measures is $(\ell/\ell')\,f_0(x/\ell')/f_0(x/\ell)$, the common weight $\rho^*$ and the common truncation at $\bar{x}$ dropping out. In log stakes the two densities are location shifts of the density $h_0$ of $\ln L_0$, and for log-concave $h_0$ the shifted pair is ordered by likelihood ratio, since $\frac{d}{dt}\ln\big[h_0(t-c')/h_0(t-c)\big] = (\ln h_0)'(t-c') - (\ln h_0)'(t-c) \geq 0$ for $c' \geq c$. The likelihood-ratio order implies first-order stochastic dominance and so orders the means (Milgrom, 1981). Hence $J_1/J_0$ is nondecreasing in $\ell$, both factors rise, and $r_{m_v}/r_{m_d}$ is nondecreasing. At a job whose vetted mass is zero, $J_0 = J_1 = 0$ and the ratio is zero, weakly below its value at any larger scale.

*Step 3: strict dominance.* At a job with positive vetted mass, differentiate under the integral in $L_0$, which is legitimate because $e^*$ is continuously differentiable in $x$ and (L) makes $f_0$ continuous on the interior of its support. Raising $\ell$ raises each vetted task's stakes and detection and enlarges the vetted set, so $J_0' \geq 0$ and $J_1' > 0$, the second strict because $\rho^* > 0$ on a vetted set of positive

mass. Substituting the Step 1 formulas into the determinant of Proposition A7(v),

$$\det M = f_D F_D \big(r_{m_d}\, \partial r_{m_v}/\partial \ell - r_{m_v}\, \partial r_{m_d}/\partial \ell\big) = f_D F_D\, V\varphi' \,\frac{a'}{a}\,(1-q)\,N,$$
$$N = q\big(V\varphi\, J_0' + J_1'\big) + (1-q)\big(J_1' J_0 - J_1 J_0'\big).$$

Step 2 gives $J_1' J_0 - J_1 J_0' = J_0^2 (J_1/J_0)' \geq 0$, so $N \geq q\, J_1' > 0$. Every prefactor is strictly positive, so $\det M > 0$, which is equation (39). □

# B Additional Results, General Versions, and Extensions

None of the results collected here is stated in the main text. Conventions on assumptions and family verification follow Appendix A.

## B.1 The Rare-Catastrophic Regime

**Proposition 9** (Atrophy and the rare-catastrophic regime)**.** *Let verification skill be built on a routine bulk of stakes $\ell$, and consider a rare task of stakes $L$ lying outside that bulk. Write $\bar{\rho}(m_v; q) = \lim_{e\to\infty} \rho(m_v, e; q)$ for the detection ceiling, the most a checker of skill $m_v$ can catch at any scrutiny. Parts* (ii) *and* (iii) *impose*

$$\bar{\rho}(m_v; q) \;>\; \frac{a-q}{1-q}, \tag{40}$$

*which puts the vetted tool's residual error rate below the worker's own and so keeps the rare task delegated rather than taken back in hand. It is vacuous whenever $q \geq a$, the regime of a tool reliable enough to be worth handing work to. It binds only for a worker whose own accuracy outruns the tool's, the same worker for whom Assumption 3 fails on the routine portfolio.* (i) *As the AI grows reliable on the routine work ($(1-q)\ell \to 0$), vetting it no longer covers its setup cost and verification skill atrophies into the trap of Proposition 5. The wedge lowers vetting at every $q$ (Proposition 5(ii)), so the set of reliabilities at which a skilled state exists shrinks and, by Lemma 6, the engineer's high branch ends no later than a full-stakes benchmark's, and strictly earlier in our parameterized example (Figure 8, left).* (ii) The uncaught loss on the rare task. *The engineer's expected uncaught loss on the rare task is $(1-q)L\,(1-\rho(m_v))$. At bounded stakes it vanishes as $q \to 1$. Along a* tail family*, a sequence with $q_n \to 1$ and $L_n \to \infty$ chosen so that $(1-q_n)L_n \to c > 0$, it converges to $c$. The whole of the expected catastrophic cost goes uncaught, because $m_v$ has lapsed in the trap and $\rho \to 0$. A benchmark worker who keeps $m_v$ alive, and for whom vetting the rare task still covers its setup cost, carries $c\,\big(1-\rho(m_v, \hat{e})\big) < c$ at the bounded effort $\hat{e}$ the finite tail incentive supports. The atrophy costs the insurance $c\,\rho(m_v, \hat{e})$ that the skill used to provide. The loss diverges only along a family on which $(1-q_n)L_n$ itself diverges.* (iii) The static split, at fixed $q$ and $m_v > 0$, as $L \to \infty$. *The extensive band of Proposition 4(iii) closes. Both parties prefer Vet to Trust, each one's value of catching errors growing linearly in $L$ against the fixed setup cost $\lambda f$, and both prefer Vet to*

*Solo by equation* (40)*. In our parameterized example the intensive gap does not close: the effort shortfall converges to a positive constant while the static loss grows like* $\ln L$*. Parts* (ii) *and* (iii) *send different objects to their limits:* (iii) *fixes the tool and the skill and sends one task's stakes to infinity, while* (ii) *sends the tool to perfection along a family whose stakes rise to hold* $(1-q)L$ *fixed.*

*Proof of Proposition 9.* (i) The private value of vetting a routine task is bounded by $(1-q)(V\varphi+\theta\ell)$, which falls below the setup cost $\lambda f$ as $(1-q)\ell \to 0$, so routine vetting ceases, $x_v \to 0$, and $m_v$ falls below $m_v^\dagger$ into the trap (Proposition 5(iii)). By Proposition 5(ii) the wedge lowers $x_v$ at every $q$, so by Lemma 6 the fold under $\theta < 1$ arrives no later than under $\theta = 1$, and in our parameterized example the shift is strict (Figure 8).

(ii) On a rare task of stakes $L$ the expected uncaught loss is $(1-q)L(1-\rho(m_v))$ with $\rho \le \bar\rho(m_v; q)$. For bounded $L$, $(1-q)L \to 0$ and the loss vanishes. Along the tail family, $(1-q_n)L_n \to c$ while $m_v \to 0$ in the trap, so $\rho \to 0$ by $\rho(0, e; q) = 0$ (Assumption 1) and continuity, and the product converges to $c$. Detection has collapsed, so skill insures none of the expected catastrophic cost. Along a family holding $(1-q)L$ bounded the uncaught loss is bounded, and it diverges exactly when $(1-q_n)L_n \to \infty$.

(iii) Fix $m_v > 0$ and decompose the static loss across the regimes of Proposition 4(iii). *Vet beats Trust.* Each party's value of catching errors, $(1-q)\rho(e)(V\varphi+\theta L)$ with $\theta = 1$ for the planner, evaluated at any fixed positive effort, grows linearly in $L$ against the fixed $\lambda f$, so the gap (5) at the optimum diverges and both vet rather than trust. *Vet beats Solo.* By equation (40) and $\rho(m_v, \cdot\,; q) \uparrow \bar\rho$, there is a fixed effort $\bar e$ with $\rho(m_v, \bar e; q) > (a-q)/(1-q)$. Revealed preference gives $u_{\mathrm{Vet}}(e^w) \ge u_{\mathrm{Vet}}(\bar e)$, and

$$u_{\mathrm{Vet}}(\bar e) - u_{\mathrm{Solo}} = \theta L\Big[(1-a) - (1-q)\big(1-\rho(m_v, \bar e; q)\big)\Big] + V\varphi\big[q + (1-q)\rho(m_v, \bar e; q)\big] - aV + \lambda\big(t_H - t_A - f - c(\bar e)\big),$$

whose bracket is strictly positive by the choice of $\bar e$ while every remaining term is a constant in $L$. The right-hand side diverges to $+\infty$ at rate $L$, so Vet beats Solo for all $L$ large, at $\theta = 1$ and at any $\theta > 0$ alike. The condition (40) carries no $\theta$, because $\theta$ multiplies the whole loss bracket and so cannot change its sign. Assumption 3 is imposed on the routine portfolio, the rare task sits outside it, and Solo has to be beaten on its own terms.

*The intensive gap.* The two first-order conditions are $\lambda c'(e^*) = (1-q)(V\varphi + L)\rho_e(e^*)$ and $\lambda c'(e^w) = (1-q)(V\varphi + \theta L)\rho_e(e^w)$, each with a unique interior solution (Assumption 1). In our parameterized example each effort diverges like $\gamma^{-1}\ln L$ while their gap converges to $-\ln\theta/\gamma$, and the static loss $S(e^*) - S(e^w)$ within the vet regime grows as

$$\mathrm{WL}_{\mathrm{static}}(L) = \frac{\lambda}{\gamma^2}\Big(\frac{1}{\theta} - 1 + \ln\theta\Big)\ln L + O(\ln\ln L),$$

with a coefficient strictly positive for every $\theta < 1$, since $h(\theta) = 1/\theta - 1 + \ln\theta$ has $h(1) = 0$ and $h'(\theta) = (\theta-1)/\theta^2 < 0$ on $(0, 1)$. □

The wedge preserves Trust dominance: if Assumption 3 holds at $\theta = 1$ on a set of tasks, it holds at every $\theta \in (0, 1]$ on the same set. The Trust-minus-Solo gap is affine in $\theta$. At $\theta = 0$ it is the first inequality of equation (22), and at $\theta = 1$ it is nonnegative by the two inequalities and linearity in $L$, so it is nonnegative in between. We use this preservation without further comment below.

### B.2 Mispriced Verification: The Wedge in the Assignment

Section 6 states Corollary 3 in prose. We state and prove it here. Its final part replaces the concentration claim with an eventual-dominance result, and everything rests on the wedge derivative of the match surplus. Under the wedge the derivative of social value with respect to verification skill, taken at the engineer's own choices, is not the envelope expression of Proposition A7(i). The private effort $e^w$ is optimal against $\theta L$ while $W^w$ is valued at $L$, so no envelope removes the effort response, and the private vetting set moves with $m_v$. Impose Assumption 3, let stakes have a density $g_L$, and let the private vet set be $\{L \geq \bar{L}^w(m_v)\}$ as in Lemma 2. Then

$$\begin{aligned}\frac{\partial W^w}{\partial m_v} = &\underbrace{F_D(\kappa)\,\mathbb{E}_L\Big[(1-q)\big(V\varphi + L\big)\rho_{m_v}(m_v, e^w)\,\mathbf{1}\{\mathrm{Vet}^w\}\Big]}_{\text{direct}} \\ &+ \underbrace{F_D(\kappa)\,(1-\theta)\,\mathbb{E}_L\Big[(1-q)L\,\rho_e(m_v, e^w)\,\frac{\partial e^w}{\partial m_v}\,\mathbf{1}\{\mathrm{Vet}^w\}\Big]}_{\text{effort response}} \\ &+ \underbrace{F_D(\kappa)\,(1-\theta)(1-q)\,\bar{L}^w\,\rho\big(m_v, e^w(\bar{L}^w)\big)\,g_L(\bar{L}^w)\Big(-\frac{d\bar{L}^w}{dm_v}\Big)}_{\text{extensive boundary}}. \end{aligned} \tag{41}$$

The first line is the fixed-effort, fixed-action-set direct component. The second follows from substituting the private first-order condition $\lambda c'(e^w) = (1-q)(V\varphi + \theta L)\rho_e$ into the total derivative, which leaves only the uninternalized share $(1-\theta)L$. The third is the social value of the marginal task the engineer newly vets, positive because the engineer is privately, not socially, indifferent there. All three terms are nonnegative under Assumption 1, with $\rho_{m_v} \geq 0$, $\partial e^w/\partial m_v \geq 0$ because $\rho_{m_v e} \geq 0$, and $d\bar{L}^w/dm_v < 0$ by Lemma 2(b). The last two carry an explicit factor $(1-\theta)$ and vanish identically at $\theta = 1$, where the envelope theorem applies. The $m_d$-derivative admits the same three-term split, with $V\varphi'$ in the direct term and its own correction terms through $\partial e^w/\partial m_d$ and the movement of the private threshold in $m_d$, likewise nonnegative and vanishing at $\theta = 1$.

**Corollary 3** (Mispriced verification)**.** *Let $\theta < 1$, impose Assumptions 1 and 3, and write $W^w$ for the social value of a match evaluated at the worker's private choices and $\mathcal{W}(m_v; \kappa, \ell) = W^{\mathrm{FB}} - W^w \geq 0$ for the per-match welfare loss.* (i) The scope column survives intact. *At every $\theta \in (0, 1]$, $\partial^2 W^w/\partial m_d\,\partial\kappa \geq 0$ and $\partial^2 W^w/\partial m_v\,\partial\kappa \geq 0$, the latter being $f_D(\kappa)$ times the per-in-scope-task object of equation* (41)*, all three of whose terms are nonnegative.* (ii) The stakes column. *Each entry of the stakes column of $M^w$ equals a nonnegative direct part plus the stakes-derivatives of the correction terms. The direct part is the interior effort channel plus the Trust-to-Vet boundary, both signed*

*as in Proposition A7(i). The correction terms are those of equation* (41) *for the* $m_v$ *entry and their analogues for the* $m_d$ *entry, and they carry an explicit factor* $(1-\theta)$ *and are not signed. The componentwise no-reversal result of Proposition A7(iii) therefore survives the wedge wherever those two entries remain nonnegative. Suppose the first-best stakes column is bounded below by* $c > 0$ *on a compact set of types and jobs, and that the entries of the wedge matrix* $M^w$ *exist and are continuous in* $\theta$ *at* $\theta = 1$*, uniformly on that set. Then there is* $\bar{\theta} < 1$ *such that both entries stay nonnegative for every* $\theta \in (\bar{\theta}, 1]$*. The two entries behave differently. In our parameterized example the verification–stakes entry stays positive at moderate accountability, weakened but not reversed, so the* ray *result of Proposition A7(ii), that across workers sharing direction skill the strongest verifiers take the highest-stakes jobs, survives the wedge there. The stakes–direction entry is not robust: it falls as* $\theta$ *falls and changes sign at moderate accountability, which breaks a* sufficient *condition for the two-dimensional theorem, though no computed assignment carries a strict reversal.* (iii) Concentration, read along the assignment. *Along the matched assignment the loss* $\mathcal{W}(m_v(\ell), \ell)$ *rises across the over-reliance band of Lemma 2 and drops at its top, so it is not monotone in the stakes. Part* (iv) *gives the eventual-dominance form of concentration instead.* (iv) Eventual dominance. *Let* $q \geq a$*, so that Assumption 3 is satisfiable at every stakes scale, and let* $\ell \mapsto m_v(\ell)$ *be a matched selection along the stakes ray that is nondecreasing on* $[\ell_*, \infty)$ *(Proposition A7(ii) provides one) and strictly positive at* $\ell_*$*, so that* $\underline{m_v} \equiv m_v(\ell_*) > 0$ *bounds it below there. Suppose in addition that the per-match loss grows without bound in the stakes scale, uniformly in the type,*

$$\liminf_{\ell \to \infty} \inf_{m_v \in [\underline{m_v}, 1]} \mathcal{W}(m_v; \kappa, \ell) \;=\; \infty, \tag{42}$$

*as it does in our parameterized example. Then for every* $\ell_0$ *there is* $\ell_1$ *such that* $\mathcal{W}(m_v(\ell), \ell) > \mathcal{W}(m_v(\ell'), \ell')$ *for every* $\ell \geq \ell_1$ *and every* $\ell' \leq \ell_0$*. In our parameterized example the map* $\ell \mapsto \mathcal{W}(m_v(\ell), \ell)$ *is not monotone at moderate stakes.*

*Proof of Corollary 3.* (i) On a vetted task the $m_d$-derivative of social value at the private action is the split noted below equation (41), each term nonnegative, and on a trusted task $s_{m_d} = qV\varphi' > 0$. Solo does not arise, the wedge preserving Trust dominance (appendix B.1). Hence $r^w_{m_d} \geq 0$ and, by the separation (21) applied to $W^w$, $\partial^2 W^w / \partial m_d \, \partial \kappa \geq 0$. For verification, equation (41) is $F_D(\kappa)$ times a per-in-scope-task object that carries no $\kappa$, so differentiating in $\kappa$ gives $f_D(\kappa)$ times that object. All three of its terms are nonnegative under Assumption 1 and Lemma 2(b), so $\partial^2 W^w / \partial m_v \, \partial \kappa \geq 0$.

(ii) Differentiate the direct term of equation (41) in $\ell$. On the private vetted interior it gives $L_0(1-q)\big[\rho_{m_v} + (V\varphi + L)\rho_{m_v e}\, de^w/dL\big] \geq 0$, using $\rho_{m_v e} \geq 0$ and $de^w/dL = \theta(1-q)\rho_e / D > 0$, with $D$ the second-order-condition denominator of equation (31). Because $\frac{d}{dL}(u^w_{\text{Vet}} - u^w_{\text{Trust}}) = \theta(1-q)\rho > 0$, the private vetted set grows with $\ell$, so the boundary contribution is a nonnegative jump. The same two channels sign the direct part of the $m_d$ entry, with the boundary jump $V\varphi'(1-q)\rho$. The remaining two terms of equation (41) each carry the factor $(1-\theta)$, and their $\ell$-derivatives involve $\partial e^w / \partial m_v$, the density $g_L$ at the moving threshold, and $d\bar{L}^w / dm_v$, none of which is signed in $\ell$. Where those entries stay nonnegative, the argument of Proposition A7(iii) applies verbatim to $W^w$.

Under the stated hypotheses $M^w \to M$ uniformly on the compact set as $\theta \to 1$, and $M$'s stakes column is at least $c$ there, so both entries are positive for $\theta$ near 1. Beyond that the claim is a family fact: in our parameterized example the direction entry falls as $\theta$ falls and crosses zero at moderate accountability, while the discretized assignment carries no strict componentwise reversal, so the sufficient condition fails before the conclusion does.

(iii) $\mathcal{W} \geq 0$ by revealed preference, and $\mathcal{W} = 0$ at $\theta = 1$. Lemma 2 read at fixed stakes gives the non-monotonicity in $m_v$. The loss is zero below $m_v^{\mathrm{FB}}$, rises across the band, and drops at $m_v^w$ by $(1-q)\rho(m_v^w, e^w)(1-\theta)L$. Hence $\frac{d}{d\ell}\mathcal{W}(m_v(\ell), \ell)$ is not signed by $\mathcal{W}_\ell \geq 0$ and $m_v'(\ell) \geq 0$ separately, and part (iv) answers that. The distorted worker's marginal social product of $m_v$ carries the two correction terms of equation (41) valued at the true $L$ but is evaluated at the private effort and vet set, so it is not ordered against the first-best in general.

(iv) On $\{L \leq \ell_0 \bar{L}_0\}$, with $\bar{L}_0$ the top of the base-stakes support as in Proposition A7(iv), Lemma 1(d) gives $0 \leq \mathcal{W} \leq (1-\theta)\,\Delta D$, where $\Delta D$ is the increase in the expected shipped loss, itself at most $(1-q)\mathbb{E}[L]$, so

$$K_0 \;\equiv\; \sup_{m_v \in [0,1],\, \ell' \leq \ell_0} \mathcal{W}(m_v; \kappa, \ell') \;\leq\; F_D(\kappa)\,(1-\theta)(1-q)\,\ell_0\,\mathbb{E}[L_0] \;<\; \infty,$$

a bound uniform in the matched type. Above, $m_v(\cdot)$ is nondecreasing on $[\ell_*, \infty)$ by hypothesis, so $m_v(\ell) \in [\underline{m_v}, 1]$ for every $\ell \geq \ell_*$, and by equation (42) there is $\ell_1 \geq \max\{\ell_*, \ell_0\}$ with $\inf_{m_v \in [\underline{m_v}, 1]} \mathcal{W}(m_v; \kappa, \ell) > K_0$ for all $\ell \geq \ell_1$. Then $\mathcal{W}(m_v(\ell), \ell) > K_0 \geq \mathcal{W}(m_v(\ell'), \ell')$ for every $\ell \geq \ell_1$ and $\ell' \leq \ell_0$.

The growth hypothesis (42) does not follow from Assumption 1. On the both-vet regime the per-task loss is the intensive gap $S(\mathrm{Vet}, e^{\mathrm{FB}}) - S(\mathrm{Vet}, e^w)$ of the proof of Proposition 4(iii), a detection gap that shrinks as effort rises over an interval of length $(1-\theta)L$ that grows, and the limit of that product is a tail property of the detection technology the assumption does not pin down. In our parameterized example it holds uniformly: by the proof of Proposition 9(iii) the per-task loss grows like $(\lambda/\gamma^2)(1/\theta - 1 + \ln\theta)\ln L$, with a leading coefficient free of $m_v$, since verification skill enters $e^w$ only through a term of order $\ln m_v$, bounded on $[\underline{m_v}, 1]$ because $\underline{m_v} > 0$, so aggregating over the base-stakes support gives $\mathcal{W}(m_v; \kappa, \ell) \geq c \ln \ell - C$ uniformly on any $[\underline{m_v}, 1]$ with $\underline{m_v} > 0$. $\square$

### B.3 The Coupled Attention Budget

The main text prices the engineer's time at a constant shadow value $\lambda$, so the tasks are separable. We show that this is the reduced form of a model in which attention is a shared budget allocated across the whole portfolio, and we record the comparative static the coupled model adds. Let the engineer maximize total value subject to a time budget, $\max \int \mathrm{value}\, d\mu$ subject to $\int \mathrm{time}\, d\mu \leq T$, over an atomless portfolio $\mu$ of tasks $(d, L)$, choosing each task's action and effort. The coupled model needs one action the main text does not, *Skip*, which declines the task at zero time and zero value.

**Nesting.** With the portfolio atomless, the budget problem separates per task into max{value − $\lambda$ time}, the decoupled payoff at shadow price $\lambda$, the multiplier on the budget. Every per-task result of the main text therefore holds at the budget-clearing price $\lambda^*$, and the separable model is the coupled model evaluated at its equilibrium shadow price. The separation is exact because the portfolio is a continuum, by Lyapunov convexification. With finitely many tasks the discrete action choice leaves a duality gap, and in a three-task instance of our parameterized example the gap is strict: no single price supports the budget optimum. Total time is nonincreasing in $\lambda$, so $\lambda^*$ is generically unique. The set of clearing prices is an interval where the budget lands on a stretch on which total time is flat in $\lambda$, and every selection from that interval supports the same allocation, which holds in the parametric family.

**The price of attention.** With difficulty $d \sim U[0,1]$ entering only through the scope gate, total time at the optimum is $\kappa\,\bar{t}_{\text{in}}(\lambda) + (1-\kappa)\bar{t}_{\text{out}}(\lambda)$, averaging the time a reachable and an unreachable task take. Differentiating the budget condition total time$(\lambda^*, \kappa) = T$ gives $d\lambda^*/d\kappa = -(\bar{t}_{\text{in}} - \bar{t}_{\text{out}})/(\partial\,\text{total time}/\partial\lambda)$, and off the flats on which total time is insensitive to the price the denominator is negative, so $d\lambda^*/d\kappa$ shares the sign of $\bar{t}_{\text{in}} - \bar{t}_{\text{out}}$. Out of scope the choice is Solo or Skip, and that choice routes the sign. Where the out-of-scope tasks are Skipped, widening scope converts free tasks into delegated ones that take time, and the price of attention rises. Where out-of-scope time is carried by Solo work and the tasks that switch delegate faster than they ran Solo ($t_A + f + c(e^*) < t_H$ on the switching set), it falls. A worker good enough to ignore a weak AI is unaffected. Neither routing follows from skill alone, since stakes, value, and the equilibrium price all enter the Solo-versus-Skip comparison. At the calibration of Figure 10 the low-skill worker Skips every out-of-scope task and the high-skill worker carries them Solo, so as the AI's scope widens the value of a low-skill worker's attention rises and a high-skill worker's falls, the wage-side shadow of the compression and expansion of Proposition 1. A mixed out-of-scope portfolio mixes the two signs through $\bar{t}_{\text{out}}$.

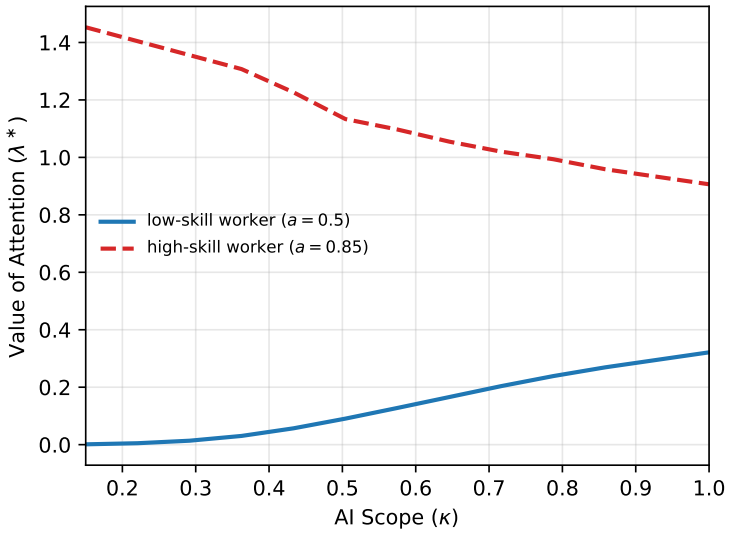



Figure 10: The price of attention

*Note:* The equilibrium shadow price of attention $\lambda^*$ as the AI's scope $\kappa$ widens, for a low-skill ($a = 0.5$) and a high-skill ($a = 0.85$) worker who both use the AI ($q = 0.8$).