Shadow AI and Knowledge Governance in Firms

Marcos Balmaceda* Pablo Paniagua†

Abstract

Organizational theory expects firms to decentralize authority when productive knowledge disperses to frontline workers. Generative AI has dispersed that knowledge to every knowledge worker, yet firms restrict access to AI, and workers conceal the gains they obtain even where access is allowed. We develop a formal model in which both responses are rational equilibria rather than governance mistakes. IT restricts because it bears the cost of any security incident but captures little of the benefit; workers conceal because disclosure invites extraction, a hold-up that strikes at the revelation of tacit knowledge, where neither ownership allocation nor contract design can reach it. What resolves both failures is not ownership form but the structure of observability. We prove an observability-asymmetry theorem: distributing the authority to observe and sanction extraction to a peer community, in the sense of Elinor Ostrom's polycentric governance, sustains a no-extraction commitment under strictly weaker conditions than any bilateral-observability design. Contribution-based screening, peer-enforced commitment, and a community-governed repository implement the design. The model reframes shadow AI as a governance problem with an institutional rather than managerial solution, and identifies the organizational conditions under which that solution holds.

Keywords: knowledge governance; polycentric governance; shadow AI; collective observability

Disclosure: This research did not receive external funding. Neither author has any employment, consulting, or ownership relationship — active within the last three years — with any firm that could be substantially affected by this research, nor any other financial relationship or potential conflict of interest to disclose. This statement applies to both authors and to their spouses and partners.

*KU Leuven, Department of Management, Strategy and Innovation, Naamsestraat 69, 3000 Leuven, Belgium. Email: marcos.balmaceda@kuleuven.be.

†Centre for the Study of Governance and Society, King's College London, Strand, London WC2R 2LS, United Kingdom. Email: pablo.paniagua_prieto@kcl.ac.uk.

1 Introduction

When the knowledge needed to apply a technology spreads from specialists to the workers who use it, organizational theory holds that authority over the technology should move outward to match it (Aghion and Tirole, 1997; Garicano, 2000). On the surface this is obvious and compelling: if frontline workers understand best how a tool applies to their work, control over its use should follow that knowledge. General-purpose AI is a sharp case, dispersing problem-solving capability to nearly every knowledge worker at once. Yet, as we argue, dispersion has not produced decentralization. Firms have tightened control over AI rather than loosened it, and many workers granted access conceal what the tools let them do. Neither response is a managerial failure; each is an equilibrium. The unit that controls access, typically IT, bears the blame for any security incident while capturing little of the productivity upside, so restriction is its dominant strategy. Workers conceal because a disclosed gain is one the firm can renegotiate away, a hold-up that strikes at the moment of revelation rather than at the relationship-specific investment the classic theory studies. The standard model treats the tacitness of AI know-how as a reason to centralize authority; the same tacitness, once incentives diverge, becomes the worker's means to withhold it. We show that a polycentric governance design (Ostrom, 1990),[1] which distributes the authority to observe and sanction rather than merely delegating it, can dissolve both failures at once, and we identify the conditions under which it dominates hierarchy. The stakes are strategic as the design determines whether a firm captures the scalable knowledge asset that AI creates or forfeits it to unmanaged private use.

Research on knowledge governance has long treated hierarchical authority over technology as efficient when specialized knowledge is concentrated at the center of the organization (Garicano, 2000; Grant, 1996; Kogut and Zander, 1992). Across earlier generations of enterprise IT this logic matched reality: expertise in security, compliance, integration, and data infrastructure resided in specialized technical units, and adoption itself depended on their mediation, so hierarchical governance was not only efficient but, in practice, unavoidable (Grandori, 2001; Nickerson and Zenger, 2004). The information-systems literature on shadow IT documents what happened at the margins when that mediation lapsed (Behrens, 2009; Silic and Back, 2014); general-purpose AI turns that margin into the main case.

Large language models (LLMs) fundamentally alter both conditions. The relevant productive knowledge no longer concerns technical implementation alone, but the application of AI tools to highly specific tasks and workflows. Workers acquire this knowledge directly through use, experimentation, and repeated interaction with models in local production contexts. In this sense, the knowledge created through AI adoption is "sticky" in von Hippel's (1994) formulation. It resides with the individuals who use the technology and is costly to transfer to external authorities. Simultaneously, the adoption barrier that historically reinforced centralized control has largely disappeared. Any worker with a browser and a consumer account can privately experiment with AI tools at essentially zero setup cost and without organizational approval. As a result, the classical rationale for hierarchical technology governance weakens on both dimensions. The in-

[1] Polycentric governance, a concept developed by Elinor Ostrom (winner of the 2009 Nobel Memorial Prize in Economic Sciences), describes a system in which many semi-autonomous decision centers hold authority over overlapping domains under a common set of rules, with no single center in exclusive control (Ostrom, 2010). Section 4 makes the distinction from ordinary decentralization precise.

formational advantage once held by specialized technical authorities erodes, while the practical constraints that historically enforced centralized control dissolve.

Yet firms have not decentralized. The reason is two reinforcing failures, each an equilibrium of the institutional structure firms already possess (North, 1990). The first emerges at the level of IT governance. Because IT departments bear the full organizational cost of AI-related security failures while capturing little of the productivity upside associated with experimentation, restricting access becomes the dominant strategy irrespective of aggregate organizational welfare (Niskanen, 1971; Tullock, 1965). This incentive does not derive from ignorance about AI's productive potential. Even when senior management and IT leaders recognize the gains associated with generative AI, accountability structures continue to reward restriction over experimentation. Shadow adoption[2] therefore emerges as workers' rational response to institutional incentives that discourage formal access.

A second failure arises at the level of knowledge disclosure. Workers who adopt AI productively accumulate tacit and locally embedded knowledge that may be highly valuable to the organization. Yet revealing such knowledge exposes them to managerial extraction because formal authority over evaluation, incentives, and workload allocation remains hierarchical (Aghion and Tirole, 1997). A financial analyst who develops an AI-enabled workflow that cuts reporting time in half, for example, cannot assume that disclosure will generate recognition or shared gains. The more predictable outcome may instead be higher performance targets or expanded responsibilities. Under these conditions, concealment becomes rational even when AI access itself is unrestricted. This mechanism extends classic hold-up logic from the investment stage of contractual relations to the revelation stage of knowledge disclosure (Klein et al., 1978; Williamson, 1985). Because the relevant knowledge remains tacit and non-verifiable ex ante, neither ownership allocation nor formal contract design resolves the problem. The result is a reinforcing dynamic in which restriction pushes AI adoption underground, while underground adoption further strengthens workers' incentives to conceal productive knowledge from the firm.

We formalize these two failures and derive their joint resolution. The access restriction problem is modeled as a simultaneous-move game between IT and a representative worker, while the disclosure problem is analyzed as a hold-up game solved through backward induction. We examine an alternative polycentric governance structure in a repeated-game setting where collective observability enables forms of multilateral punishment that hierarchical governance cannot sustain. The model is built from standard primitives, the knowledge-hierarchy cost structure, the hold-up logic of incomplete contracts, and the folk theorem for repeated games; its three mechanisms are predictions of that construction, not generalizations from a case. We then read one observed configuration against them. BBVA, one of Europe's largest banks, with about 125,000 employees across 25 countries, moved from widespread shadow adoption under restrictive governance to a large-scale knowledge commons of thousands of community-developed AI tools, a trajectory in which each mechanism is independently visible (Alfaro et al., 2026). We use it as a theory-informing case in

[2]We define shadow AI operationally to include three phenomena that the governance literature has treated separately: (i) *unauthorized external use*, in which workers use consumer AI tools outside enterprise data agreements entirely without IT awareness; (ii) *informal internal use*, in which organizational awareness exists but no formal governance structure has been established; and (iii) *unreported productivity gains*, in which workers hold formal access but conceal the magnitude of AI-enabled output improvements. The first phenomenon generates the security exposure σ; the third generates the organizational knowledge suppression that the hold-up model captures. The second is the intermediate state the access-category analysis in Definition 3 characterizes.

the sense of Siggelkow's (2007): it illustrates and disciplines the model's logic without testing or generating it. Appendix B summarizes the case, and we refer to it as the relevant mechanisms appear.

The analysis generates three principal results. First, we show that a contribution requirement can produce a separating equilibrium in which workers whose AI use generates sufficiently high productivity gains voluntarily self-identify through willingness to incur the costs of participation, even in the absence of direct monitoring or mandatory disclosure. Second, we show that a Community of Practice (Wenger, 1998) can render non-extraction commitments self-enforcing through collective observability. Under hierarchical governance, extraction is typically a bilateral and partially private event, limiting punishment to the withdrawal of a single worker. Under polycentric governance, by contrast, attempts at extraction become simultaneously observable to all contributors, allowing collective withdrawal to function as a credible punishment mechanism. Third, we establish an observability-asymmetry theorem (Theorem 1): the set of organizational environments in which non-extraction commitments are self-enforcing is strictly larger under polycentric governance than under any design that relies on bilateral observability of extraction events. The revelation-stage hold-up creates the need for collective observability; Theorem 1 is the result that makes it sufficient. Together, these three mechanisms generate organizational payoffs that exceed the hierarchical optimum whenever the gains from knowledge sharing, reduced security exposure, and community externalities outweigh the costs of governance.

The result is conditional. Polycentric governance dominates only when productive AI knowledge is dispersed and tacit rather than concentrated in a specialist unit, when the adoption barrier is low enough that shadow adoption is viable, and when monitoring costs scale with firm size; where AI serves a single concentrated task, gains are homogeneous, or regulated outputs require centralized accountability, Garicano's (2000) hierarchical logic remains efficient. As a work of theory, the paper draws the line between its contribution and the empirical agenda deliberately. It states the model's predictions in operational, falsifiable form and specifies, in Section 6, the cross-sectional variation in governance and accountability that would identify them, handing the field a precise test rather than awaiting one.

The paper contributes to three streams of research. First, knowledge-hierarchy theory characterizes how a firm should optimally organize knowledge work, and recent work extends it to AI (Garicano, 2000; Ide and Talamàs, 2025, 2026). These are efficiency benchmarks, in which the unit that designs the organization maximizes output and the structure that emerges is the structure that should emerge. We show that the decentralizing adjustment they imply for AI does not occur inside real firms, for two reasons they abstract from: the unit that controls access pursues its own accountability rather than firm output, and workers conceal the tacit knowledge AI generates because disclosure invites extraction. Because generative AI applies across functions simultaneously, the resulting suppression is organizational rather than localized, an exploitation trap in March's (1991) sense in which governance routines persist even as worker-level capabilities accumulate outside the firm's visible knowledge base.

Second, the paper extends hold-up theory beyond relationship-specific investment at the contracting stage (Hart and Moore, 1990; Klein et al., 1978; Williamson, 1985). Tacit knowledge generates an analogous extraction hazard at the revelation stage, the hazard the ratchet-effect and expropriable-ideas literatures study under bilateral contracting (Anton and Yao, 1994; Laffont and Tirole, 1988; Weitzman, 1980).

Our contribution is the substitution result: when disclosed knowledge has value super-additive across the workforce, the contractual commitment those literatures rely on is bounded, and a collective-observability architecture takes its place (Theorem 1). Concealment here is structural rather than relational (Černe et al., 2014; Connelly et al., 2012), so the remedy is organizational; the full differentiation is in Section 3.

Third, the paper extends polycentric governance theory (Lee and Edmondson, 2017; Ostrom, 1990, 2010; Puranam et al., 2014). Prior work establishes that distributed structures dominate hierarchy when monitoring costs are high and local knowledge is dispersed, and treats this dominance as a general property of distributed organizing. We identify what is specific to AI. The advantage does not come from monitoring cost alone but from the revelation-stage hold-up, which bilateral commitment cannot resolve because no single worker's withdrawal threatens the firm enough to deter extraction. The structure of observability and sanction does the work, not ownership form: the result holds for any arrangement that distributes them, so it travels beyond AI to settings such as whistleblowing channels and worker councils, and narrows the polycentric-versus-decentralized distinction to a precise margin. A signaling analysis then identifies the minimum credible investment required to initiate the transition.

2 Knowledge Governance and the Hierarchy Premise

The knowledge hierarchy and its efficiency conditions

We model the firm's knowledge production problem following Garicano (2000). Workers face a stream of problems drawn from the exponential distribution $f(z) = \lambda e^{-\lambda z}$, where $z \geq 0$ indexes complexity and $\lambda > 0$ governs the tail of the distribution. Workers are organized into a production layer and a sequence of specialist layers. Production workers handle the problems they know and escalate the remainder to specialists, who process only the escalated tasks.

Two cost parameters govern how knowledge is organized. The communication cost $h \in (0,1)$ is the share of a specialist's time consumed per consultation. The knowledge acquisition cost $c > 0$ is the cost of learning to solve a class of problems. Because communication is costly, knowledge is allocated in frequency order. Production workers learn the common problems, and higher layers learn the progressively rarer ones.

A social planner chooses thresholds (z_w, z_s) to maximize per-capita net output subject to the specialists' capacity constraint. Solving this problem under the exponential distribution yields the efficient thresholds

$$z_w^* = \frac{1}{\lambda} \ln\left(\frac{h\lambda}{c} - h\ln h\right), \tag{1}$$

$$z_s^* = \frac{1}{\lambda} \ln\left(\frac{\lambda}{c} - \ln h\right). \tag{2}$$

The closed forms matter less than what they locate. The threshold z_w^* is the complexity up to which a production worker solves problems unaided, and problems above it are escalated. Because $\ln h < 0$, we have $z_w^* < z_s^*$, so specialists optimally cover the broader and rarer knowledge bands.

One comparative static drives the rest of the paper. Both thresholds rise as the acquisition cost c falls ($\partial z_w^*/\partial c < 0$, $\partial z_s^*/\partial c < 0$), because a lower cost lets each layer master a wider band of problems without

escalation (Garicano, 2000). The effect on organizational shape is asymmetric: as $c \to 0$, z_w^* grows without bound, production workers come to handle almost the entire distribution, and the escalated share falls toward zero. Span of control rises, layers shrink, and the specialist tier contracts until it is redundant, flattening the hierarchy into production-level problem solving.

This is the efficient benchmark, and it assumes a social planner who sets the thresholds to maximize output. Real firms have no such planner. The unit that actually controls access to a new technology, the IT department, sets the effective threshold and pursues its own objective rather than firm output. The wedge between the efficient threshold and the one IT implements is the governance gap we formalize below (Definition 1).

[Insert Figure 1 about here]

General-purpose AI as a knowledge cost shock

LLMs are a general-purpose technology in Bresnahan and Trajtenberg's (1995) sense: broadly applicable, improving over time, and complementary with organizational innovation. What distinguishes AI is where its productivity effect lands. Earlier automation targeted codifiable, routine tasks at the production margin, leaving the knowledge-intensive work hierarchies coordinate untouched; LLMs acquire tacit knowledge and perform noncodifiable cognitive tasks (Ide and Talamàs, 2025), so their effect arrives precisely where knowledge hierarchies concentrate authority. That is what makes AI consequential for governance.

AI enters the framework of the preceding section as a uniform downward shift in c across tasks and workers, over the full problem distribution rather than a narrow domain.[3] As acquisition costs fall, the efficient threshold for production workers rises, the specialist layer's comparative advantage shrinks, and the information concentration that justified centralized authority erodes (equations (1)–(2)). The model predicts that firms facing this shock should loosen hierarchical access controls and let governance authority follow capability outward. Earlier information technologies produced weaker versions of this effect in bounded domains. Bloom et al. (2014) decompose ICT into an information component that lowers acquisition costs and decentralizes and a communication component that lowers the helping cost and centralizes, and find decision rights tracking both margins; Rajan and Wulf (2006) and Gulati et al. (2026) document the associated flattening. In that decomposition a general-purpose LLM operates on the information margin: the worker consults the model rather than escalating to a specialist, so the shock falls on c, and the predicted adjustment is sharper than for earlier technologies because the cost reduction is broader and deeper. The decomposition also sharpens the puzzle. Cheaper communication would predict efficient centralization, with knowledge flowing to the center; what firms exhibit instead is restriction with no knowledge flow in either direction, which neither margin predicts.

The evidence contradicts this prediction at every level. At the worker level, Brynjolfsson et al. (2025) find that customer-service agents given a generative AI assistant resolved 15 percent more issues per hour,

[3] We model general-purpose AI as a reduction in the knowledge acquisition cost c, the channel through which a tool lets frontline workers solve more problems without escalation. Ide and Talamàs (2025) study a complementary channel in which AI is an autonomous solver in its own right; the cost-reduction channel is the one relevant to how firms allocate authority over a tool that disperses capability to incumbent workers, and the two are not mutually exclusive.

with gains largest among less-experienced workers and negligible among the most experienced, exactly what a downward shift in c implies: workers below the knowledge frontier gain most. The cost shock is real; the governance response it should trigger has not materialized. Survey evidence shows why. A McKinsey survey of about 500 organizations finds nearly two-thirds cite security and risk as the top barrier to scaling AI (Asaftei et al., 2026), and Renieris et al. (2023), across 1,240 organizations, find that 78 percent use third-party AI tools, more than half of AI failures originate there, and leadership is often unaware which tools are in use, a pattern they term shadow AI. Firms whose workers have acquired meaningful AI capability are simultaneously failing to capture it and bearing the security exposure of unmanaged adoption.

Shadow AI belongs to a documented class. The information-systems literature identifies shadow IT, employee-initiated adoption outside IT sanction, across enterprise software and unsanctioned cloud services (Behrens, 2009; Silic and Back, 2014), tracing it to friction between worker needs and IT provisioning. Shadow AI differs in two respects that make the governance problem structural. First, the adoption barrier is effectively zero: a browser and a consumer account suffice, putting unmanaged adoption below any monitoring threshold, so restriction cannot work. Second, and more consequentially, the productive knowledge workers develop is task-specific and tacit, genuinely valuable to the firm yet difficult to extract without their cooperation. Zero adoption cost makes restriction ineffective; tacit productive knowledge makes disclosure strategically hazardous.

The governance gap: authority and incentives

The preceding analysis treats AI as a uniform cost shock that a social planner would accommodate by expanding worker discretion. That conclusion holds only if the unit controlling AI access optimizes firm value, and in practice it does not. The authority the knowledge hierarchy model assigns to a planner resides, in firms, with IT, a bureau whose incentives reward restriction rather than output. The divergence between IT's objective and firm value generates the governance gap the model now derives.

Let $b > 0$ denote IT's private benefit from restriction and $\sigma > 0$ the expected security cost under permissive access. Since $b > 0$, IT's dominant strategy is to block access to AI tools, effectively fixing the knowledge acquisition cost at the pre-AI baseline $c_0 > c$. Workers can still solve problems at the baseline threshold $z_w^*(c_0)$; what IT forecloses is the AI-enabled increment that would raise the threshold to $z_w^*(c)$.

Definition 1 (Governance gap)**.** *Let $c_0 > c$ denote the pre-AI knowledge acquisition cost, $z_w^*(c)$ the efficient threshold under AI, and $z_w^*(c_0)$ the threshold IT implements by restricting access. The governance gap is* $G(c) \equiv z_w^*(c) - z_w^*(c_0) \geq 0$.

Since $\partial z_w^*/\partial c < 0$ and c_0 is fixed, $G(c)$ is strictly increasing as c falls. The more capable the AI, the larger the wedge between efficient and implemented knowledge allocation. As $c \to 0$, $z_w^*(c) \to \infty$ while $z_w^*(c_0)$ remains finite, so $G(c) \to \infty$.

We interpret b as an *audit-defensibility premium*. When IT formally prohibits AI use, a breach attributable to unauthorized employee use becomes the liability of the policy-violating worker rather than of IT's governance decision, insulating IT from accountability and transferring breach risk downward, independent of whether restriction is efficient for the firm. This differs from Niskanen's (1971) budget-

maximization: what distinguishes the AI case is that breach liability is large, personalizable to IT leadership, and deflectable through a restriction policy, so even a budget-maximizing unit restricts when added infrastructure expands the breach surface by more than the insulation it provides. The parameter b aggregates audit defensibility, career protection, and regulatory blame avoidance.

Definition 2 (Audit-defensibility premium)**.** *Let $p(c) > 0$ denote the per-period probability of a breach event when AI is in use, weakly increasing in AI capability ($\partial p/\partial c \leq 0$: as knowledge cost c falls, AI processes more sensitive data and expands the attack surface, raising breach likelihood). Let $L > 0$ denote the personal liability exposure of IT leadership per breach event (regulatory fine, disciplinary action, or reputational cost). The* audit-defensibility premium *is:*

$$b \equiv p(c) \cdot L \cdot (1 - \alpha).$$

Under Permit, IT's expected per-period accountability cost is $p(c)L$. Under Restrict, the formal prohibition shifts breach attribution to policy-violating workers, leaving IT with expected cost $\alpha p(c)L$. The premium $b = p(c)L(1-\alpha)$ is the per-period accountability saving from restriction, strictly positive whenever $\alpha < 1$, $p(c) > 0$, and $L > 0$.

Remark 1 (Endogenous payoff parameters)**.** *The knowledge surplus V_F and the security exposure σ are not fixed primitives; both grow as the governance gap widens, because a larger gap means more foregone knowledge to recover and a larger unmanaged shadow band to expose. Writing $V_F(G) = \gamma G(c)$ and $\sigma(G) = \beta G(c)$ with $\gamma, \beta > 0$, the dominance margin $\Delta\Pi = V_F(G) + N \cdot e + \sigma(G) - N \cdot c_c$ therefore rises with AI capability, so the capability threshold above which polycentric governance dominates is determinate and finite.*

[Insert Figure 2 about here]

The implication is structural: the same fall in c that widens the wedge between efficient and implemented allocation also raises b, so IT's incentive to restrict hardens exactly as the welfare cost of restriction grows. The two channels are independent. Restriction fixes the implemented threshold below the efficient level regardless of worker behavior, while hold-up blocks the organizational surplus even when access is permitted, so resolving either alone leaves value unrealized. The next section derives both.

3 Two Channels of Governance Failure

We analyze a governance comparison game $\mathcal{G}(R)$, parameterized by regime $R \in \{H, P\}$, with three classes of players: the IT department, N workers indexed $i = 1, \ldots, N$, and firm management. Under polycentric governance, $M \leq N$ workers hold contributor licenses. We write N throughout the main text for clarity; terms that scale with the contributor pool ($N \cdot e$, $N \cdot c_c$) take M when $M < N$, and since the dominance condition is monotone in M, no qualitative conclusion changes. Let $\omega > 0$ denote per-worker productive output that accrues to the firm through normal production channels under any governance regime. The governance comparison is cross-sectional. For each regime, we characterize equilibrium outcomes. Transition

dynamics between regimes are analyzed in Section 6. Proofs not given in the text are collected in the online Appendix A, which also reports the numerical calibration; Appendix B summarizes the BBVA case drawn on throughout.

Under hierarchical governance $\mathscr{G}(H)$, the interaction proceeds through three sequential stages. In the first stage, the IT department sets access policy (choosing between Restrict and Permit) and each worker decides whether to comply or adopt AI through shadow channels. AI-enabled workers then choose between revealing and concealing the tacit knowledge their AI use has generated. In the final stage, the firm observes any disclosed knowledge and chooses between commitment and extraction, given extraction gain $\tau > 0$. Propositions 1 and 3 characterize the equilibria of this game; the terminal organizational payoff is $\Pi_H = N\omega - \sigma$.

Under polycentric governance $\mathscr{G}(P)$, the structure changes in two essential ways. Before worker play begins, the firm designs the governance architecture. It extends basic managed AI access to all employees and applies scarcity only to the contributor tier, setting license capacity $M < N$, a contribution requirement $c_c > 0$, an amnesty commitment A, and Community of Practice rules governing contribution and tacit knowledge sharing (Wenger, 1998). Workers self-select into the contributor tier. Worker i participates when $(\theta_i - 1)\pi \geq c_c$, with managed non-contributor access at π the relevant outside option. Disclosure and response then repeat across an infinite horizon with discount factor $\delta \in (0,1)$; extraction is publicly observable to the community, triggering peer sanction $s > 0$ and immediate commons collapse. Propositions 4 and 6 characterize the equilibria of this repeated game; the terminal payoff is $\Pi_P = N\omega + V_F + N \cdot e - N \cdot c_c$, where $e > 0$ is the per-contributor Community of Practice externality. Proposition 7 evaluates $\Delta\Pi = \Pi_P - \Pi_H$ across both regimes and derives the dominance condition (4).

Definition 3 (Three access categories)**.** *Three access states recur across regimes. Under* shadow adoption *the worker uses consumer AI without authorization, capturing π at concealment cost k while the firm bears security exposure $\sigma > 0$; this is the equilibrium of $\mathscr{G}(H)$ (Proposition 1). Under* managed access *the worker holds enterprise credentials, which removes σ but leaves $V_F = 0$, since the hold-up of Proposition 3 leaves disclosure incentives unchanged ($\Pi_I = N\omega$). Under* contributor access *the worker holds a license conditioned on contribution under $\mathscr{G}(P)$, which generates $V_F > 0$ through the repository and resolves the hold-up by collective observability (Proposition 6). Thus σ arises only from shadow adoption and V_F only from contributor access. Polycentric governance moves every worker out of shadow adoption, non-contributors into managed access and productive adopters into contributor access, eliminating σ and releasing V_F at once.*

Channel one: the IT–worker game

The interaction is a simultaneous-move game between the IT department and a representative worker. IT chooses between *Permit* and *Restrict*; the worker chooses between *Comply* and *Shadow*. Four parameters shape the payoffs: the worker's productivity gain $\pi > 0$ from AI use, IT's private accountability benefit $b > 0$ from restriction, the firm-level security cost $\sigma > 0$ from unmanaged shadow adoption (data processed outside enterprise agreements, AI services not subject to organizational compliance controls), and the worker's concealment cost $k > 0$. Table 1 presents the full payoff matrix.

[Insert Table 1 about here]

The structure of IT's payoffs makes Restrict a dominant strategy. For any worker action, restriction yields a strictly higher IT payoff than permission, because private benefit b is collected regardless of whether the worker complies, while the blame-deflection logic eliminates IT's residual accountability even under shadow adoption. Given restriction, a productive worker prefers Shadow over Comply whenever the productivity gain exceeds the concealment cost.

Proposition 1 (Hierarchy Failure)**.** *When IT's audit-defensibility premium is positive ($b > 0$, which holds for all $\alpha < 1$), restriction is a dominant strategy independent of worker behavior. The unique equilibrium is* (Restrict, Shadow)*: IT restricts access; workers adopt AI through informal channels rather than comply. Workers privately capture productivity gains; the firm captures no scalable organizational knowledge value and bears the full security cost of unmanaged shadow adoption.*

The efficient outcome (*Permit, Comply*) yields higher worker surplus and eliminates concealment costs, but is not implementable because IT strictly prefers Restrict due to b. The result corresponds to a Niskanen (1971)-type incentive problem within the firm, where the governance unit optimizes accountability payoffs rather than organizational output.

The accountability mechanism is not the only candidate explanation for restriction; inertia, regulatory pressure, isomorphic conformity, and simple unawareness are others. What distinguishes b is its observable implications. It predicts that restriction persists even when senior management is fully informed of the gains, that it is stronger where personal liability for breaches is higher (financial services, healthcare, legal), and that the governance gap widens as AI capability grows. Inertia and awareness-deficit accounts instead predict that restriction fades as firms learn, and regulatory accounts predict it tracks regulatory stringency rather than capability. The persistence-and-intensification pattern is consistent with Alfaro et al. (2026) and Asaftei et al. (2026) and is distinctive to the accountability mechanism.

The equilibrium is robust to the degree of blame deflection. Restrict dominates for all $\alpha \in [0,1)$, since any deflection makes $b = p(c)L(1-\alpha) > 0$. Only at $\alpha = 1$, where restriction provides no blame deflection and IT bears full accountability regardless of policy, does $b = 0$ and the equilibrium require an independent private benefit, and such complete insulation failure is not observed in practice.

Proposition 2 (Managerial override is insufficient)**.** *A senior management mandate to permit AI access does not eliminate IT's restriction incentive. Because $b = p(c) \cdot L \cdot (1-\alpha) > 0$ derives from IT's* institutional position *as the unit to which AI-related breach accountability is attributed, mandating permission preserves the accountability structure even as it removes the restriction policy. Three interventions can reduce b: reassigning governance authority to a body whose accountability structure does not generate the audit-defensibility premium (eliminating b at source by changing the institutional position); indemnifying IT against personal liability (reducing L); or establishing fuller attribution symmetry (raising α). Governance reassignment is the most credible because it operates on the institutional position itself, rather than negotiating around payoff parameters within it.*

This follows from Definition 2: b depends on (p, L, α), not on access policy, so a mandate changes the equilibrium action but not b, and Proposition 1 re-applies once the mandate lapses. Of the three remedies,

governance reassignment is the most credible: indemnification needs complete contracts over breaches before standards stabilize, and attribution symmetry needs tracing infrastructure at odds with differentiated responsibility, while reassignment changes the structure of accountability rather than a parameter within it.

Corollary 1 (Comparative statics of b)**.** *From Definition 2, $b = p(c) \cdot L \cdot (1-\alpha)$, so:*

$$\frac{\partial b}{\partial L} = p(c)(1-\alpha) > 0, \qquad \frac{\partial b}{\partial c} = \frac{\partial p}{\partial c} \cdot L(1-\alpha) \leq 0, \qquad \frac{\partial b}{\partial \alpha} = -p(c)L < 0.$$

(i) b is strictly increasing in personal liability L: legal regimes that attach breach liability to named IT officers rather than the organization as a whole produce higher b and stronger restriction regardless of AI's productive potential. (ii) b is weakly increasing in AI capability (weakly decreasing in c): falling c raises breach probability $p(c)$, expanding the breach surface that a restriction policy insulates against. (iii) b is decreasing in the residual blame fraction α: more precise attribution of unauthorized AI access lowers α (attribution more completely shifts to workers under restriction), raising b.
Empirical implication. Restriction intensity is cross-sectionally higher in industries with personal liability regimes for data governance (financial services, healthcare, legal) and increases over time as AI capability grows, even under constant regulatory frameworks. The governance gap $G(c)$ widens monotonically as c falls under any $b > 0$: the more capable the AI, the more pressing the governance redesign.

Channel two: the worker–firm hold-up

Suppose IT permits AI use. A second failure then arises at disclosure: workers accumulate tacit, task-specific knowledge through AI use that is valuable to the firm but costly to reveal, because revealing it invites renegotiation.

The interaction is a game between worker and firm. The worker chooses between *Reveal* and *Hide*; the firm responds with *Commit* or *Extract*. Key parameters are the organizational surplus $V_F > 0$ the firm captures by scaling workers' disclosed AI-enabled workflow improvements through knowledge diffusion, the extraction gain $\tau > 0$ from renegotiating terms once productivity is revealed, the worker's disclosure cost $\kappa > 0$, and the concealment cost $k > 0$. Table 2 presents the payoff matrix.

[Insert Table 2 about here]

In the one-shot game, the firm's dominant strategy is to extract whenever the worker reveals, since positive extraction gain $\tau > 0$ makes the commitment not credible. Anticipating extraction, the worker conceals.

Proposition 3 (Hold-Up Equilibrium)**.** *Knowledge disclosure fails. Anticipating managerial extraction, workers conceal AI-enabled productivity rather than reveal it. The firm captures no organizational knowledge surplus. Workers bear the private cost of sustained concealment.*

The cooperative outcome (*Reveal, Commit*) yields $V_F > 0$ and eliminates concealment costs, but is not self-enforcing in a one-shot interaction. This differs from the canonical hold-up problem (Klein et al., 1978; Williamson, 1985) in that the distortion operates at the revelation stage rather than through ex ante

underinvestment in specific assets, so governance must target disclosure incentives rather than ownership allocation (Hart and Moore, 1990) or contract design (Magelssen et al., 2022).

The closer antecedent is the ratchet effect of dynamic incentive theory, in which a worker conceals high productivity because revealing it tightens tomorrow's targets (Freixas et al., 1985; Laffont and Tirole, 1988; Weitzman, 1980); Gibbons (1987) models the piece-rate version of the financial-analyst example above, and Anton and Yao (1994, 2002) study the adjacent hazard of disclosing a valuable but non-contractible idea that the counterparty can then expropriate. Our revelation-stage hold-up shares this logic but differs in the object and the remedy. There the relationship is bilateral and the distortion is under-revelation, which contract-theoretic commitment devices address; here the disclosed object is tacit knowledge whose value is super-additive across the workforce (Assumption 1), so bilateral commitment is bounded by the single-relationship value v_i and cannot reach the commons value. A collective-observability architecture can (Theorem 1). The theorem thus offers that literature the conditions under which governance design substitutes for the contractual commitment it has shown to be fragile.

This mechanism is distinct from the knowledge-hiding literature in organizational behavior. Connelly et al. (2012) identify motivational hiding driven by distrust and competition, Černe et al. (2014) show it is self-reinforcing, and Edmondson (1999) establishes psychological safety as the precondition for team learning, implying concealment in low-trust settings even absent any extraction threat. Here the driver is structural: workers conceal because extraction is the firm's dominant response to revelation. Trust-building and psychological-safety interventions therefore leave the hold-up intact while extraction remains dominant, whereas polycentric governance alters the game, making extraction unprofitable rather than merely unlikely. Foss et al. (2010) pose governance for knowledge sharing as an open problem precisely because motivational fixes fail when the strategic structure is unchanged; the design here addresses that gap.

Interaction between the two failures

The two failures reinforce each other. A worker who surfaces AI productivity after unofficial adoption risks penalty for the prior unauthorized use, raising the disclosure cost from κ to $\kappa + d$ ($d > 0$, a disciplinary risk premium). Disclosure also carries an authenticity risk $a > 0$, as colleagues may question whether AI-assisted output reflects genuine competence (Lifshitz-Assaf, 2018; Raisch and Krakowski, 2021).[4] In the other direction, prolonged concealment generates an inauthenticity cost $k_a > 0$ rising with the duration of shadow use, so both options worsen as shadow adoption persists. Partial interventions fail: permitting access without resolving the hold-up yields concealed productivity, and resolving the hold-up without amnesty leaves workers unwilling to surface prior use. Effective governance must do all three at once, make extraction unprofitable, remove the disciplinary risk, and lower the concealment cost, which the next section shows polycentric design achieves.

[4]For tractability, the authenticity risk a and the disciplinary risk d are folded into the effective disclosure cost κ in the baseline hold-up game (Table 2), which captures the total private cost a worker faces when surfacing AI-enabled productivity. The resolution analysis in Proposition 6 unfolds these two components, because the amnesty mechanism operates directly on d and a. The baseline results are unchanged whether the components are modeled separately or folded in.

4 The Polycentric Resolution

The two failures share a structural feature. Both arise from concentrating formal authority in a single center whose incentives are misaligned with organizational welfare. Ostrom (1990) identifies this as the characteristic failure of monocentric governance, where a single authority must simultaneously acquire dispersed information, bear monitoring costs, and align private incentives with collective welfare. When local knowledge is the binding constraint, no single center can hold all the relevant information.

Definition 4 (Polycentric governance)**.** *Polycentric governance is a structure in which multiple overlapping decision-making units hold authority over defined domains, operate under shared overarching rules, and interact without any single center holding exclusive control.*

This differs from decentralization, which merely delegates authority downward through the hierarchy. Polycentric governance distributes authority by design. The governance response to the two failures operates through three institutional moves, each corresponding to a distinct mechanism in the theoretical framework.

Move one: artificial scarcity as a screening mechanism

The first mechanism is artificial scarcity. The model calls for a contribution requirement that screens productive adopters without monitoring or self-report; BBVA instantiates it by allocating fewer licenses than potential users and conditioning renewal on active contribution, rather than offering licenses to all employees or restricting access to an approved group. Workers who failed to use the license or contribute to the shared tool repository forfeited access; the license was reallocated to the next applicant.

This design resolves a fundamental information problem. The firm faces three obstacles at once: it cannot observe which workers use AI productively, asking them to self-report reinstates the hold-up of Section 3, and monitoring use is prohibitively costly at scale. Artificial scarcity sidesteps all three, without monitoring and without disclosure.

Workers differ in how productively they use AI (Caplin et al., 2025; Tambe, 2026). A worker of type $\theta \in \{\theta_H, \theta_L\}$, with $\theta_H > \theta_L > 0$, generates a productivity gain $\theta\pi$ from formal AI access; holding a license requires a contribution cost $c_c > 0$ per period. The relevant outside option is not abstention but managed access: the firm extends basic AI access to all employees, so a worker who declines the contributor tier still uses AI openly at net payoff π, with nothing to conceal. A worker of type θ therefore prefers contributor access when $\theta\pi - c_c \geq \pi$, that is, when the productivity premium $(\theta - 1)\pi$ from contribution exceeds the contribution cost c_c.

Proposition 4 (Screening via Artificial Scarcity)**.** *A contribution threshold set between the effective participation costs of low- and high-productivity AI users induces a separating equilibrium. Workers with genuinely productive AI applications self-select into the formal governance system, while marginal users remain at the managed-access tier. The firm identifies its most productive AI contributors without monitoring or self-disclosure of productivity.*

The license here governs the contribution tier (who builds tools, documents workflows, and joins the Community of Practice), not basic AI access, which a firm can extend universally while applying scarcity to

the contributor pool. The mechanism works because task-specific know-how is sticky (von Hippel, 1994): a threshold that screens on willingness to bear the cost c_c separates types without asking workers to reveal how productive they are, and without the monitoring that is infeasible at scale (Section 6).

At BBVA, 83 percent of license holders used the platform weekly, averaging 50 prompts per week (Alfaro et al., 2026), consistent with self-selection by high-productivity types. The Quality Score, an automated evaluation of tool quality along dimensions such as guardrails, clarity, and context, served as the contribution filter, replacing a formal risk-admission process that would have been prohibitively slow at scale.

Proposition 5 (Separating interval bounds)**.** *Let $\bar{c}_L \equiv (\theta_L - 1)\pi$ and $\bar{c}_H \equiv (\theta_H - 1)\pi$. The separating equilibrium of Proposition 4 exists if and only if $c_c \in (\bar{c}_L, \bar{c}_H)$.*

Proof. See Appendix A.

The design implication is asymmetric. Setting c_c below $\bar{c}_L$ induces pooling, which lowers screening precision but preserves the governance advantage; setting it above $\bar{c}_H$ drives participation, and hence V_F, to zero. The community need not know the type distribution to land inside the interval: an adaptive rule that raises c_c when participation exceeds the high-type share and lowers it otherwise converges to $[\bar{c}_L, \bar{c}_H]$ while observing only participation rates (Lemma 1, Appendix A). This is consistent with Ostrom's collective-choice principle and with the community-administered standard Alfaro et al. (2026) describe at BBVA, where the contribution requirement was set through the Adoption Network rather than by central IT. Because falling participation signals over-restriction, the rule self-corrects downward but cannot recover an upward deviation once contributors have exited.

Move two: the no-punishment commitment

Proposition 3 establishes that workers will not disclose tacit productivity knowledge if they anticipate extraction. Resolving this requires a credible commitment by the firm not to extract. Credibility is the binding constraint. A bare announcement of amnesty is not credible because the firm has a dominant strategy to defect in the one-shot game.

The repeated game changes the calculus. In an infinitely repeated interaction with discount factor $\delta \in (0,1)$, a firm can sustain commitment to non-extraction as a subgame-perfect equilibrium if the long-run value of the knowledge commons exceeds the one-shot gain from extraction. The firm's continuation value from the cooperative relationship is $V_F/(1-\delta)$; its one-shot gain from extraction is τ. Cooperation is self-enforcing when:

$$\delta \cdot \frac{V_F}{1-\delta} \geq \tau \qquad \Longleftrightarrow \qquad \delta \geq \frac{\tau}{V_F + \tau}. \tag{3}$$

The organizational design problem is to raise V_F and reduce τ simultaneously. The Community of Practice raises V_F through network externalities, as each additional contributor increases the value of the commons to all members. Peer observability serves two distinct roles. First, it makes the continuation punishment credible. Defection triggers immediate commons collapse, eliminating all future V_F and ensuring the punishment is executed rather than merely threatened. Second, the Wizard community[5] imposes a reputational sanction

[5]The Wizard role is defined within the Community of Practice in Wenger's (1998) sense. Wizards are the technically skilled

$s > 0$ in the period of defection itself. Public documentation of the extraction event within the Community of Practice reduces the firm's standing as a trustworthy governance counterpart, making it harder to attract future contributors and sustain leadership support for the governance program, and thereby lowering the net extraction gain from τ to $\tau - s$. Writing $\delta^* = \tau/(V_F + \tau)$ for the threshold in (3) without peer sanctions, the self-enforcement condition becomes $\delta \cdot V_F/(1-\delta) \geq \tau - s$, that is $\delta \geq \delta^{**} = (\tau - s)/(V_F + \tau - s) < \delta^*$, satisfied for a strictly larger set of discount factors than in the absence of peer sanctions; for the general theory of relational contracts, see Baker et al. (2002).

Proposition 6 (Hold-Up Resolution)**.** *When the organizational value of the knowledge commons sufficiently exceeds the one-shot extraction gain, the no-extraction commitment becomes self-enforcing as a subgame-perfect equilibrium. Three mechanisms operate jointly. The Community of Practice raises the organizational cost of defection through collective commons collapse, peer observability makes that punishment credible rather than merely threatened, and immediate reputational sanctions reduce the net extraction gain. Workers disclose AI-enabled productivity; the firm captures the full organizational knowledge surplus V_F.*

The repeated-game analysis selects the grim-trigger equilibrium, under which cooperation prevails until the first deviation and the commons collapses permanently thereafter. This is standard in the relational contracting literature (Baker et al., 2002) and selects the Pareto-superior cooperative outcome among the multiple equilibria the folk theorem supports.

Optimal relational contracts with partial (stick-and-carrot) punishment can sustain cooperation at lower discount factors than grim-trigger strategies (Levin, 2003), but this strengthens rather than weakens the argument. Any punishment available in a bilateral relationship is bounded by v_i, the per-period value of that single relationship, whatever the punishment strategy. Switching to an optimal relational contract lowers the hierarchical threshold δ_i^H but does not close the gap $\delta_i^H - \delta^{**} > 0$, because under polycentric governance the binding constraint is the commons collapse, not the punishment strategy (Theorem 1).

The grim-trigger equilibrium is also robust to renegotiation. Commons collapse is automatic rather than coordinated: a contributor who keeps disclosing after extraction earns $\pi - \tau - \kappa$, while reverting to concealment earns $\pi - k$, and since $\tau > k - \kappa$, withdrawal strictly dominates for every contributor regardless of what others do (Step 3 in the proof of Proposition 6). Repository quality therefore deteriorates without any collective decision. The peer sanction $s > 0$ shrinks the firm's one-time deviation gain to $\tau - s$ against a forfeited continuation value of $V_F/(1-\delta)$, even if workers would renegotiate, since reputational damage persists, and each member can enforce the collapse unilaterally, so the coordination problem that renegotiation-proofness usually raises does not arise.

Why does the same logic not discipline extraction under hierarchy? Observability. Under hierarchy extraction is private: HR adjusts one worker's quota, others cannot verify it, and punishment reduces to that worker withdrawing, which the firm absorbs while extracting from others in turn. Because Wizards work in the open and share a repository, any extraction is immediately visible and collapses V_F in one period, so a hierarchical amnesty with identical words is empty without that observability. Section 6 (Theorem 1) makes the asymmetry precise.

workers who build and maintain shared AI tools, train others, and collectively govern the repository. The term is used consistently with Alfaro et al.'s (2026) description of BBVA's governance structure.

Alfaro et al. (2026) document both features: shadow use reframed as productive demand, removing the disciplinary dimension, and a public augmentation pledge reducing authenticity risk, with credibility deriving from the Adoption Network investment itself, which replacement-based restructuring would have destroyed. In Puranam et al.'s (2012) terms, polycentric governance restructures epistemic interdependence, not merely payoffs: contribution thresholds make types observable through actions, and the community creates predictive knowledge that dissolves the mutual uncertainty sustaining the hold-up.

Move three: the Adoption Network structure

The governance structure constitutes a Community of Practice in Wenger's (1998) sense, defined by mutual engagement in a shared domain, joint enterprise, and a shared repertoire of tools and procedures. Three overlapping roles organize participation. Champions are departmental coordinators who translate organizational AI strategy into local practice. Co-Champions assist at the team level. Wizards are technically skilled workers who build and maintain custom AI tools and train others. No single center holds authority over which tools are built, how they are governed, or how knowledge is shared.

This maps directly onto polycentric governance. Senior leadership manages the membership boundary, setting the license cap, issuing amnesty, and retaining the right to revoke, but it governs who participates, not what is produced. Tool production, local rule adaptation, and dispute resolution are distributed by design across Champions, Co-Champions, and Wizards, in line with Ostrom's (1990) Principle 8 (governance organized in nested layers, each with defined scope, as Table 3 maps). The difference from a delegated hierarchy is that this authority inheres in the roles by the rules of the Community of Practice rather than being a revocable grant from above; the difference from platform governance is that a platform owner retains editorial control over what is produced, while the Adoption Network does not. This is distributed authority by design rather than by delegation (Ostrom, 2010).

The three institutional moves yield payoffs that can be compared directly. Under hierarchical governance the organizational payoff is $\Pi_H = N\omega - \sigma$, where workers generate individual AI gains via shadow use, the firm bears unmanaged security cost σ, and no scalable organizational knowledge value is created. Under polycentric governance satisfying Propositions 4 and 6, the payoff is $\Pi_P = N\omega + V_F + N \cdot e - N \cdot c_c$: individual gains are retained, the organizational knowledge surplus $V_F > 0$ is released by credible commitment, Community of Practice externalities $N \cdot e > 0$ accrue, and $N \cdot c_c$ is the aggregate contribution cost. The net organizational gain from adopting polycentric governance is:

$$\Delta\Pi \;=\; \Pi_P - \Pi_H \;=\; V_F + N \cdot e + \sigma - N \cdot c_c. \tag{4}$$

Individual output ω is regime-invariant and cancels from $\Delta\Pi$, so the design problem reduces to the incremental value the structure captures or forfeits, namely V_F, $N \cdot e$, and σ. The break-even surplus and a full numerical calibration are reported in Appendix A.

That calibration traces Theorem 1 at representative values: $\delta^{**} \approx 0.0006$ under polycentric governance versus $\delta_i^H = 0.50$ under hierarchy at the conservative benchmark $v_i = \tau$, an order-of-magnitude gap that widens as AI capability grows, since the polycentric threshold falls with $V_F(G(c))$ (Remark 1) while the

hierarchical one does not.

Proposition 7 (Polycentric Dominance)**.** *Polycentric governance strictly dominates hierarchical governance when the organizational knowledge surplus, eliminated security risk, and Community of Practice network externalities jointly exceed aggregate contribution costs ($\Delta\Pi > 0$, equation (4)). The condition is non-trivial. It fails when productive AI knowledge is not dispersed, security exposure from shadow adoption is low, or governance costs exceed per-contributor externalities. It holds when productive AI knowledge is widely dispersed across the workforce, shadow adoption poses meaningful unmanaged security risk, and the contributor community is large enough to generate substantial network externalities.*

Proposition 8 (Worker Welfare Under Polycentric Governance)**.** *Under polycentric governance satisfying Propositions 4 and 6:*

(i) High-type workers are strictly better off. *Each active contributor receives $\theta_H\pi - c_c$, which exceeds the hierarchical shadow-adoption payoff $\pi - k$ by $(\theta_H - 1)\pi + k - c_c$; the screening condition $(\theta_H - 1)\pi > c_c$ makes this strictly positive.*

(ii) Low-type workers are also strictly better off. *Declining the contributor tier, they take managed access at π, which exceeds their hierarchical shadow payoff $\pi - k$ by $k > 0$; the move out of shadow also removes the disciplinary risk $d > 0$ borne under unauthorized use.*

Polycentric governance is therefore a strict Pareto improvement over the hierarchical equilibrium for all workers.

Proof. See Appendix A.

Equivalently, dominance requires $V_F + \sigma > N(c_c - e)$: the knowledge surplus plus the eliminated security cost must exceed the net governance cost across contributors. An intermediate baseline sharpens the comparison. Permitting access but leaving the hold-up unresolved yields $\Pi_I = N\omega$: managed access removes the security exposure σ, but concealment still eliminates V_F and the externalities $N \cdot e$. Polycentric governance dominates this baseline whenever $V_F + N \cdot e > N \cdot c_c$, so the three-way ranking $\Pi_H < \Pi_I < \Pi_P$ holds when $\sigma > 0$ and the dominance condition is satisfied.

These dominance results hold even when the hierarchical benchmark is strengthened to an optimized hierarchy, in which senior management overrides IT's restriction preference, permits managed AI access, and invests in centralized monitoring at per-worker cost $m > 0$. To make the comparison precise, ask how much of a worker's productive method a manager could reconstruct just by watching output. Routine, templated tasks are almost fully recoverable this way; tasks built on contextual judgment or workflow-embedded adaptation are not. Let $\rho_j \in [0,1]$ be this *codifiability* in task j (near 1 for routine tasks, near 0 for judgment-heavy ones), and $\phi \equiv \mathbb{E}_j[\rho_j]$ the worker's *monitoring recoverability*, the share of method a manager could recover on average. The strict inequality $\phi < 1$ follows from Proposition 3: V_F consists of tacit, task-specific knowledge that workers acquire through use and can withhold even while complying with output-based oversight (Crémer et al., 2007; Szulanski, 1996; von Hippel, 1994); the hold-up logic applies to productive methods exactly as it applies to explicitly disclosed knowledge.

The tacit share $1-\phi$ rises with AI capability: the problems newly accessible to production workers, the band $(z_w^*(c_0), z_w^*(c))$, are the low-codifiability tasks ($\rho_j \approx 0$) needing contextual judgment, so as c falls ϕ falls while V_F widens (Remark 1). The surplus beyond monitoring's reach, $(1-\phi)V_F$, therefore grows more than proportionally with capability, strengthening the polycentric advantage as the technology improves.

Two mechanisms underlie this. The first is genuine tacitness: some task-specific knowledge cannot be articulated even when workers want to, because productive AI use settles into workflow habits and judgment calls, the prompt phrasings and the sense of when to trust a model that a worker reaches for without quite being able to write them down (von Hippel, 1994). The second is strategic suppression: workers withhold method knowledge they could articulate, because disclosure invites extraction. Polycentric governance reaches only the second. By removing the extraction threat it makes suppression individually irrational, while the irreducible tacit floor ($\phi < 1$) remains. Its value is the recovery of the strategically suppressed share, which monitoring cannot reach at any intensity. Let $\sigma_H \in [0, \sigma]$ denote the residual security risk under managed hierarchical access. The optimized hierarchical payoff is:

$$\Pi_H^{\text{opt}} = N\omega + \phi \cdot V_F - N \cdot m - \sigma_H. \tag{5}$$

Polycentric governance dominates even this strengthened benchmark whenever:

$$\Delta\Pi^{\text{opt}} \;=\; (1-\phi)\cdot V_F + N\cdot e + N\cdot m - N\cdot c_c + \sigma_H \;>\; 0. \tag{6}$$

Each term has a direct interpretation: $(1-\phi)V_F$ is the tacit surplus monitoring cannot recover, $N \cdot e$ the network externalities no hierarchy can replicate, $N \cdot m$ the monitoring costs community self-governance avoids, and σ_H the residual security exposure optimized hierarchy still bears. The condition $\Delta\Pi^{\text{opt}} > 0$ holds under the same scope conditions as Proposition 7.

The advantage is largest when the tacit share $(1-\phi)$, the surplus V_F, the community size N, and the per-worker monitoring cost m are all high, and it erodes when c_c rises well above $e+m$. Large firms with dispersed, tacit, hard-to-codify knowledge and high oversight costs gain most; where AI knowledge is easily codifiable, communities are small, or monitoring is cheap, the two payoffs converge and the governance choice matters less.

These results are consistent with, though not tested by, the BBVA case: organizational knowledge value is predicted to be zero under hierarchy and strictly positive under polycentric governance once the dominance condition holds, the pattern the case displays. Cross-organizational evidence points the same way. Kathuria et al. (2026) find that centrally governed AI implementations underperform decentralized ones, with outcomes improving as governance authority moves to frontline units, as Proposition 7 predicts.

5 The GPT Repository as a Knowledge Commons

The shared GPT repository poses a collective governance problem familiar from knowledge management research (Brown and Duguid, 1991). Workers must jointly populate, maintain, and improve a shared repository whose value depends on the quality and quantity of contributions, yet each individual act of contributing

is costly and the benefits accrue to all. The repository constitutes a non-tangible knowledge commons in the sense of Hess and Ostrom (2007), who extend Ostrom's (1990) framework to digital knowledge systems: it pairs low excludability within the licensed group with a subtractable input, though the subtractable element is not the tools themselves.

The GPT tools are non-rival, but *curatorial attention*, the finite capacity of Wizards to evaluate, document, and maintain tools, is not: a low-quality, undocumented, or duplicate tool consumes the community's evaluation and search capacity. The governed object is therefore quality-adjusted curatorial capacity, and free-riding on it is the appropriation problem the design must solve, which is also why the repository is a common-pool resource rather than a pure club good. Excludability is low within the licensed group, since formal exclusion requires revoking licenses, a sanction but not a default. Lerner and Tirole (2002) identify the same structure in open-source production, but the intra-firm repository differs in two respects: the Community of Practice creates repeated interaction with observable defection that anonymous open-source networks lack, making collective punishment more credible; and the contribution requirement resolves the Lerner–Tirole selection problem through the artificial scarcity that screens types in Proposition 4, not through external-market reputation.

Table 3 maps the eight design principles (the institutional features associated with long-enduring, self-governing common-pool resource arrangements) onto the governance design under study.

[Insert Table 3 about here]

Within this commons, Wizards face a coordination problem the design principles alone do not resolve. Each can contribute a tool or withhold it for private use; contribution is costly (documentation, maintenance, explanation), while withholding is cheaper but destroys the network externality the commons depends on. This is the classic free-rider problem in team production (Holmström, 1982): contributing is collectively optimal but individually rational only when each Wizard expects enough others to contribute, the structure of the Stag Hunt analyzed by Ostrom (1998). Resolving the hold-up (Proposition 6) is a precondition for the cooperative equilibrium: workers will not build for a commons they expect to be extracted from, and removing the extraction threat removes the reason to withhold regardless of what others do. Denrell et al. (2026) formalize a related obstacle as Rogers' paradox (frictionless sharing does not raise collective payoffs above the individual baseline) and show that an absorptive-capacity condition resolves it: individual AI use must precede meaningful contribution. The contribution requirement functions precisely as this condition, making participation individually rational only for workers who already use the technology productively.

Two mechanisms then resolve the coordination problem. Wizard status is a publicly visible rank tied to contribution, so the reputational return raises the private payoff to cooperation; and the Community of Practice creates repeated interaction with observable defection and graduated sanctions (Principle 5, Table 3), which by Proposition 6 supports cooperation as a subgame-perfect equilibrium.

BBVA's Wizard community produced more than 4,800 custom GPT tools (Alfaro et al., 2026). Ungoverned adoption would have produced a sparser, lower-quality stock concentrated in informal pockets; hierarchical production would have concentrated tool development in formally authorized specialist units.

We treat BBVA as a theory-informing case in Siggelkow's (2007) sense: the model stands on its own derivation, and the case illustrates its mechanisms rather than testing or generating them, on evidence that

is a single practitioner account (Alfaro et al., 2026). Table 4 records, for each mechanism, the observable implication, the corresponding evidence, and the most plausible alternative the case cannot rule out. Evidentiary strength is uneven. Screening and the repository leave directly observable traces, such as participation rates and the tool stock. The no-extraction commitment is hardest to document, because the relevant counterfactual is a managerial action that did not occur; the account reports a public amnesty and an augmentation pledge, but whether these were a credible commitment or merely preceded disclosure cannot be settled from one case, so we treat that mechanism as inferred from participation and mark it as such.

[Insert Table 4 about here]

[Insert Figure 3 about here]

6 Implications and Scope Conditions

The framework applies most directly where three conditions hold: productive AI knowledge is dispersed across the workforce rather than concentrated in a specialist unit; the adoption barrier is low enough that shadow adoption is viable; and monitoring costs scale with firm size.

Dispersed productive knowledge is the basic condition (Hayek, 1945): workers hold real authority (Aghion and Tirole, 1997) over how the tool applies to their tasks, and that knowledge is sticky (von Hippel, 1994), costly for IT to evaluate from outside. Where AI instead serves a single, well-defined task with knowledge concentrated in a specialist unit, Garicano's (2000) logic applies and hierarchy remains efficient: $\Delta\Pi > 0$ fails because V_F is small, monitoring recovers most of it (ϕ near 1), and the contributor community is too small for network externalities ($N \cdot e \approx 0$). The customer-service deployment of Brynjolfsson et al. (2025) is such a case: the cost shock is genuine, but the task is concentrated and low in tacitness, so the gain materializes without governance redesign.

A low adoption barrier is equally necessary: shadow adoption is viable only when a browser and a consumer account suffice, not when the technology needs significant IT infrastructure or integration, where the hold-up of Table 2 does not arise. Finally, the welfare case grows with firm size, since monitoring cost scales with headcount while the community externality $N \cdot e$ scales more than proportionally, so the analysis applies most forcefully to large, distributed organizations where centralized monitoring is infeasible.

Why no hierarchical design achieves the polycentric outcome

A natural challenge is whether an *optimized* hierarchy (senior management overriding IT, risk-adjusted licensing, a corporate AI program, or an internal center of excellence) could produce the same outcomes as the polycentric design, or whether less-hierarchical organizational forms short of full polycentric governance would suffice (Lee and Edmondson, 2017; Puranam et al., 2014). The dominance condition $\Delta\Pi > 0$ would be uninteresting if a well-managed hierarchy could achieve $V_F + N \cdot e$ through centralized means. Three arguments establish that it cannot, and each corresponds to one of the three polycentric mechanisms.

A hierarchical licensing program can select workers but cannot screen on productive use without recreating the hold-up, since any request to self-report productivity triggers strategic suppression (Proposition 3);

a centrally announced amnesty is not self-enforcing, because the firm still prefers extraction given revelation (Proposition 6); and a central repository captures generic applications but misses the workflow-specific adaptations that constitute V_F, the knowledge-integration problem in Grant's (1996) sense, because that knowledge is sticky (Hansen, 1999; von Hippel, 1994). Each gap traces to the same root, the authority to observe and sanction, which a hierarchy cannot distribute without ceasing to be one. The argument rests on one property that distinguishes a commons from a portfolio of bilateral exchanges.

Assumption 1 (Super-additive commons value)**.** *Let $v_i > 0$ be the firm's per-period gain from worker i's disclosure in isolation. The commons is worth at least the sum of its standalone bilateral relationships, $\sum_i v_i \leq V_F$, because each contribution raises the value of every other through the network externality e. With symmetric contributors this is $v_i \leq V_F/N$, strict whenever $N > 1$ and the externality is strictly super-additive.*

Under hierarchical governance, worker i's bilateral self-enforcement condition requires the firm's patience to satisfy:

$$\delta \;\geq\; \frac{\tau}{v_i + \tau} \;\equiv\; \delta_i^H, \tag{7}$$

where v_i is the firm's gain from that single bilateral relationship alone. In words, each threshold is the patience the firm needs to keep its promise: δ_i^H when punishment is bilateral, so only the wronged worker withdraws, and δ^{**} when punishment is collective. By Assumption 1, $v_i \leq V_F/N < V_F$, so with $s > 0$ the bilateral threshold is strictly harder to clear than $\delta^{**} = (\tau - s)/(V_F + \tau - s)$: punishment scaled to the whole commons asks far less patience than punishment scaled to one relationship. The full chain $\delta^{**} < \delta^{*} \leq \delta_i^H$ is derived in Appendix A.

Theorem 1 (Observability asymmetry)**.** *Hold the governance primitives fixed and let the operative variable be the* observability-and-sanction structure*: whether an extraction event is observable to, and sanctionable by, the affected worker alone (bilateral) or the full contributor community (distributed). Under a bilateral structure, punishment is capped at a single worker's withdrawal, a threat calibrated to the bilateral value $v_i \leq V_F/N$ (Assumption 1); under a distributed structure, any extraction triggers collective withdrawal calibrated to the full commons value V_F. The set of environments in which the non-extraction commitment is self-enforcing is therefore strictly larger under the distributed structure, $\delta^{**} \leq \delta_i^H$, with strict inequality whenever $N > 1$ and $s > 0$. The binding constraint is the observability-and-sanction structure, not ownership form.*

What ties this structure to polycentric governance, and severs it from hierarchy, is that distributed observability is not separately contractible. A firm that retains centralized sanction authority gains nothing by publicizing its own extraction and will not do so, so bilateral observability is the default whenever the party that sanctions also controls disclosure. Distributed observability is instead a byproduct of the production architecture: when contribution is public and the repository is shared (Section 5), extraction is visible without anyone choosing to reveal it, and the authority to sanction already resides with the peers who observe it. This sharpens the result rather than weakening it. Ownership is orthogonal: arrangements hierarchical in ownership that nonetheless cede extraction observability and sanction authority to a peer community, such as anonymous reporting channels, gain-sharing with peer monitoring, or worker councils, are polycentric

in the operative sense and inherit the lower threshold δ^{**}. What no hierarchy can do is retain centralized control over sanctions and still manufacture credible collective observability, and that is the combination Theorem 1 rules out.

Corollary 2 (Hierarchical ceiling on organizational value)**.** *No governance design that preserves bilateral observability of extraction events and centralized control over sanctions achieves the polycentric payoff* Π_P*. Three structural barriers bind simultaneously.*

Screening barrier. *Conditioning selection on self-reported productivity triggers strategic suppression (Proposition 3); only the cost-based threshold of Proposition 4 separates types without disclosure, and IT- or HR-run programs cannot replicate it.*

Commitment barrier. *By Theorem 1, hierarchical commitment requires strictly greater patience, and a non-empty range of contexts exists in which polycentric commitment is self-enforcing while any design with bilateral observability fails.*

Knowledge ceiling. *Centralized monitoring recovers at most* ϕV_F *with* $\phi < 1$*; the tacit share* $(1-\phi)V_F > 0$ *stays inaccessible regardless of intensity.*

Consequently, $\Pi_H^{\text{opt}} < \Pi_P$ *whenever* $\Delta\Pi^{\text{opt}} > 0$ *(equation* (6)*). The three mechanisms are jointly necessary, and their joint implementation is what constitutes polycentric governance.*

When polycentric governance fails

Regulatory override is the most binding limit. Where AI feeds regulated outputs, such as credit decisions, clinical assistance, or model-risk-managed processes, external authorities may require centralized accountability that no distributed design can satisfy, and personal liability on IT officers can make the no-punishment commitment impermissible. BBVA illustrates the boundary: the Adoption Network governed productivity tools, while credit scoring and anti-money-laundering models stayed under centralized oversight. Regulated firms therefore face a bifurcated problem, with polycentric design serving the internal productivity layer and centralized accountability the regulated-output layer.

Commitment credibility is a second, independent limit. The no-punishment commitment must be believed, not merely announced (Proposition 6). Where a firm has used disclosures against workers before, an amnesty will not change behavior; the Community of Practice makes defection observable but cannot substitute for a track record, so such firms face a higher threshold and may need structural separation between the Adoption Network and HR.

The screening mechanism also has limits. It depends on genuine heterogeneity in worker types (Proposition 4): where gains are broadly similar, as in high-volume repetitive processing, the threshold cannot separate types. The design is therefore most valuable where work is diverse and judgment-intensive and the variance in AI gains across individuals is high.

The model also abstracts from AI's effect on workers' general capabilities, which transfer across employers. This creates two tensions outside the framework: productivity gains may surface as discretionary time rather than firm output, and stronger general capabilities improve outside options, sharpening both IT's incentive to restrict and workers' to conceal. Since V_F is firm-specific, these spillovers are a natural extension.

Public-sector organizations face a distinct version: employment protections, standardized pay, and political oversight shape adoption differently (Pieroni and Rossello Roig, 2026), so governance reform there is a political as much as a managerial decision and needs separate analysis.

A more fundamental limit concerns the transition itself. The model treats polycentric governance as an equilibrium once in place, but moving there from (Restrict, Shadow) requires a change no single actor has an incentive to start: IT restricts until authority is reallocated, workers will not disclose until they believe the amnesty, and the amnesty is not credible until the observability structure is already operating. The transition is a coordination game whose Pareto-superior equilibrium needs simultaneous moves, with no first-mover gain.

Proposition 9 (Credible transition to polycentric governance)**.** *Suppose firms differ in an unobservable primitive, the discount factor: a* committed *firm is patient, $\delta_C \geq \delta^{**}$, so honoring the no-extraction agreement is sequentially rational for it once community observability is operative (Proposition 6); an* opportunistic *firm is impatient, $\delta_O < \delta^{**}$, and extracts whenever profitable. Workers observe the firm's investment but not its discount factor, and the signal operates in the transition window before the community's observability is established, when the discipline of Proposition 6 does not yet bind. A separating equilibrium in which committed firms successfully initiate polycentric governance exists if and only if the firm publicly undertakes an investment I^* in the separating range*

$$I^* \in \left(\tau, \ \frac{V_F}{1-\delta} + \sigma\right].$$

The lower bound is the deterrence threshold: an opportunistic firm will not mimic, because mimicry costs I^* upfront and yields at most τ from extraction, a net loss once $I^* > \tau$. The upper bound is the committed firm's break-even, at which the cooperative continuation value plus the recovered security cost, $V_F/(1-\delta)+\sigma$, just covers the outlay. The range is non-empty whenever $V_F/(1-\delta) > \tau - \sigma$, which holds in particular when $\delta \geq \delta^* = \tau/(V_F+\tau)$ (Proposition 6).

Proof. See Appendix A.

The threshold I^* tells a committed firm what separation requires: a visible, costly investment in governance infrastructure before any disclosure is asked; a new governance unit publicly separated from the HR functions that previously used disclosures against workers; and an amnesty that explicitly covers prior shadow use, recasting it as evidence of demand rather than grounds for discipline. The BBVA evidence fits all three, and firms with a history of extraction face a compounded version, pushing I^* toward its upper bound.

Design implications

The analysis carries direct implications for how organizations respond to shadow AI, and those implications run counter to common practice.

The most immediate is that restriction is not a safe default. The (Restrict, Shadow) equilibrium is strictly inferior to permission, generating unmanaged security risk with no governance benefit because shadow adoption moves AI use outside the firm's data agreements; restriction only looks safe from IT's perspective. The question for leadership is not whether to allow AI but how to govern it.

Access is not sufficient. Permission removes the restriction failure but leaves the revelation-stage hold-up intact, so a firm that opens access without resolving it observes individual use without sharing. Screening and commitment must be implemented jointly: artificial scarcity selects productive adopters, and the no-punishment commitment makes it safe for them to share.

The deeper implication concerns IT's role. Under polycentric governance IT's comparative advantage is infrastructure, enterprise contracts, data-governance standards, and security architecture, not access control. Dessein (2002) shows delegation to informed agents is optimal when they hold a decisive local information advantage; the Adoption Network is such delegation, IT holding the infrastructure layer and the community the knowledge layer. That infrastructure also enables what shadow adoption forecloses: grounding models on proprietary data through retrieval-augmented systems, an asset dispersed consumer-tool use cannot produce. Consistent with this, Asaftei et al. (2026) report that firms assigning explicit AI-governance accountability reach higher maturity. The model yields two testable predictions: (i) firms with centralized AI governance should show lower adoption, higher shadow use, and lower knowledge capture than comparable firms with polycentric governance; and (ii) the governance gap should widen as AI capability grows.

A further implication concerns managerial roles: as production-worker problem-solving expands, demand shifts from routine coordination toward strategic decision-making. Alekseeva et al. (2026) document a positive association between AI adoption and demand for strategic managerial skills in job-posting data; polycentric governance is the mechanism through which that shift is realized rather than merely anticipated.

Testing these predictions requires operationalizing three constructs: the governance gap $G(c)$, measurable as the deviation between observed adoption and the knowledge-hierarchy benchmark; the knowledge surplus V_F, proxied by repository size, contribution quality scores, or productivity divergence between contributor and non-contributor units; and a regime indicator coding whether governance authority sits with IT or with contributor communities, whether access is conditioned on contribution, and whether community sanctions operate independently of HR. Shadow-to-formal use ratios provide a complementary signal, and the plant-level autonomy and span instruments of Bloom et al. (2014) offer validated templates for measuring where AI decision rights actually sit.

Four questions remain open. First, how does the value gap between polycentric and hierarchical governance vary across firm types and regulatory environments, and can the dominance condition be tested with the measures above? Second, under what conditions can senior leadership credibly initiate redesign against IT's institutional resistance? Third, the static framework leaves dynamics unaddressed: as the contributor community grows, V_F and $N \cdot e$ compound, so early polycentric adoption may be a path-dependent source of advantage. Fourth, when AI enhances general rather than firm-specific capabilities, stronger outside options interact with retention and disclosure in ways the present framework does not capture.

7 Conclusion

This paper began with a governance gap. The knowledge hierarchy model identifies an efficiency condition under which, as knowledge acquisition costs fall, the efficient organization expands production worker discretion, contracts the specialist layer, and flattens. General-purpose AI constitutes such a cost reduction, at

a scale the model did not contemplate.

The dominant response has nonetheless been restriction. IT is not the social planner the model assumes but a bureau whose audit-defensibility incentives reward restriction regardless of organizational welfare. Workers respond in two ways: they adopt outside the formal system, generating shadow use the firm cannot govern, and they conceal the gains adoption produces, denying the firm the tacit knowledge value the technology creates. The second response is independent of the first. Even where firms have moved past restriction and granted broad access, concealment persists as long as disclosure invites extraction. Access policy resolves the restriction failure but leaves the revelation-stage hold-up intact.

Together, these failures leave the firm unable to capture the scalable organizational knowledge AI can produce, while under restriction it also bears the security cost of unmanaged shadow adoption. Polycentric governance resolves both jointly, converting a security liability into a governed knowledge commons through three institutional moves: a contribution requirement that turns access from a compliance problem into a productivity signal, a no-punishment commitment made self-enforcing through collective observability, and a community-governed repository that aggregates the tacit knowledge centralized collection cannot reach. Each corresponds to Ostrom's (1990) design principles for self-governing shared resources.

The failures identified here are structural, not managerial. Changing who sits at the top of the hierarchy does not alter the incentives that generate restriction and concealment. What changes them is the architecture of observability, and that is what polycentric governance provides. AI is what makes the stakes large enough to force the question, but the result is general. Wherever dispersed tacit knowledge meets self-interested gatekeeping, whether it is the analyst who has quietly halved her reporting time or any worker who has learned what a tool is good for, what determines whether the firm captures that knowledge is who can observe and sanction its extraction, not who owns the firm. How much organizational knowledge is currently suppressed across governance regimes, and whether polycentric redesign recovers it at scale, are the empirical questions the framework is designed to frame.

References

Aghion P, Tirole J (1997) Formal and real authority in organizations. *J. Political Econom.* 105(1):1–29.

Alekseeva L, Azar J, Giné M, Samila S (2026) Artificial intelligence adoption and the demand for managerial expertise. *Strategic Management J.*, ePub ahead of print, https://doi.org/10.1002/smj.70099.

Alfaro E, Cabrales A, Durán Roa JE, Garicano L, Pérez del Caño I, Roldán Monés T, Vieira de Santiago G (2026) The hidden demand for AI inside your company. *Harvard Bus. Rev.* (April 14), https://hbr.org/2026/04/the-hidden-demand-for-ai-inside-your-company.

Anton JJ, Yao DA (1994) Expropriation and inventions: appropriable rents in the absence of property rights. *Amer. Econom. Rev.* 84(1):190–209.

Anton JJ, Yao DA (2002) The sale of ideas: strategic disclosure, property rights, and contracting. *Rev. Econom. Stud.* 69(3):513–531.

Asaftei GM, Roberts R, Sticha A, Prinsen C (2026) State of AI trust in 2026: Shifting to the agentic era. Report, McKinsey & Company, New York. Accessed May 31, 2026, https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era.

Baker G, Gibbons R, Murphy KJ (2002) Relational contracts and the theory of the firm. *Quart. J. Econom.* 117(1):39–84.

Behrens S (2009) Shadow systems: the good, the bad and the ugly. *Comm. ACM* 52(2):124–129.

Bloom N, Garicano L, Sadun R, Van Reenen J (2014) The distinct effects of information technology and communication technology on firm organization. *Management Sci.* 60(12):2859–2885.

Bresnahan TF, Trajtenberg M (1995) General purpose technologies "engines of growth"? *J. Econometrics* 65(1):83–108.

Brown JS, Duguid P (1991) Organizational learning and communities-of-practice: toward a unified view of working, learning, and innovation. *Organ. Sci.* 2(1):40–57.

Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *Quart. J. Econom.* 140(2):889–942.

Caplin A, Deming D, Li S, Martin D, Marx P, Weidmann B, Ye KJ (2025) The ABCs of who benefits from working with AI: Ability, beliefs, and calibration. *Management Sci.*, ePub ahead of print, https://doi.org/10.1287/mnsc.2024.08994.

Černe M, Nerstad CGL, Dysvik A, Škerlavaj M (2014) What goes around comes around: knowledge hiding, perceived motivational climate, and creativity. *Acad. Management J.* 57(1):172–192.

Connelly CE, Zweig D, Webster J, Kelloway EK (2012) Knowledge hiding in organizations. *J. Organ. Behav.* 33(1):64–88.

Crémer J, Garicano L, Prat A (2007) Language and the theory of the firm. *Quart. J. Econom.* 122(1):373–407.

Denrell J, Luukkonen J, Chater N, Liu C (2026) The need for absorptive capacity alleviates the free-rider problem in knowledge production. *Management Sci.*, ePub ahead of print, https://doi.org/10.1287/mnsc.2024.06471.

Dessein W (2002) Authority and communication in organizations. *Rev. Econom. Stud.* 69(4):811–838.

Edmondson AC (1999) Psychological safety and learning behavior in work teams. *Admin. Sci. Quart.* 44(2):350–383.

Foss NJ, Husted K, Michailova S (2010) Governing knowledge sharing in organizations: levels of analysis, mechanisms, and research directions. *J. Management Stud.* 47(3):455–482.

Freixas X, Guesnerie R, Tirole J (1985) Planning under incomplete information and the ratchet effect. *Rev. Econom. Stud.* 52(2):173–191.

Garicano L (2000) Hierarchies and the organization of knowledge in production. *J. Political Econom.* 108(5):874–904.

Gibbons R (1987) Piece-rate incentive schemes. *J. Labor Econom.* 5(4):413–429.

Grandori A (2001) Neither hierarchy nor identity: knowledge-governance mechanisms and the theory of the firm. *J. Management Governance* 5(3–4):381–399.

Grant RM (1996) Toward a knowledge-based theory of the firm. *Strategic Management J.* 17(S2):109–122.

Gulati P, Marchetti A, Puranam P (2026) Collaborative work management technologies and managerial intensity in U.S. corporations: An examination. *Management Sci.*, ePub ahead of print, `https://doi.org/10.1287/mnsc.2023.04127`.

Hansen MT (1999) The search-transfer problem: the role of weak ties in sharing knowledge across organization subunits. *Admin. Sci. Quart.* 44(1):82–111.

Hart O, Moore J (1990) Property rights and the nature of the firm. *J. Political Econom.* 98(6):1119–1158.

Hayek FA (1945) The use of knowledge in society. *Amer. Econom. Rev.* 35(4):519–530.

Hess C, Ostrom E (2007) *Understanding Knowledge as a Commons: From Theory to Practice* (MIT Press, Cambridge, MA).

Holmström B (1982) Moral hazard in teams. *Bell J. Econom.* 13(2):324–340.

Ide E, Talamàs E (2025) Artificial intelligence in the knowledge economy. *J. Political Econom.* 133(12):3762–3800.

Ide E, Talamàs E (2026) The impact of AI on global knowledge work. *J. Monetary Econom.* 157:103876.

Kathuria A, Karhade PP, Malik O, Baruri VKP, Konsynski BR (2026) AI governance and the decentralization of technology production: An investigation of AI-based IPA bots. *Inform. Systems Res.*, ePub ahead of print, `https://doi.org/10.1287/isre.2023.0588`.

Klein B, Crawford RA, Alchian AA (1978) Vertical integration, appropriable rents, and the competitive contracting process. *J. Law Econom.* 21(2):297–326.

Kogut B, Zander U (1992) Knowledge of the firm, combinative capabilities, and the replication of technology. *Organ. Sci.* 3(3):383–397.

Laffont JJ, Tirole J (1988) The dynamics of incentive contracts. *Econometrica* 56(5):1153–1175.

Lee MY, Edmondson AC (2017) Self-managing organizations: exploring the limits of less-hierarchical organizing. *Res. Organ. Behav.* 37:35–58.

Lerner J, Tirole J (2002) Some simple economics of open source. *J. Indust. Econom.* 50(2):197–234.

Levin J (2003) Relational incentive contracts. *Amer. Econom. Rev.* 93(3):835–857.

Lifshitz-Assaf H (2018) Dismantling knowledge boundaries at NASA: the critical role of professional identity in open innovation. *Admin. Sci. Quart.* 63(4):746–782.

Magelssen C, Rich B, Mayer K (2022) The contractual governance of transactions within firms. *Organ. Sci.* 33(6):2226–2249.

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

Nickerson JA, Zenger TR (2004) A knowledge-based theory of the firm: the problem-solving perspective. *Organ. Sci.* 15(6):617–632.

Niskanen WA (1971) *Bureaucracy and Representative Government* (Aldine Atherton, Chicago).

North DC (1990) *Institutions, Institutional Change and Economic Performance* (Cambridge University Press, Cambridge, UK).

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, Cambridge, UK).

Ostrom E (1998) A behavioral approach to the rational choice theory of collective action. *Amer. Political Sci. Rev.* 92(1):1–22.

Ostrom E (2010) Beyond markets and states: polycentric governance of complex economic systems. *Amer. Econom. Rev.* 100(3):641–672.

Pieroni L, Rossello Roig M (2026) AI adoption in bureaucracies. *J. Institutional Econom.* 22:e15.

Puranam P, Raveendran M, Knudsen T (2012) Organization design: the epistemic interdependence perspective. *Acad. Management Rev.* 37(3):419–440.

Puranam P, Alexy O, Reitzig M (2014) What's "new" about new forms of organizing? *Acad. Management Rev.* 39(2):162–180.

Raisch S, Krakowski S (2021) Artificial intelligence and management: the automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Rajan RG, Wulf J (2006) The flattening firm: evidence from panel data on the changing nature of corporate hierarchies. *Rev. Econom. Statist.* 88(4):759–773.

Renieris EM, Kiron D, Mills S (2023) Building robust RAI programs as third-party AI tools proliferate. Report, MIT Sloan Management Review and Boston Consulting Group, Cambridge, MA. Accessed May 31, 2026, `https://sloanreview.mit.edu/projects/scholars/building-robust-rai-programs-as-third-party-ai-tools-proliferate/`.

Siggelkow N (2007) Persuasion with case studies. *Acad. Management J.* 50(1):20–24.

Silic M, Back A (2014) Shadow IT – a view from behind the curtain. *Comput. Security* 45:274–283.

Szulanski G (1996) Exploring internal stickiness: impediments to the transfer of best practice within the firm. *Strategic Management J.* 17(S2):27–43.

Tambe PB (2026) Reskilling the workforce for AI: domain expertise and algorithmic literacy. *Management Sci.* 72(1):515–537.

Tullock G (1965) *The Politics of Bureaucracy* (Public Affairs Press, Washington, DC).

von Hippel E (1994) Sticky information and the locus of problem solving: implications for innovation. *Management Sci.* 40(4):429–439.

Weitzman ML (1980) The "ratchet principle" and performance incentives. *Bell J. Econom.* 11(1):302–308.

Wenger E (1998) *Communities of Practice: Learning, Meaning, and Identity* (Cambridge University Press, Cambridge, UK).

Williamson OE (1985) *The Economic Institutions of Capitalism* (Free Press, New York).

Table 1

Table 1: The IT–worker game

	Worker: Comply	**Worker: Shadow**
IT: Permit	$(0,\ \pi)$	$(-\sigma,\ \pi)$
IT: Restrict	$(b,\ 0)$	$(b-\alpha\sigma,\ \pi-k)$

Note: Payoffs are ordered (IT, Worker). $\alpha \in [0,1]$ is the residual blame fraction borne by IT when it has formally restricted access but shadow adoption still occurs: $\alpha = 0$ is full blame deflection; $\alpha = 1$ is no deflection. Under (Permit, Shadow), IT bears $-\sigma$ in full since no restriction policy exists to deflect accountability. Under Permit the worker earns π in either cell, since permitted unofficial use requires no concealment and k drops out; we resolve the indifference toward Comply. Restrict strictly dominates Permit for all $b > 0$ and $\alpha \leq 1$. Against Comply, $b > 0$ suffices; against Shadow, $b - \alpha\sigma > -\sigma$ iff $b > (\alpha - 1)\sigma \leq 0$, which holds whenever $b > 0$. Here b is treated as an exogenous parameter; under the microfoundation $b = p(c)L(1-\alpha)$ (Definition 2), the knife-edge $\alpha = 1$ yields $b = 0$ and only weak dominance, consistent with Proposition 1.

Table 2

Table 2: The worker–firm hold-up game

	Firm: Commit	**Firm: Extract**
Worker: Reveal	$(V_F,\ \pi-\kappa)$	$(V_F+\tau,\ \pi-\tau-\kappa)$
Worker: Hide	$(0,\ \pi-k)$	$(0,\ \pi-k)$

Note: Payoffs are ordered (Firm, Worker). V_F = organizational value from scaling AI-enabled productivity knowledge; τ = extraction gain from renegotiation; κ = revelation cost; k = concealment cost (notation unified with Table 1). Extract weakly dominates Commit: it yields $V_F + \tau > V_F$ strictly when the worker reveals, and 0 whether or not the worker hides (identical to Commit). By backward induction, the firm extracts upon any revelation; anticipating this, the worker hides when $\tau > k - \kappa$ (assumed throughout).

Table 3

Table 3: Ostrom's design principles and BBVA's GPT commons governance

Design principle	Standard CPR form	BBVA implementation	Mapping
1. Defined boundaries	Boundaries of the resource and the user group are clearly specified	Licensed users only; Adoption Network defines membership; non-members cannot contribute or access GPT tools	Literal
2. Rules fit local conditions	Appropriation and provision rules match local resource conditions	Task-specific GPTs; Champions adapt rules to departmental use cases rather than imposing a universal standard	Analogy
3. Collective-choice arrangements	Those affected by the rules participate in modifying them	Wizards and Champions co-design contribution requirements; the Quality Score system is governed by community consensus, not IT mandate	Literal
4. Monitoring	Monitors audit resource conditions and user behavior	Automated Quality Score evaluates tool guardrails, clarity, and context; peer review by Wizards	Analogy
5. Graduated sanctions	Violations receive sanctions scaled to severity and context	License revocation for non-contribution; reduced Wizard status for undocumented tools; amnesty for legacy shadow AI	Analogy
6. Conflict-resolution mechanisms	Low-cost arenas for dispute resolution exist among users	Community of Practice provides a governance forum; Champions mediate disputes between departments; escalation to management is a last resort	Literal
7. Recognition of rights	External authorities recognize the right of users to self-organize	Senior leadership publicly endorsed the Adoption Network's autonomy; IT department's formal authority is bounded by the amnesty commitment	Literal
8. Nested enterprises	Governance is organized in nested layers for larger systems	Team-level (Co-Champions) $\rightarrow$ Department-level (Champions) $\rightarrow$ Firm-wide (Community of Practice); each layer has defined scope	Literal

Note: Design principles follow Ostrom (1990). BBVA implementation details are drawn from Alfaro et al. (2026), a practitioner account published in Harvard Business Review; the mapping is a qualitative illustration of the model mechanisms, not a coded case study. The **Mapping** column marks whether each principle transfers literally or by analogy. Principles governing the rivalrous appropriation of resource units (2, 4, 5) transfer by analogy, because the GPT tools are non-rival in use; following Hess and Ostrom (2007), the subtractable input is curatorial attention rather than the resource itself, so these principles apply to contribution quality and congestion rather than to extraction. Principles governing boundaries and governance structure (1, 3, 6, 7, 8) transfer literally. CPR = common-pool resource; GPT = generative pre-trained transformer (used here as a generic label for custom AI tools built on a foundation model).

Table 4

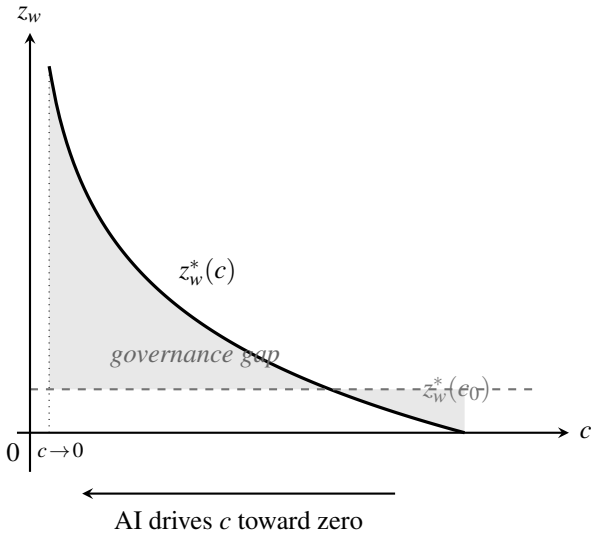
Table 4: BBVA as a theory-informing case: mechanism-evidence mapping

Model mechanism	Observable implication	BBVA evidence	Alternative explanation
Screening via artificial scarcity (Prop. 4)	Productive users self-select into the contributor pool; marginal users remain in informal use	83% of license holders used the platform weekly, averaging 50 prompts per week; Quality Score gated contribution	Licenses were allocated to already-motivated early adopters, so selection reflects enthusiasm rather than the contribution threshold
No-extraction commitment (Prop. 6) *[inferred]*	Workers disclose tacit workflows; management does not raise targets or cut headcount in response to disclosure	Public amnesty reframed prior shadow use as productive demand; leadership pledged augmentation rather than replacement	Disclosure was driven by intrinsic or career motives; the augmentation pledge may not have constituted a binding commitment
Collective observability (Thm. 1)	Extraction would be visible to the whole community and trigger collective withdrawal	Wizards interact publicly, share a common repository, and observe governance events jointly	Observed cooperation may reflect a short horizon without any tested extraction event
Community repository as knowledge commons (Prop. 7)	A large, maintained stock of community-built tools accumulates under distributed curation	4,800 custom GPT tools built and maintained by the Wizard community	Tool count reflects firm size and centralized IT promotion rather than commons governance
Governance gap under restriction (Prop. 1)	Restriction precedes adoption; shadow use emerges at scale before formal governance	The workforce adopted AI through informal channels before any formal governance existed	Ordinary diffusion lag rather than a restriction equilibrium

Note: Evidence is drawn from Alfaro et al. (2026), a practitioner account published in Harvard Business Review. The table records the case as theory-informing in Siggelkow's (2007) sense, not as a test. The mechanism marked *[inferred]* is the no-extraction commitment, whose evidence rests on a managerial action whose counterfactual cannot be observed in a single case; it is inferred from participation rather than directly documented. The alternative-explanation column lists the rival account that the case alone cannot exclude, which the testable predictions in Section 6 are designed to separate.

Figure 1

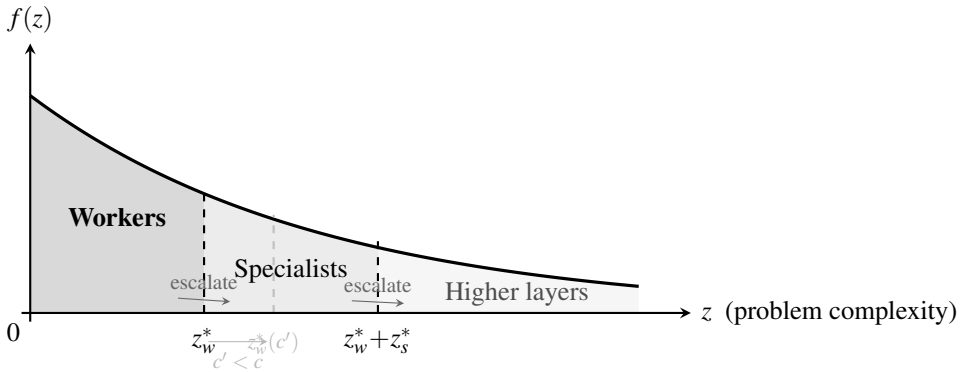
Figure 1: The institutional governance gap



Note: The curve $z_w^*(c)$ is the socially efficient production worker knowledge threshold derived from Garicano (2000); it rises without bound as the knowledge acquisition cost c falls toward zero. The dashed line $z_w^*(c_0)$ is the threshold IT implements by fixing knowledge acquisition costs at the pre-AI baseline $c_0 > c$; it remains constant regardless of AI capability because IT's payoff depends on accountability minimization, not organizational output. The shaded region is the governance gap: the AI-enabled knowledge allocation forgone by operating at $z_w^*(c_0)$ rather than $z_w^*(c)$. General-purpose AI constitutes a large downward shock to c, shifting the economy toward the left of the figure and expanding the governance gap.

Figure 2

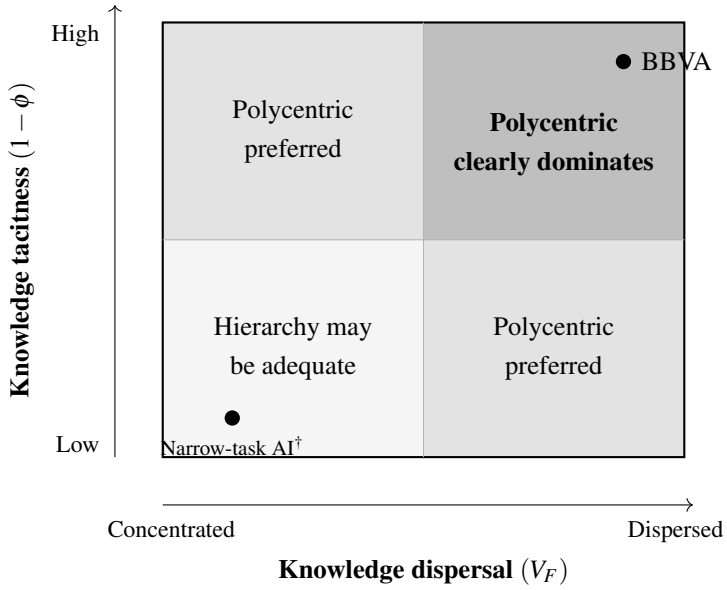
Figure 2: The knowledge hierarchy: partitioning the problem distribution



Note: The curve plots the exponential problem distribution $f(z) = \lambda e^{-\lambda z}$, where $z \geq 0$ indexes problem complexity. The dark dashed lines mark the efficient knowledge thresholds z_w^* and $z_w^* + z_s^*$ derived in equations (1)–(2). Workers handle the most frequent (low-complexity) problems in $[0, z_w^*]$; specialists handle exceptions in $[z_w^*, z_w^* + z_s^*]$; higher layers handle the rare tail. The lighter dashed line shows $z_w^*(c')$, the efficient worker threshold under a lower knowledge-acquisition cost $c' < c$; the rightward shift expands the worker zone and absorbs problems previously requiring specialist escalation. Figure 1 shows the governance cost of failing to allow this shift.

Figure 3

Figure 3: Scope conditions for polycentric governance dominance



Note: The horizontal axis is the dispersal of productive AI knowledge across the workforce, which drives the recoverable organizational surplus V_F; the vertical axis is knowledge tacitness $(1-\phi)$. Polycentric governance clearly dominates in the upper-right quadrant (dispersed and tacit), where both $(1-\phi)V_F$ and the network externality are large. It is preferred but less decisively in the off-diagonal quadrants, where only one of the two conditions holds. The lower-left quadrant (concentrated and codifiable) approximates settings where Garicano's (2000) hierarchical model remains efficient. †Narrow-task AI (e.g., Brynjolfsson et al. 2025): knowledge concentrated in a defined task, low tacitness, small V_F; the dominance condition is unlikely to bind.

A Proofs

Proof of Proposition 1. IT compares payoffs across its two strategies. Against a Compliant worker, Restrict yields $b > 0$ and Permit yields 0; Restrict dominates since $b > 0$. Against a Shadow worker, Restrict yields $b - \alpha\sigma$ and Permit yields $-\sigma$; Restrict dominates iff $b - \alpha\sigma > -\sigma$, i.e. $b > (\alpha - 1)\sigma$. Since $\alpha \leq 1$, the right-hand side is $(\alpha - 1)\sigma \leq 0 < b$, so the condition holds for any $b > 0$ and $\alpha \in [0, 1]$. Restrict therefore strictly dominates Permit regardless of the worker's strategy and regardless of the degree of blame deflection. Given Restrict, the worker chooses Shadow whenever $\pi - k > 0$; since $\pi > k$ by assumption, Shadow is the unique best response. The pair (Restrict, Shadow) is the unique Nash equilibrium. Workers capture private AI gains $\pi - k > 0$ through shadow use; the firm captures individual AI output $N\omega$ through production but no scalable organizational knowledge value, and bears the full organizational security cost σ from unmanaged shadow use (σ is the total organizational exposure; $\alpha\sigma$ in Table 1 is IT's private accountability share, not the organizational cost). □

Proof of Proposition 3. Table 2 shows that when the worker reveals, Extract yields $V_F + \tau$ while Commit yields V_F; since $\tau > 0$, the firm strictly prefers Extract at the revelation node. When the worker hides, both Extract and Commit yield 0 (no knowledge is available to extract), so the firm is indifferent. Extract therefore weakly dominates Commit, with strict dominance conditional on revelation. By backward induction, the worker anticipates extraction upon any revelation and compares: $\pi - k$ (Hide) versus $\pi - \tau - \kappa$ (Reveal). Hide is strictly preferred when $\tau > k - \kappa$, which is satisfied by assumption. Since the firm extracts upon revelation, Reveal is never a best response; (Hide, Extract) is the unique subgame-perfect equilibrium. No AI-enabled productivity knowledge is disclosed, so the firm captures zero organizational knowledge value. □

Proof of Proposition 4. Worker i of type $\theta_i \in \{\theta_L, \theta_H\}$ with $\theta_L < \theta_H$ prefers contributor access over managed non-contributor access iff $\theta_i\pi - c_c \geq \pi$, i.e. $(\theta_i - 1)\pi \geq c_c$. Set $c_c \in \big((\theta_L - 1)\pi,\ (\theta_H - 1)\pi\big)$. This interval is non-empty because $\theta_H > \theta_L$ implies $(\theta_H - 1)\pi > (\theta_L - 1)\pi$. Individual rationality: $(\theta_H - 1)\pi > c_c$ so the high type participates; $(\theta_L - 1)\pi < c_c$ so the low type prefers managed access, yielding π. Incentive compatibility: the high type has no incentive to mimic the low type (managed access yields $\pi < \theta_H\pi - c_c$); the low type has no incentive to mimic the high type (contributor access yields $\theta_L\pi - c_c < \pi$, a net loss relative to managed access). Both conditions hold strictly for any c_c in the interior of the stated interval. The equilibrium is separating and unique given c_c. □

Proof of Proposition 6.

Step 1 (Worker's individual rationality post-amnesty). Let $d' \geq 0$ and $a' \geq 0$ denote the residual disciplinary risk and authenticity risk after amnesty, with $d' < d$ and $a' < a$. The amnesty design sets $d' = 0$ (unauthorized use is reframed as untapped demand, eliminating the disciplinary component) and reduces a to some $a' < a$ (public designation of AI as standard practice lowers the perception cost of disclosure). The concealment cost rises from k to $k + k_a$ over time as the authenticity burden of prolonged shadow use accumulates. Under the firm's no-extraction commitment (established in Step 2 below), the worker compares:

$$\begin{aligned} \text{Reveal:} \quad & \pi - (\kappa + d' + a') = \pi - \kappa - a', \\ \text{Hide:} \quad & \pi - (k + k_a). \end{aligned}$$

Revelation is individually rational iff $k + k_a \geq \kappa + a'$, i.e. the total concealment cost exceeds the post-amnesty revelation cost. This condition holds by design: the amnesty sets a' such that $a' < k + k_a - \kappa$, which is feasible since $k_a > 0$ and, by assumption, $k + k_a > \kappa$ (sustained concealment is more costly than disclosure once the authenticity burden accumulates). Hence, given credible commitment, every worker for whom the above holds strictly prefers Reveal.

Step 2 (Firm's self-enforcement condition). Consider the infinitely repeated worker–firm game with common discount factor $\delta \in (0,1)$. On the cooperation path, the worker reveals each period and the firm does not extract; the firm receives V_F per period, yielding discounted value $V_F/(1-\delta)$. If the firm deviates in period t, it gains $\tau - s$ net of peer sanctions $s > 0$ (public censure within the Wizard network); the Community of Practice collapses immediately, yielding zero thereafter. The firm's gain from deviation relative to cooperation is:

$$\underbrace{\tau - s}_{\text{net extraction gain}} \quad - \quad \underbrace{\delta \cdot \frac{V_F}{1-\delta}}_{\text{discounted future loss}} \quad .$$

The no-punishment commitment is self-enforcing as a subgame-perfect equilibrium if and only if this deviation gain is non-positive:

$$\delta \cdot \frac{V_F}{1-\delta} \geq \tau - s \quad \Longleftrightarrow \quad \delta \geq \frac{\tau - s}{V_F + \tau - s} \equiv \delta^{**} < \delta^{*}.$$

Community of Practice governance raises V_F (network externalities from each additional Wizard's contribution) and introduces peer sanctions $s > 0$ that reduce the net extraction gain from τ to $\tau - s$, while commons collapse makes the continuation punishment credible. All three effects strictly decrease δ^{**}, expanding the set of discount factors supporting cooperation as a subgame-perfect equilibrium relative to a community without peer sanctions.

Step 3 (Individual incentive compatibility of the punishment). The grim-trigger threat is credible only if each contributor finds it individually rational to withdraw after observing extraction, so that the commons collapse requires no coordinated decision. Consider contributor i in the period after an extraction event is publicly observed. Withdrawing reverts i to concealment (the Hide option of Proposition 3), yielding $\pi - k$ in every subsequent period. Continuing to contribute exposes i to the firm's extraction strategy, which extracts upon any revelation, yielding $\pi - \tau - \kappa$ per period. Withdrawal is individually rational whenever

$$\pi - k > \pi - \tau - \kappa \qquad \Longleftrightarrow \qquad \tau > k - \kappa,$$

which holds by the standing assumption (Table 2). Withdrawal therefore strictly dominates continuation for every contributor after observed extraction, independent of what other contributors do. The collective punishment is a dominant-strategy response rather than a coordinated one, so the commons collapse is subgame-perfect and the threat sustaining the firm's commitment is credible.

Combining Steps 1–3. Step 1 establishes that workers strictly prefer Reveal given Commit; Step 2 establishes that the firm strictly prefers Commit given $\delta \geq \delta^{**}$; Step 3 establishes that the punishment supporting the firm's incentive is itself individually incentive-compatible for each contributor. The pair (Reveal, Commit) is therefore a subgame-perfect equilibrium for all $\delta \in [\delta^{**}, 1)$, and workers disclose AI-enabled productivity, allowing the firm to capture the full organizational knowledge surplus V_F. □

Proof of Proposition 7. *Step 1: Hierarchical payoff.* By Proposition 1, the equilibrium under hierarchy is (Restrict, Shadow). N workers each produce individual AI output $\omega > 0$ through shadow use, so the firm captures $N\omega$ through normal production channels. The firm bears security cost $\sigma > 0$ from unmanaged shadow access. No worker discloses AI-enabled productivity knowledge: under Restrict there is no formal disclosure channel, and by Proposition 3 workers would conceal even if access were permitted (anticipating extraction upon any revelation). Hence $V_F = 0$ and network externalities are zero. Thus $\Pi_H = N\omega - \sigma$.

Step 2: Polycentric payoff. Under polycentric governance with Propositions 4 and 6 satisfied, managed access eliminates σ. High-type workers reveal AI-enabled productivity knowledge under the credible no-punishment commitment, releasing $V_F > 0$. Each of the $N \geq 1$ active contributors generates network externality $e > 0$ and bears contribution cost c_c. Individual AI output ω continues under managed access.

Thus $\Pi_P = N\omega + V_F + N \cdot e - N \cdot c_c$.

Step 3: Dominance condition.

$$\Delta\Pi = \Pi_P - \Pi_H = V_F + N \cdot e + \sigma - N \cdot c_c.$$

$\Delta\Pi > 0$ iff $V_F + \sigma > N(c_c - e)$. Two sufficient conditions: (i) $e \geq c_c$: each contributor generates weakly more externality than cost, so $\Delta\Pi \geq V_F + \sigma > 0$; (ii) $V_F + \sigma$ is large relative to $N(c_c - e)$ when $c_c > e$. The gap $\Delta\Pi$ is strictly increasing in N when $e > c_c$, and increasing in V_F, σ, and e for any N.

Implementability. For Π_P to be achievable, high-type workers must voluntarily participate. Under the screening condition, $(\theta_i - 1)\pi + k > c_c$ for each active contributor, so each participant's productivity premium strictly covers her contribution cost net of the shadow outside option, and participation is individually rational. □

Proof of Proposition 8.

Part (i): High-type workers. In the hierarchical equilibrium (Restrict, Shadow), a high-type worker receives $\pi - k$. Under polycentric governance the screening condition $(\theta_H - 1)\pi > c_c$ gives $\theta_H \pi - c_c > \pi > \pi - k$, so each active contributor strictly prefers contributor access. The strict gain is $(\theta_H - 1)\pi + k - c_c > 0$.

Part (ii): Low-type workers. A low-type worker does not contribute (by construction $(\theta_L - 1)\pi < c_c$) and takes managed access at π, which strictly exceeds the hierarchical shadow-adoption payoff $\pi - k$. The amnesty design additionally removes the disciplinary risk $d > 0$ borne under prior unauthorized use.

Combining parts. Both types are strictly better off, so polycentric governance is a strict Pareto improvement over the hierarchical equilibrium for all workers. □

Proof of the hierarchical optimization result (Section 4).

Step 1: Bounding ϕ. We show $\phi < 1$ through two independent channels. First, *genuine tacitness*: monitoring reveals output $y_i = v_i + \omega + \varepsilon_i$ (where ε_i is noise), but not the underlying productive method v_i. Even with unlimited monitoring intensity, the inference problem is unsolvable: y_i identifies the level of $v_i + \omega$ up to noise, not the method v_i that constitutes the transferable knowledge. By von Hippel (1994), task-specific AI knowledge is sticky in exactly this sense: it resides with the worker and is costly to articulate regardless of incentives. Second, *strategic suppression*: even the codifiable share of v_i would be withheld under monitoring if the firm could use it for extraction. By Proposition 3, workers conceal productive method knowledge whenever disclosure invites surplus renegotiation. Monitoring does not resolve this channel; it merely changes what the firm can observe, not the strategic incentive to conceal. Polycentric governance addresses the second channel (removing the extraction threat makes strategic suppression individually irrational) but not the first (tacit knowledge remains inaccessible regardless of governance design). The observable component of V_F is therefore $\phi \cdot V_F$ where $\phi \in [0, 1)$ is the codifiable share; the tacit share $(1 - \phi)V_F$ remains inaccessible regardless of monitoring intensity. Hence $\phi < 1$ strictly whenever V_F has a positive tacit component, which holds by the dispersed-knowledge assumption.

Step 2: Hierarchical payoff under optimization. Senior management permits access, reducing shadow use and its associated security cost to $\sigma_H \leq \sigma$. Monitoring at per-worker cost $m > 0$ captures $\phi \cdot V_F$. Individual AI output ω continues. Thus $\Pi_H^{\text{opt}} = N\omega + \phi V_F - Nm - \sigma_H$ (equation (5)).

Step 3: Dominance condition.

$$\begin{aligned}
\Delta\Pi^{\text{opt}} &= \Pi_P - \Pi_H^{\text{opt}} \\
&= \big(N\omega + V_F + N \cdot e - N \cdot c_c\big) - \big(N\omega + \phi V_F - N \cdot m - \sigma_H\big) \\
&= (1 - \phi)V_F + N \cdot e + N \cdot m - N \cdot c_c + \sigma_H.
\end{aligned}$$

$\Delta\Pi^{\text{opt}} > 0$ iff $(1-\phi)V_F + N(e + m - c_c) + \sigma_H > 0$. Since $\phi < 1$ (Step 1), $e > 0$, $m > 0$, and $\sigma_H \geq 0$, a sufficient condition is $e + m \geq c_c$. The monitoring cost m enters the condition *positively*: it represents expenditure that

polycentric governance avoids entirely through community self-governance, making polycentric governance more attractive relative to the optimized hierarchy than relative to the pathological one. Under the scope conditions of Section 6 (dispersed knowledge, large N, low adoption barrier), all four terms are strictly positive and $\Delta\Pi^{\text{opt}} > 0$. □

Proof of the governance gap amplification result (Definition 1). $G(c) = z_w^*(c) - z_w^*(c_0)$, where c_0 is fixed by IT restriction. Since c_0 does not vary with c, $\partial G/\partial c = \partial z_w^*/\partial c$. Differentiating equation (1):

$$\frac{\partial z_w^*}{\partial c} = \frac{1}{\lambda} \cdot \frac{-h\lambda/c^2}{h\lambda/c - h\ln h} = \frac{-1/c^2}{\lambda/c - \ln h} \; < \; 0,$$

where the denominator is positive because $\ln h < 0$; interiority of the solution ($z_w^* > 0$) additionally requires $h\lambda/c - h\ln h > 1$, which we assume throughout. Hence $G(c)$ is strictly decreasing in c, i.e. strictly increasing as c falls. As $c \to 0$, $z_w^*(c) \to \infty$ while $z_w^*(c_0)$ remains finite, so $G(c) \to \infty$. □

Proof of Theorem 1 (observability asymmetry). Under hierarchical governance, consider the bilateral repeated game between the firm and worker i. Let v_i denote the firm's payoff per period from worker i's cooperation alone (worker i's knowledge component revealed). Network externalities are generated across all workers jointly. Under symmetric contributors each standalone bilateral relationship has the same value v_i, and Assumption 1 gives $N \cdot v_i \leq V_F$, hence $v_i \leq V_F/N$, with strict inequality whenever $N > 1$ and the externality is strictly super-additive.

If the firm deviates against worker i in period t, it extracts τ from that worker. Because extraction is bilateral and unobservable to other workers, no collective punishment is triggered; only worker i withdraws cooperation, reducing the firm's continuation value by $v_i/(1-\delta)$ (not by the full $V_F/(1-\delta)$). The no-deviation condition for the bilateral relationship is therefore:

$$\delta \cdot \frac{v_i}{1-\delta} \; \geq \; \tau,$$

yielding $\delta_i^H = \tau/(v_i + \tau)$. Since $v_i \leq V_F/N < V_F$ and $s > 0$:

$$\delta^{**} = \frac{\tau - s}{V_F + \tau - s} < \delta^* = \frac{\tau}{V_F + \tau} \leq \frac{\tau}{V_F/N + \tau} \leq \frac{\tau}{v_i + \tau} = \delta_i^H,$$

where the first inequality uses $s > 0$ and the remainder use $V_F \geq V_F/N \geq v_i > 0$. The cooperation set $[\delta^{**}, 1)$ under polycentric governance strictly contains $[\delta^*, 1)$ (without peer sanctions) which weakly contains $[\delta_i^H, 1)$ under hierarchy. The argument depends only on the observability of extraction, not on ownership form: the bilateral threshold δ_i^H binds for any governance design in which extraction against worker i is observable only to worker i, because the punishment available is then capped at the single-relationship value $v_i \leq V_F/N$. Any design that makes extraction observable to the full contributor community achieves the lower threshold δ^{**}, regardless of whether ownership is centralized. □

Proof of Corollary 2 (hierarchical ceiling). Each part follows directly from the established propositions.

Part (i). Suppose a hierarchical mechanism $\mathscr{M}$ conditions admission on a worker's self-reported AI productivity $\hat{\pi}_i$. Under Extract (the firm's dominant strategy by Proposition 3), a worker's payoff from revealing $\hat{\pi}_i = \pi_i$ is $\pi_i - \tau - \kappa < \pi_i - k$ (since $\tau > k - \kappa$ by assumption). Hence every worker reports $\hat{\pi}_i = 0$ regardless of type, and $\mathscr{M}$ achieves no separation. The contribution threshold c_c screens on the observable action (contribute or shadow-adopt) rather than on the report $\hat{\pi}_i$, and is therefore not subject to this strategic suppression; Proposition 4 shows it achieves full separation.

Part (ii). By Theorem 1, a hierarchical non-extraction commitment requires $\delta \geq \delta_i^H > \delta^{**}$. For any $\delta \in [\delta^{**}, \delta_i^H)$, the polycentric condition $\delta \geq \delta^{**}$ is satisfied while the hierarchical condition $\delta \geq \delta_i^H$ is not.

Hence there is a non-empty set of firms for which polycentric governance achieves credible commitment and hierarchy cannot.

Part (iii). By Step 1 of the hierarchical optimization proof above, $\phi < 1$ strictly. The firm therefore captures $\phi V_F < V_F$ regardless of monitoring intensity, and $\Pi_H^{\text{opt}} = N\omega + \phi V_F - Nm - \sigma_H$. Subtracting from $\Pi_P = N\omega + V_F + N \cdot e - N \cdot c_c$ gives $\Pi_P - \Pi_H^{\text{opt}} = \Delta\Pi^{\text{opt}} > 0$ (equation (6)) under the stated scope conditions. □

Proof of Proposition 9. A committed firm (type C) that invests I and sustains polycentric governance obtains net payoff $V_F/(1-\delta) - I + \sigma$: the cooperative continuation value $V_F/(1-\delta)$, minus the investment cost I, plus the recovered security cost σ that polycentric governance eliminates. An opportunistic firm (type O) that mimics the investment obtains $\tau - I$: because $\delta_O < \delta^{**}$, its continuation value from cooperation falls short of the one-shot extraction gain, so it extracts τ after workers disclose and exits; the security cost σ is not recovered because the firm does not sustain governance.

Participation constraints. Type C invests I iff its net payoff is non-negative: $V_F/(1-\delta) - I + \sigma \geq 0$, so $I \leq V_F/(1-\delta) + \sigma$. Type O mimics iff it earns a non-negative net payoff from mimicry: $\tau - I \geq 0$, so $I \leq \tau$.

Bounding I^*. Separation requires I^* to deter type O ($I^* > \tau$, so mimicry is loss-making) while remaining acceptable to type C ($I^* \leq V_F/(1-\delta) + \sigma$). The separating range is therefore $(\tau, V_F/(1-\delta) + \sigma]$, non-empty iff $\tau < V_F/(1-\delta) + \sigma$, i.e. $V_F/(1-\delta) > \tau - \sigma$; since $\sigma > 0$ this holds a fortiori whenever $V_F/(1-\delta) \geq \tau$, i.e. $\delta \geq \tau/(V_F + \tau) = \delta^*$. At $\delta = \delta^*$: $1 - \delta^* = V_F/(V_F + \tau)$, so $V_F/(1-\delta^*) = V_F + \tau$, giving the non-degenerate range $(\tau, V_F + \tau + \sigma]$. □

Lemma 1 (Convergence of community governance)**.** *Let $\eta_H \in (0,1)$ denote the population share of high-productivity workers, and let P_t denote the equilibrium participation rate at period t. Suppose the community adjusts the contribution threshold by the adaptive rule $c_c^{t+1} = c_c^t + \zeta(P_t - \eta_H)$ with step size $\zeta > 0$. Starting from any $c_c^0 > 0$, if the step size satisfies $\zeta < (\bar{c}_H - \bar{c}_L)/(1 - \eta_H)$, the sequence $\{c_c^t\}$ enters the separating interval $[\bar{c}_L, \bar{c}_H]$ in finite steps; for larger step sizes it may overshoot and oscillate, with convergence preserved in the limit but finite-step entry not guaranteed. The rule requires the community to observe P_t but not η_H directly.*

Proof of Lemma 1. The equilibrium participation rate is a step function of c_c:

$$P(c_c) = \begin{cases} 1 & \text{if } c_c < \bar{c}_L \quad \text{(pooling zone)}, \\ \eta_H & \text{if } c_c \in [\bar{c}_L, \bar{c}_H] \quad \text{(separating zone)}, \\ 0 & \text{if } c_c > \bar{c}_H \quad \text{(exclusion zone)}. \end{cases}$$

Case 1 ($c_c^t < \bar{c}_L$): $P_t = 1 > \eta_H$, so $c_c^{t+1} = c_c^t + \zeta(1 - \eta_H) > c_c^t$; the threshold rises toward the separating interval. *Case 2* ($c_c^t > \bar{c}_H$): $P_t = 0 < \eta_H$, so $c_c^{t+1} = c_c^t - \zeta\eta_H < c_c^t$; the threshold falls toward the interval. *Case 3* $\left(c_c^t \in [\bar{c}_L, \bar{c}_H]\right)$: $P_t = \eta_H$, so $c_c^{t+1} = c_c^t$; stationary.

The per-step adjustment in Case 1 has magnitude $\zeta(1 - \eta_H)$, which is the largest step the rule can take. If $\zeta(1 - \eta_H) \leq \bar{c}_H - \bar{c}_L$, equivalently $\zeta \leq (\bar{c}_H - \bar{c}_L)/(1 - \eta_H)$, a single upward step from below cannot jump past the entire interval, so the sequence enters $[\bar{c}_L, \bar{c}_H]$ in finite steps and remains there. If the step size violates this bound, the sequence may overshoot the interval into the exclusion zone, from which Case 2 returns it downward; the sequence then oscillates around the interval. Convergence to a neighborhood of $[\bar{c}_L, \bar{c}_H]$ is preserved in the limit, but finite-step entry requires the stated step-size bound. □

Proof of Proposition 5. *Existence iff.* For $c_c \in (\bar{c}_L, \bar{c}_H)$ the participation and incentive conditions verified in the proof of Proposition 4 hold strictly, so the separating equilibrium exists. Outside the interval separation fails: for $c_c < \bar{c}_L$ both types contribute (pooling), and for $c_c > \bar{c}_H$ neither does (exclusion). At a boundary one type is exactly indifferent, so separation there turns on the tie-breaking convention; stating the interval

as open makes existence independent of any such convention. Lemma 1's closed separating zone differs only at these two points, which its dynamics never distinguish.

Directional asymmetry. Pooling ($c_c < \bar{c}_L$): both types participate; V_F falls to $V_F^{\text{pool}} \leq V_F$. $\Delta\Pi = V_F^{\text{pool}} + \sigma - N(c_c - e) > 0$ under the same condition as Proposition 7, so the governance advantage survives. *Exclusion* ($c_c > \bar{c}_H$): no worker contributes, so $V_F = 0$ and no contribution cost is paid; managed access remains universal, so σ stays eliminated and $\Pi_P^{\text{excl}} = N\omega = \Pi_I$, the intermediate baseline, which exceeds $\Pi_H = N\omega - \sigma$ but forfeits the knowledge surplus. The asymmetry is therefore directional: under-restriction (pooling) attenuates V_F but preserves the advantage, whereas over-restriction drives V_F to zero and the payoff back to Π_I, and the two boundaries differ because $\bar{c}_L < \bar{c}_H$. □

Numerical calibration (Theorem 1). The following calculations are illustrative rather than structural estimates. They trace the qualitative properties of the equilibrium at representative parameter values and verify that the polycentric advantage is not an artifact of extreme calibration choices. We treat V_F and σ here as reduced-form primitives. Remark 1 gives their structural interpretation, in which both scale with the governance gap $G(c)$, and the values chosen below correspond to an implied gap with $\gamma/\beta = V_F/\sigma = 50$. Set $N = 1{,}000$, $\pi = \$10{,}500$ (15% average effect from Brynjolfsson et al. (2025) applied to an assumed \$70,000 annual compensation), $\tau = \$5{,}000$, $\sigma = \$100{,}000$, $c_c = \$3{,}000$, $e = \$500$, $s = \$2{,}000$, and $V_F = \$5{,}000{,}000$ (above the break-even threshold $V_F^* = \$2{,}400{,}000$).
Welfare. $\Delta\Pi = \$5{,}000{,}000 + \$500{,}000 + \$100{,}000 - \$3{,}000{,}000 = \$2{,}600{,}000 > 0$.
Commitment thresholds.

$$\delta^{**} = \frac{3{,}000}{5{,}003{,}000} \approx 0.0006, \qquad \delta^{*} = \frac{5{,}000}{5{,}005{,}000} \approx 0.001, \qquad \delta_i^H = \frac{5{,}000}{10{,}000} = 0.50.$$

The gap $\delta_i^H - \delta^{**} \approx 0.50$ spans virtually all empirically plausible discount factors. Firms with a discount factor in $[0.0006, 0.50)$ can sustain polycentric commitment but not hierarchical commitment.
Sensitivity to super-additivity. The gap is governed by how far the commons exceeds its bilateral parts. Write $v_i = \mu V_F$ with $\mu \in [1/N, 1]$, so $\delta_i^H = \tau/(\mu V_F + \tau)$. The benchmark $v_i = V_F/N$ ($\mu = 1/N$, so $v_i = \$5{,}000 = \tau$) is the least favorable point for the polycentric account that Assumption 1 permits, and even there the gap is order N: $\delta_i^H = 0.50$. The gap only widens for smaller v_i (at $v_i = 0.5\tau$, $\delta_i^H \approx 0.67$; at $v_i = 0.1\tau$, $\delta_i^H \approx 0.91$), and it narrows as $\mu \to 1$, the two regimes coinciding only in the degenerate case $\mu = 1$, where a single worker's withdrawal already threatens the entire surplus and observability ceases to matter. That case is precisely the one Assumption 1 rules out for a genuine commons. The polycentric threshold δ^{**} is unaffected by v_i throughout, since collective punishment is calibrated to V_F rather than to any single relationship. The order-of-magnitude separation is therefore a property of the degree of super-additivity, not an artifact of the benchmark point.

B The BBVA Enterprise AI Program

This appendix summarizes the case we draw on throughout the paper. The account follows Alfaro et al. (2026); it is background for the reader, and the analytical mapping between the case and the model appears in the main text and in Table 4.

BBVA is one of Europe's largest banks, with about 125,000 employees in 25 countries and more than 77 million clients. Like many large organizations, it first met generative AI not as a managed program but as unsanctioned use: employees were running personal ChatGPT and similar tools for work, often outside the awareness of IT and compliance. Alfaro et al. (2026) describe this "shadow AI economy" as widespread and read it as a signal of latent demand rather than a compliance failure.

In April 2024 the bank reached an agreement with OpenAI and deployed ChatGPT Enterprise in a secured private cloud. The decision rested on a judgment that unmanaged hidden use was more dangerous than a managed alternative deployed quickly. Senior leadership compressed the usual risk-assessment, legal-review, and GDPR-compliance steps into roughly two months and assigned a dedicated senior team with authority to act, a speed made possible by data infrastructure the bank had built since 2017, with its data organization reporting to the CEO.

Scaling followed a logic of attraction rather than mandate. The bank issued only 3,000 licenses at first, allocating them to business-area leaders who passed them to their most motivated people under a "use it or lose it" rule, with under-used licenses reassigned to colleagues on a waiting list. Active use was defined by contribution as well as frequency, and workers who built and shared custom GPTs (user-configured assistants with set instructions and tools) were prioritized. Demand soon exceeded supply, and access became a privilege rather than an obligation.

The organizing structure was the Adoption Network. About 25 Champions (heads of areas and countries) and more than 100 Co-Champions handled diffusion and resource allocation, while about 200 Wizards (power users selected for skill and enthusiasm) acted as local experts and built high-value tools. A Community of Practice served as the horizontal forum where members shared tools, solved problems together, and surfaced uses developed elsewhere. Because central risk assessment of thousands of employee-built GPTs was infeasible, the bank used an automated Quality Score that evaluated each tool on guardrails, clarity, specificity, ambiguity, and context, in place of a formal admission process that would have stalled the effort. One principle anchored governance: AI is an assistant, not an autonomous agent. Outputs could not write directly to core systems without human validation, the employee owned the result, and every user had to complete training, pass an exam, and sign a statement of responsibility.

The program scaled from 3,000 to 11,000 active users in under a year, with about 80 percent of use through direct prompting and 20 percent through employee-built GPTs. More than 83 percent of license holders used the system weekly, averaging about 50 prompts per week, and reported time savings of two to five hours. More than 4,800 custom GPTs were built internally and used roughly three times as often as the enterprise average. Adoption was deepest in internal audit, where management funded full coverage and 99 percent of the bank's 600 auditors became active users. Comparable gains appeared across functions and geographies, with the tools built by frontline employees rather than by a central IT team.

Not all AI fell under this design. Regulated, high-stakes models such as credit scoring and anti-money-laundering systems stayed under centralized oversight, consistent with the bifurcated governance the model predicts for regulated outputs (Section 6).