

# Designing Enterprise AI Systems: Hallucination, Creativity, and Moral Hazard

Tinglong Dai\*

Terry Taylor†

\*Carey Business School, Johns Hopkins University, Baltimore, Maryland 21202; Data Science and AI Institute, Johns Hopkins University, Baltimore, Maryland 21218. [dai@jhu.edu](mailto:dai@jhu.edu)

†Haas School of Business, University of California, Berkeley, California 94720. [taylor@haas.berkeley.edu](mailto:taylor@haas.berkeley.edu)

---

**Abstract.** Generative AI (GenAI) is reshaping business operations, yet enterprise deployments still confront hallucinations. As organizations shift toward enterprise-controlled deployments, they can tune temperature, a hyperparameter that mediates a trade-off between creativity and reliability: higher values generate more diverse and exploratory outputs, increasing both the scope for insight and likelihood of hallucinations; lower values yield more consistent and deterministic responses, reducing variability and error risk. Greater variability makes human effort—in vetting, correcting, refining, and building on AI-generated content—more consequential for success. We study incentives when GenAI output and costly, unobservable effort jointly determine performance. In a principal–agent model, the firm chooses temperature and a success-contingent bonus. Temperature and effort are complements: higher temperature increases the marginal effect of effort on the probability of success. We show that endogenizing temperature can reverse classical results: decentralization can distort the effort and success probability *upward*. This occurs because higher temperature strengthens the link between effort and outcomes, reducing required bonuses and information rents. Despite using temperature to curb information rents, the principal’s discretion can make the agent strictly *better off* by enabling stronger incentives. Finally, decentralization alters comparative statics with respect to task stakes and effort costs. In the centralized benchmark, the optimal temperature and effort depend on task stakes and effort costs in an intuitively monotone fashion. In the decentralized setting, the interaction between system design and incentive provision can *break* this monotonicity. Together, these results shed light on how enterprise AI system design and incentive provision shape one another.

**Key words:** Generative AI, principal–agent problem, incentive design, human–AI interaction.

---

## 1. Introduction

Artificial intelligence (AI) has been around for nearly seven decades (Bringsjord and Govindarajulu 2024; Simon 1987), but its wide adoption—across virtually all industries and functional areas—is relatively recent. Among the various forms of AI, generative AI (GenAI) systems have not only captured outsized public attention but have also seen substantial real-world use. By late 2025, nearly 90% of organizations had reported integrating GenAI into at least one business function (McKinsey & Company 2025). The promise of GenAI to reimagine business operations is compelling, yet it comes

with a host of new challenges. One immediate challenge is organizational: firms must determine how to use GenAI without exposing proprietary or regulated information.

This concern has been salient since the start of the current GenAI wave. The wave began in November 2022 with the release of OpenAI's ChatGPT, and it quickly raised questions about whether non-enterprise AI tools could be used safely inside organizations. By early 2023, concerns about data breaches and security risks had led many organizations to restrict employees' use of public GenAI tools for work (Abril 2023; DeRose 2023): Samsung, for example, imposed a company-wide ban after an engineer inadvertently uploaded sensitive internal source code to ChatGPT, raising fears that proprietary data could be retained on external servers or exposed beyond the firm (Ray 2023). Amazon likewise warned employees not to share confidential information with public GenAI tools after observing outputs that appeared to mirror internal proprietary material. Financial institutions, including Bank of America, Goldman Sachs, and JPMorgan Chase, adopted similar restrictions, citing regulatory and security concerns. In healthcare, many organizations also prohibited the use of public GenAI tools, pointing to potential noncompliance with the Health Insurance Portability and Accountability Act (HIPAA), which governs the handling of electronic protected health information. Public versions of tools such as ChatGPT, Claude, and Gemini generally do not satisfy these requirements (Alder 2024).

Beginning in late 2023, however, outright bans increasingly gave way to a more strategic embrace of GenAI. Organizations began investing in secure, enterprise-controlled deployments that could be integrated into internal workflows while preserving control over data, access, and compliance (Korn 2023). This shift reflects a broader recognition that the relevant question is not whether GenAI will be used, but where it will sit inside the firm's production process and how it will be governed. Enterprise systems are therefore designed to operate behind corporate firewalls, incorporate firm-specific knowledge bases, enforce permissioning and auditing, and satisfy sector-specific regulatory constraints. Walmart, for instance, introduced "My Assistant," an internal tool designed to support routine tasks and idea generation. Consulting firms such as McKinsey, PwC, and EY similarly developed proprietary GenAI systems—often via internal, private deployments—to raise productivity while maintaining tighter control over data. Academic medical centers have moved in the same direction, rolling out HIPAA-compliant enterprise GenAI platforms that can be used with sensitive health information. In short, enterprise systems aim to mitigate key vulnerabilities while still capturing the gains from GenAI.

This shift toward proprietary, enterprise-controlled deployments gives organizations the opportunity to select key system hyperparameters that shape employees' productive environment. One such hyperparameter is *temperature*, a setting available through large language models' (LLMs') application programming interfaces (APIs) that governs generative behavior. Temperature mediates a trade-off between creativity and reliability. Higher values generate more diverse and exploratory outputs,

increasing both the scope for insight and the likelihood of hallucinations; lower values yield more consistent and deterministic responses, reducing variability and the risk of error. Despite rapid advances in LLM capabilities, hallucinations remain a central concern in enterprise deployments (Cummings 2025). Moreover, they reflect fundamental features of current LLM architectures and are unlikely to be eliminated in the near term (Rothman 2023). This concern is especially salient in regulated, high-stakes domains such as healthcare and finance, where firms often mandate low temperature settings to support compliance and reliability (Mantel Group 2024; McKinsey 2023). For example, C3.ai configures its enterprise AI systems to meet domain-specific operational requirements. In a recent earnings call, the company’s CEO noted that for mission-critical applications, the firm sets temperature low explicitly to suppress hallucinations (C3.ai 2024).

Across enterprise GenAI applications, the interaction between system design and human effort is first order. Output quality depends on both the user’s effort and the system’s configuration. In particular, temperature governs how strongly effort translates into success. Empirical work shows that increasing temperature widens the dispersion of LLM outputs in clinical tasks (Jarrett et al. 2025): when temperature is high, the system generates a broader range of outputs, including both valuable possibilities and errors requiring human correction. This variability makes the agent’s effort—in vetting, refining, and building on AI-generated content—more consequential for ultimate success. In our framework, this effect is captured parsimoniously by modeling temperature and effort as complements in the success probability: success becomes more sensitive to human effort at higher temperature. This complementarity is not specific to clinical tasks; it arises when GenAI produces candidate outputs that agents must vet and refine. Crucially, the choice of temperature is therefore not merely an engineering decision; because effort-exerting human agents interact with the technology, the design of the AI system and the provision of incentives are fundamentally intertwined.

The same logic appears in multiple contexts. In wealth management and related financial services, advisors increasingly use GenAI as a co-production tool for portfolio recommendations (Mehta et al. 2024). Temperature governs how diverse the system’s candidate suggestions are, while the advisor’s effort governs how effectively those suggestions are vetted, corrected, and tailored to the client before any recommendation is delivered. When temperature is low, the system tends to generate conservative and homogeneous candidate allocations; when temperature is high, it generates a broader set of candidates that can include both valuable ideas and problematic elements. In turn, the marginal impact of advisor effort on the quality (and perceived success) of the final recommendation is larger at higher temperature. A parallel pattern appears in large contact centers, where agents and GenAI jointly produce communications to customers (Allon 2024; Brynjolfsson et al. 2025). Assistant tools surface candidate language and relevant supporting materials, while the agent decides what to use and how to revise it before anything reaches the customer; as in wealth management, higher temperature

settings generate more creative but potentially inappropriate responses, making the agent’s effort more consequential for the quality of customer interactions.

Motivated by enterprise settings in which AI system design and agency considerations are intertwined, we develop a parsimonious principal–agent model of an AI-augmented productive environment. A principal configures an enterprise GenAI system by specifying its temperature and offers a success-contingent bonus, whereas an agent privately chooses effort. Temperature shapes how strongly effort translates into performance. Effort and temperature are complements, so that higher temperature makes incremental human effort more effective at converting AI-assisted work into success. The framework then addresses four interlinked questions about enterprise AI design under moral hazard: how discretion over temperature affects (*i*) incentive distortions and effort provision, (*ii*) the choice of the temperature itself, (*iii*) the welfare of the agent, and (*iv*) the way optimal temperature and effort adjust to changes in task stakes or effort costs. Throughout, “centralized” (“first-best”) denotes the full-information benchmark in which the principal observes and contracts on effort. “Decentralized” (“second-best”) denotes the incentive-constrained benchmark in which effort is hidden and non-contractible.

First, how does the principal’s discretion over the system hyperparameter of temperature alter the incentive effects of decentralization? In the classical moral-hazard setting—where the production technology is fixed from the principal’s perspective—decentralization is known to distort effort and the probability of success downward relative to the centralized benchmark. In contrast, when the principal–manager can choose the AI system’s temperature, this conclusion can reverse: decentralization may distort effort and the success probability *upward*. The mechanism is that temperature affects not only expected performance but also the informativeness of outcomes about unobserved effort. By raising temperature, the principal can make outcomes more sensitive to effort, strengthening incentives precisely in environments where standard moral-hazard logic predicts attenuation.

Second, how does decentralization shape the principal’s choice of temperature? Our analysis shows that decentralization systematically distorts the principal’s system-design choice. Holding fixed the effort level the principal seeks to implement, decentralization weakly distorts temperature *upward*. Intuitively, temperature becomes an instrument for limiting information rents: by increasing the sensitivity of success to effort, the principal can reduce the bonus required to induce a given effort level, making higher temperature relatively more attractive under decentralization than under the centralized benchmark.

Third, given that the principal uses its discretion over temperature to limit the agent’s information rents, can the agent ever benefit from the principal’s discretion over temperature? Our analysis reveals discretion over temperature need not operate solely against the agent. Under some parameter values, the agent can be strictly *better off* when the principal has discretion over temperature. Although

the principal uses temperature strategically to reduce incentive costs, the resulting equilibrium can expand surplus enough that the agent's expected payoff rises, because higher temperature and stronger incentives increase the likelihood of success and, in turn, expected compensation.

Fourth, as the task stakes (i.e., the magnitude of the reward for a successful outcome) rise or effort becomes more costly, how do the principal's optimal temperature and the induced effort respond, and how do these comparative statics differ from those in the centralized benchmark? In the centralized system, the optimal temperature and effort are monotone in the task stakes, as one might expect. In contrast, in the decentralized system, the principal's optimal temperature and effort can be *nonmonotone* in the task stakes. As stakes rise, the principal values success more, but the cost of eliciting effort also changes because temperature and incentives move together. The optimal response can therefore involve raising temperature and effort over some ranges and lowering one or both over others. Parallel remarks apply to the impact of the costliness of effort.

Our primary contribution is to show that endogenizing temperature can reverse classical predictions about decentralization: effort and success probability can be distorted upward, and task-stakes comparative statics can become non-monotonic, in ways that do not arise when temperature is exogenous. These departures stem from a common mechanism: temperature governs how informative outcomes are about hidden effort, giving the principal an instrument to reduce agency costs beyond what is possible when technology is exogenous. Thus, temperature is not only a technological design choice that governs the creativity–reliability trade-off in AI-assisted work; it also shapes the marginal cost of eliciting human effort and therefore the structure of incentives inside the firm. Together, these results shed light on the joint design of enterprise GenAI configuration and incentives.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 presents the model and benchmarks. Section 4 analyzes how decentralization shapes optimal temperature, induced effort, and success probability. Section 5 studies how task stakes and effort costs shape AI system configuration and incentive provision. Section 6 examines how the firm's discretion over system design affects the agent's utility. Section 7 concludes. The appendices provide proofs and supplemental materials.

## 2. Literature

Our paper builds on and contributes to three streams of literature: (1) operations management with moral hazard, (2) economics and operations of AI, and (3) human–AI interaction.

First, our model builds on the canonical moral hazard framework (Grossman and Hart 1983; Holmstrom 1979; Laffont and Martimort 2002), in which a principal contracts with an agent whose effort is unobservable but affects a payoff-relevant outcome. Departing from the canonical moral hazard framework, we allow the principal to make a system-level design decision—specifically, setting

the generative AI system’s temperature—prior to the agent’s choice of effort. This feature connects our model to a strand of literature in economics (Bester and Krähmer 2008; Demougin and Fluet 2001; Kirkegaard 2022) and operations management (Alp and Sen 2021; Baiman et al. 2010; Dai et al. 2021; de Véricourt and Gromb 2018, 2019; Li et al. 2020; Long and Nasiry 2020; Nikoofal and Gümüş 2018; Plambeck and Taylor 2006), in which the principal endogenously specifies aspects of the operating environment—such as monitoring intensity, task complexity, or operational constraints—to shape agent incentives. These studies emphasize that environment design is not merely a backdrop to the incentive problem but a central part of it. For instance, Demougin and Fluet (2001) analyze the optimal mix of monitoring and performance-based pay under moral hazard with limited liability, and show how this mix varies with monitoring costs and the agent’s liability limit. More recent work pushes this logic further by treating system features as active design levers. For example, de Véricourt and Gurkan (2023) and Gurkan and de Véricourt (2022) show how attributes such as algorithmic complexity or data richness can change the cost-effectiveness of inducing effort. In the GenAI setting, a system-wide attribute is the model’s temperature. Temperature affects the dispersion of generated outputs and, with it, the informativeness of what the principal can observe. In our model, temperature is a hyperparameter that modulates the effective observability of effort and thereby the cost of providing incentives. Put differently, system design enters the incentive problem directly, and the principal’s choice of temperature shapes optimal effort inducement.

Our framework highlights a channel through which system design affects incentive provision. By adjusting temperature, the principal not only modifies how effort maps into success but also alters the informativeness of observed outcomes. This mechanism parallels prior work showing that the structure of the monitoring environment can influence agent behavior (Dye 1986; Bester and Krähmer 2008). While classical moral hazard models typically assume diminishing marginal returns to effort, our setting introduces a countervailing force: higher temperature makes outcomes more sensitive to—and therefore more informative of—effort. This channel is consistent with a broader literature emphasizing the role of information systems in mitigating agency frictions. As Holmstrom (1979) and Diamond (1984) argue, improvements in the granularity of outcome signals—whether through monitoring or technology—enable more efficient incentive schemes. Similarly, Jensen and Meckling (1976) suggest that enhanced monitoring can raise the value of high effort by tightening the link between actions and outcomes. Temperature plays an analogous role by amplifying performance dispersion and improving the signal-to-noise ratio of observed outcomes. Our paper also connects to the operations literature on incentive design, which studies how production features such as inventory constraints, quality thresholds, and workflow structure shape contract design (Atasu et al. 2021; Balachandran and Radhakrishnan 2005; Ke and Ryan 2018; Plambeck and Zenios 2000, 2003; Plambeck and Taylor 2006; Serpa and Krishnan 2017; Song et al. 2024; Yu and Kong 2020). We contribute to this line of

work by treating an AI system’s configuration as a strategic design variable within a principal–agent relationship.

Our model also yields results that depart from standard predictions. First, while most principal–agent models predict downward distortions in effort under decentralization, we show that—under certain conditions—the firm may optimally induce *higher* effort in a decentralized setting than in a centralized one. This occurs despite the fact that, in our model, all effort levels are theoretically *implementable*. By contrast, in much of the literature, upward distortions arise only because certain interior effort levels are not incentive compatible (Laffont and Martimort 2002). Second, while one might expect a higher system-wide reward for success to lead the firm to elicit greater effort, we identify cases in which a higher reward prompts the principal to induce *lower* effort—a result, to our knowledge, not previously documented in the one-agent setting. Third, we show that the principal may endogenously choose to degrade system performance—for example, by raising temperature to increase output dispersion and hallucination risk—to reduce incentive costs. This underscores a core trade-off in AI-augmented work systems: firms may sacrifice output quality to obtain more informative signals. Together, these findings highlight how incentive design and system configuration are jointly determined, and how moral hazard interacts with machine-level parameters.

Second, our paper contributes to the literature on the economics and operations of AI. Gurkan and de Véricourt (2022) examine the AI flywheel effect, where larger training datasets improve predictive accuracy, leading to broader adoption and further data generation. They show that firms tend to oversupply data relative to the social optimum, enhancing monitoring of developer effort and reinforcing product improvement cycles. Dai and Singh (2020, 2025) study how AI adoption shapes physicians’ professional reputation and malpractice liability, while Agrawal et al. (2024) highlight the need for system-wide organizational adaptation to integrate AI effectively. More broadly, Acemoglu and Restrepo (2018) and Autor et al. (2003) examine how technology reshapes labor markets and organizational structures, emphasizing that AI’s impact depends on incentive alignment and skill complementarities. Our paper advances this literature by analyzing how enterprise AI tools interact with incentive structures, a topic that, to our knowledge, remains largely unexplored. Recent industry-facing work (e.g., McKinsey & Company 2023; C3.ai 2024) documents how firms increasingly treat model parameters such as temperature as levers to balance creativity, compliance, and productivity. While these insights have yet to be fully formalized in economic models, they point to the relevance of our framework for practical design decisions.

Third, by examining an enterprise setting in which GenAI tools and human decision-makers jointly produce outcomes, our work contributes to research on human–AI interaction. de Véricourt and Gurkan (2023) model a decision-maker who learns about AI accuracy through iterative interaction. Grand-Clément and Pauphilet (2023), in a continuous prediction framework, argue that AI systems

should account for user adherence behavior rather than optimize purely for predictive accuracy. Recent studies emphasize the complementarity between AI and human judgment in improving decision-making, particularly in healthcare (Mullainathan and Obermeyer 2021; Orfanoudaki et al. 2022) and retail (Karlinsky-Shichor and Netzer 2024). This aligns with Acemoglu and Restrepo (2018) and Autor et al. (2003), who highlight AI’s dual role in augmenting and displacing human decision-making, depending on the degree of complementarity between AI and human expertise. One mechanism for synergy is that AI enhances predictive accuracy while humans provide critical interpretation and contextualization (Agrawal et al. 2018). Boyaci et al. (2024) show accuracy gains from human–AI collaboration, albeit with cognitive costs. Other studies suggest that human oversight in AI-assisted decision-making improves outcomes and remains important given evidence that LLMs can exhibit human-like decision biases (Chen et al. 2023, 2025; Kim et al. 2024). Relatedly, Bastani and Cachon (2025) study how the need for costly inspection and rework of AI output shapes compensation and adoption. Lu and Tomlin (2025) show that firms may optimally limit the disclosure of a prediction machine’s signal when disclosure discourages managerial effort. To our knowledge, however, the literature has yet to explore how incentive design shapes hyperparameter choices in GenAI settings. In this regard, our paper foregrounds the trade-off between hallucination, creativity, and moral hazard in AI-augmented decision environments.

### 3. Model

We consider a principal–agent model inspired by the setting in which a GenAI system and a human agent jointly produce a final outcome. Our model introduces two key features specific to GenAI: the possibility of hallucination and the ability to tune a system hyperparameter known as “temperature.” A higher temperature makes the AI’s output more diverse and exploratory, increasing the scope for value but also the likelihood of hallucinations. The human agent reviews the AI output and exerts effort to filter, refine, correct and/or build on it. Effort improves performance, and it is especially productive when the AI output is more variable; that is, the marginal return to effort increases with temperature. As is standard in the principal–agent literature, effort is costly and unobservable to the principal.

For tractability, we assume that the outcome of the human–AI collaboration is binary: either a success or a failure. Let  $\tau \in \mathcal{T}$  denote the temperature chosen by the firm, and let  $e \in \mathcal{E}$  denote the agent’s privately chosen effort level, where  $(\mathcal{E}, \mathcal{T}) \subseteq [0, \infty)^2$ . The probability of success  $p(e, \tau)$  is increasing and concave in effort  $e$ . The agent incurs a cost  $ke$ , where  $k > 0$  reflects the *costliness of effort*. We normalize  $k = 1$  without loss of generality. For simplicity in exposition, we suppose that the agent has the option to take a zero-cost action, formally,  $0 \in \mathcal{E}$ . Unless stated otherwise, all

monotonicity and curvature claims (increasing/decreasing, concave/convex) are strict. When  $\mathcal{E}$  or  $\mathcal{T}$  is continuous, we assume  $p(e, \tau)$  is differentiable in the relevant argument(s).

We impose a strict supermodularity assumption on  $p(e, \tau)$ . Specifically, when  $e$  and  $\tau$  are both continuous, this means

$$\frac{\partial^2 p(e, \tau)}{\partial e \partial \tau} > 0 \quad (1)$$

and, when  $\tau$  is continuous,

$$\frac{\partial p(\bar{e}, \tau)}{\partial \tau} > \frac{\partial p(\underline{e}, \tau)}{\partial \tau}, \quad (2)$$

for any  $\bar{e} > \underline{e}$ . In more general settings, this assumption implies that  $p(e, \tau)$  exhibits strictly increasing differences:

$$p(\bar{e}, \bar{\tau}) - p(\underline{e}, \bar{\tau}) > p(\bar{e}, \underline{\tau}) - p(\underline{e}, \underline{\tau}) \quad (3)$$

for any  $\bar{e} > \underline{e}$  and  $\bar{\tau} > \underline{\tau}$ . In words, temperature  $\tau$  and agent effort  $e$  are strict complements. When temperature is high, the GenAI system produces a wide range of outputs, including both novel insights and hallucinations. In this case, success depends heavily on the agent's effort: high effort corrects errors and leverages valuable ideas, while low effort results in failure. When temperature is low, output is more deterministic but less rich, and success becomes less sensitive to effort; see Figure 1 later in the paper for an illustration of the supermodularity assumption. (In principle, temperature and effort could be substitutes. We discuss how our results change under this alternative setup at the end of Section 6.)

Our information structure abstracts from direct monitoring of either the GenAI output or the agent's effort. In many enterprise deployments, the principal (e.g., the firm) observes only the final customer-facing outcome and aggregate performance metrics, not the intermediate drafts produced by GenAI nor the detailed effort exerted by a given agent. Intermediate outputs may be non-verifiable or proprietary to the agent's workflow, and agents often commit effort to filtering or refining AI-generated content before its quality can be assessed independently. To highlight the core incentive forces, we analyze the limiting case in which all information about the production process is private to the agent (our decentralized setting). At the opposite extreme, if the principal can observe and contract on the agent's effort directly, the problem reduces to the centralized benchmark (see Section 3.2). Settings with partial monitoring lie between these two poles; analyzing such intermediate information structures is an interesting direction for future research but lies beyond our scope.

### 3.1. A Microfoundation for the Success Probability Function

We now present a microfoundation for the success probability function  $p(e, \tau)$ , consistent with the assumptions that  $p(e, \tau)$  is strictly supermodular, meaning that it satisfies eqs. (1) to (3), and that it is concave and increasing in effort  $e$ . Since our results rely only on these properties rather than on

any specific microfoundation, readers who find the assumptions plausible may skip the remainder of this section without loss of continuity. There are many ways one could conceive the details of how temperature and effort interact to produce a success probability possessing these properties. We present one such way, first with a general framework and then with two special cases.

We model the production process as a joint endeavor between the GenAI system and the human agent: the system generates content, and the agent expends effort to filter, refine, correct, and build on this output; and only the final outcome (success or failure) is contractible. The sequence of these actions is immaterial for our results; what matters is that system design and agent effort jointly determine success. For the general framework, suppose  $\mathcal{E} = [\underline{e}, 1]$  and  $\mathcal{T} = [\underline{\tau}, 1]$ , where  $(\underline{e}, \underline{\tau}) \in [0, 1]^2$ . The probability of a hallucination  $h(\tau)$  increases in the temperature  $\tau$ . The agent identifies and corrects a hallucination with probability  $g(e)$ , which increases in the agent's effort  $e$ . If a hallucination is not identified and corrected, then the outcome is a failure; otherwise, the outcome is a success with probability  $f(e, \tau)$ , which weakly increases in  $e$ . Hence, the success probability

$$p(e, \tau) = [1 - h(\tau) + g(e)h(\tau)]f(e, \tau). \quad (4)$$

In the first special case,  $g(e)$  is concave in  $e$  and  $f(e, \tau)$  is invariant to  $e$ . This corresponds to the case where agent effort is devoted exclusively to identifying and correcting hallucinations. For this special case, the success probability  $p(e, \tau)$  is strictly supermodular and concave, increasing in effort  $e$ . If, in addition,  $h(\tau) = \tau$  and  $f(e, \tau)$  is concave in  $\tau$ , then  $p(e, \tau)$  is concave in  $\tau$ .

In the second special case,  $g(e) = e$ ,  $h(\tau) = \tau$  and  $f(e, \tau) = 1 - s/(e\tau)$ , where  $s \in (0, e^2\tau^2)$ . The functional form  $f(e, \tau)$  is consistent with the following: The raw output of the enterprise AI system  $r$  is uniformly distributed on  $[0, \tau]$ . As in the general framework, a hallucination occurs with probability  $h(\tau)$  and is identified and corrected with probability  $g(e)$ . The agent's effort transforms the raw output to  $re$  if there is no uncorrected hallucination and to 0 otherwise. The outcome is successful if the transformed output exceeds the threshold  $s$ . For this special case, the success probability  $p(e, \tau)$  is strictly supermodular and concave, increasing in effort  $e$ . Further,  $p(e, \tau)$  is concave in  $\tau$ .

### 3.2. Centralized and Decentralized Settings

Both parties are risk neutral, and the agent has limited liability.<sup>1</sup> We analyze both a centralized (first-best) benchmark, in which the firm can dictate effort, and a decentralized setting in which effort must be elicited via incentives.

<sup>1</sup> Following Laffont and Martimort (2002, p. 120), a limited-liability constraint on rents “plays a similar role as risk aversion,” and can be interpreted as if the agent were infinitely risk-averse below a wealth bound; see their discussion in the adverse-selection environment. We adopt the standard risk-neutral-with-limited-liability formulation here.

**Centralized Benchmark.** In the centralized benchmark, the firm (principal) selects both effort and temperature directly. The objective is to maximize expected surplus:

$$\Pi(e, \tau) = p(e, \tau) \cdot R - e, \quad (5)$$

where  $R$  is the reward the firm receives upon a successful outcome. The reward for failure is normalized to zero. Let  $(e^{FB}, \tau^{FB})$  denote the first-best (centralized, full-information) choices of effort and temperature that maximize eq. (5). Let  $e^{FB}(\tau)$  denote the optimal effort given temperature  $\tau$ , and let  $\tau^{FB}(e)$  denote the optimal temperature given effort  $e$ .

**Decentralized Setting.** The principal chooses temperature  $\tau$  and a success-contingent bonus  $b \geq 0$ . The agent then privately chooses effort  $e$ . Under  $(e, \tau, b)$ , the principal's expected utility is

$$\pi(e, \tau, b) = p(e, \tau) (R - b), \quad (6)$$

and the agent's expected utility is

$$u(e, \tau, b) = p(e, \tau) b - e. \quad (7)$$

The principal chooses  $(\tau, b)$  anticipating the agent's best response in effort. Equivalently, we write the problem as a choice over  $(e, \tau, b)$  subject to incentive compatibility and participation:

$$\max_{e \in \mathcal{E}, \tau \in \mathcal{T}, b \geq 0} \pi(e, \tau, b) \quad (8)$$

$$\text{s.t. } u(e, \tau, b) \geq u(e', \tau, b) \quad \forall e' \in \mathcal{E} \quad (9)$$

$$u(e, \tau, b) \geq 0. \quad (10)$$

Constraint (9) requires that  $e$  be optimal for the agent given  $(\tau, b)$ , and (10) ensures participation.

Let  $(e^*, \tau^*, b^*)$  denote the solution to the principal's problem. We refer to  $(e^*, \tau^*)$  as the second-best (decentralized, incentive-constrained) effort and temperature. To clarify comparative statics and benchmarks, we define the following auxiliary objects. Let  $(e^*(\tau), b^*(\tau))$  denote the optimal effort and bonus when the principal takes temperature  $\tau$  as given, that is, the solution to (8)–(10) conditional on  $\tau$ . Similarly, let  $(\tau^*(e), b^*(e))$  denote the optimal temperature and bonus when the principal fixes effort at  $e$ .

We assume that  $\tau^{FB}(e)$  and  $\tau^*(e)$  are uniquely defined. A sufficient condition for  $\tau^{FB}(e)$  to be unique is that  $p(e, \tau)$  is concave in  $\tau$ . Sufficient conditions for uniqueness of  $\tau^*(e)$  are provided in Appendix D.

## 4. Impact of Decentralization

In this section, we study the impact of decentralization. We compare the decentralized outcome, where the principal delegates to a self-interested agent who privately chooses effort, to the centralized benchmark, where effort is observable and contractible. Thus, comparing decentralized and centralized systems isolates the implications of non-contractible effort.

In the classical principal–agent moral hazard setting, the agent’s effort exhibits diminishing returns (captured in our setting by the success probability  $p(e, \tau)$  being concave in effort  $e$ ) and the temperature  $\tau$  is exogenous. A celebrated result in that classical setting is that decentralization distorts downward the effort

$$e^*(\tau) \leq e^{FB}(\tau)$$

and the success probability

$$p(e^*(\tau), \tau) \leq p(e^{FB}(\tau), \tau)$$

(see Lemma A3 in Appendix A). Moreover, when temperature is exogenous, the implemented effort is weakly increasing in the reward for success  $R$  in both the decentralized problem and the centralized benchmark (see Lemma A4(a)–(b) in Appendix A).

Below, Proposition 1A shows that when the principal is able to tune temperature  $\tau$ , the classical downward-distortion result can be reversed.

**PROPOSITION 1A.** *There exist scenarios in which decentralization distorts upward the effort  $e^* > e^{FB}$ , temperature  $\tau^* > \tau^{FB}$ , and success probability  $p(e^*, \tau^*) > p(e^{FB}, \tau^{FB})$ .*

Next, to allow for the development of sufficient conditions for upward distortion that are simple and easy to interpret, Proposition 1B considers the case where the action spaces for temperature and effort are discrete. Formally,  $\mathcal{T} = \{\tau_L, \tau_H\}$ , where  $\tau_L < \tau_H$ , and  $\mathcal{E} = \{e_L, e_M, e_H\}$ , where  $e_L < e_M < e_H$ . We say that *effort is evenly spaced* if  $e_H - e_M = e_M - e_L$ . Our assumption that the agent has a zero-cost action implies  $e_L = 0$ . Proposition 1B provides conditions under which decentralization distorts effort, temperature, and success probability upward. The proof verifies that the success probabilities characterized in eqs. (11) and (12) satisfy the assumptions in Section 3.

**PROPOSITION 1B.** *Suppose the action spaces for temperature and effort are discrete,  $(\mathcal{T}, \mathcal{E}) = (\{\tau_L, \tau_H\}, \{e_L, e_M, e_H\})$ , effort is evenly spaced, and the success probabilities are*

$$p(e_L, \tau_L) = \alpha, \quad p(e_M, \tau_L) = \alpha + \Delta, \quad p(e_H, \tau_L) = \alpha + \Delta + \varepsilon, \quad (11)$$

$$p(e_L, \tau_H) = 0, \quad p(e_M, \tau_H) = \Delta + \varepsilon, \quad p(e_H, \tau_H) = 2\Delta, \quad (12)$$

where

$$0 < \alpha < \Delta, \quad (13)$$

$$\frac{2}{2\Delta - \alpha} < \frac{R}{e_M} < \frac{1}{\Delta - \alpha}. \quad (14)$$

There exists  $\bar{\varepsilon} > 0$  such that if  $\varepsilon \in (0, \bar{\varepsilon})$ , then decentralization distorts upward the effort  $e^* > e^{FB}$ , temperature  $\tau^* > \tau^{FB}$ , and success probability  $p(e^*, \tau^*) > p(e^{FB}, \tau^{FB})$ .

The sufficient conditions for upward distortion in eqs. (11) to (14) can be interpreted as follows:

(i) the reward for success  $R$  is moderate, (ii) the success probability  $p(e, \tau)$  exhibits a strong complementarity between effort and temperature at high levels of effort, (iii) the upper envelope of the success probability  $\max_{\tau \in \mathcal{T}} p(e, \tau)$  exhibits diminishing returns in effort, and (iv) the distinct-temperature success probability curves cross once. Next, we consider each condition sequentially, explaining its role in contributing to upward distortion. We begin by explaining why decentralization distorts effort upward  $e^* = e_H > e_M = e^{FB}$  and temperature upward  $\tau^* = \tau_H > \tau_L = \tau^{FB}$ . The intuition is most transparent when  $\varepsilon$  is small; throughout we assume  $\varepsilon < \Delta/2$ . Figure 1 illustrates Proposition 1B.

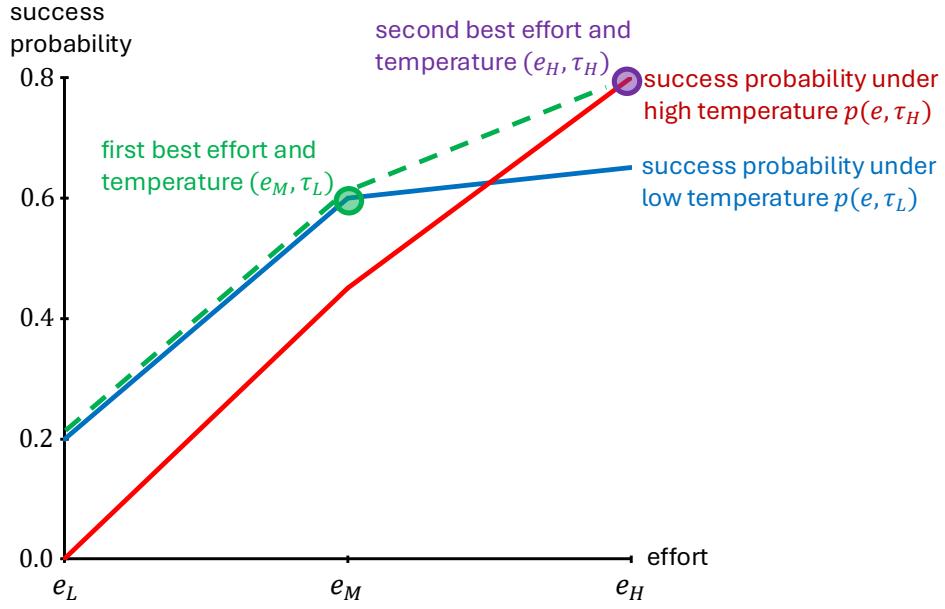
A necessary condition for an upward distortion with discrete effort is that the success reward  $R$ , relative to effort costs, lies in an intermediate range; see (14). If  $R$  is sufficiently large (respectively, small), then both the centralized and decentralized systems select the highest (respectively, lowest) effort level. The left inequality in (14) ensures that  $R$  is large enough that the optimum effort under either regime is in  $\{e_M, e_H\}$ . Moreover, as  $\alpha \uparrow \Delta$ , the upper bound in (14) diverges, so the restriction becomes mild: upward distortion arises for any sufficiently large reward  $R$ .

As illustrated in Figure 1, the success probability  $p(e, \tau)$  displays strong effort–temperature complementarity at higher effort. In particular, the gain from raising effort from moderate to high is much smaller under low temperature than under high temperature:

$$p(e_H, \tau_L) - p(e_M, \tau_L) < p(e_H, \tau_H) - p(e_M, \tau_H).$$

Consequently, when temperature is high, the minimal bonus required to induce high effort is not too large  $b = e_M/(\Delta - \varepsilon)$ . This moderate bonus, which induces high effort under high temperature, induces only moderate effort under low temperature (formally, when temperature is low, the minimal bonus required to induce high effort is the larger  $b = e_M/\varepsilon$ ). The underlying logic reflects the informativeness principle (Holmstrom 1979): high temperature amplifies the dependence of success on effort, making the outcome more informative about effort. Greater informativeness reduces the bonus required to implement high effort and thus lowers the information rent. That is, higher temperature lowers the cost of motivating high effort. It is optimal for the principal to choose high temperature  $\tau^* = \tau_H$  and induce high effort  $e^* = e_H$  because doing so yields a high success probability with only a moderate cost to induce the effort.

The information rents that shape the principal’s choice of temperature are absent under centralization. Fix an effort level  $e$ . The centralized system then chooses  $\tau \in \mathcal{T}$  to maximize the success



**Figure 1** Parameters are  $\alpha = 0.20$ ,  $\Delta = 0.40$ ,  $\varepsilon = 0.05$ , and  $R = 4.8$ . The success probability  $p(e, \tau)$  is plotted against effort levels  $(e_L, e_M, e_H) = (0, 1, 2)$  under low and high temperatures  $\tau_L$  and  $\tau_H$ . Under centralized decision-making, the first-best outcome is  $(e_M, \tau_L)$ , yielding a success probability of  $p(e_M, \tau_L) = \alpha + \Delta = 0.60$  (green point on the low-temperature curve). Under decentralization, the equilibrium shifts to  $(e_H, \tau_H)$  with  $p(e_H, \tau_H) = 2\Delta = 0.80$  (purple point on the high-temperature curve), thus distorting both effort and success probability upward.

probability, so its attainable performance is  $\max_{\tau \in \mathcal{T}} p(e, \tau)$ , the upper envelope of  $\{p(e, \tau) : \tau \in \mathcal{T}\}$ , shown by the dashed green curve in Figure 1. In contrast to the nearly constant returns to effort under high temperature that prompt the principal to choose high effort  $e^* = e_H$ , the upper envelope faced by the centralized system exhibits diminishing returns, prompting the centralized system to choose moderate effort  $e^{FB} = e_M$ . Moreover, at  $e_M$  the success probability is higher under low temperature, so the centralized system sets  $\tau^{FB} = \tau_L$ . By contrast, the decentralized system selects  $(e^*, \tau^*) = (e_H, \tau_H)$ . Therefore, relative to decentralization, the centralized benchmark features lower effort and lower temperature:  $e^{FB} = e_M < e_H = e^*$  and  $\tau^{FB} = \tau_L < \tau_H = \tau^*$ . Finally, because  $p(e, \tau_H)$  and  $p(e, \tau_L)$  cross once, high temperature dominates at high effort. Hence the probability of success is higher under decentralization:  $p(e^*, \tau^*) = p(e_H, \tau_H) > p(e_M, \tau_L) = p(e^{FB}, \tau^{FB})$ .

Although the precise conditions in Proposition 1B are stylized, the discussion above highlights the key qualitative features they embody. This suggests that in more complex settings (e.g., where the effort and temperature action spaces are more expansive), the general conditions (i)-(iv) may be sufficient for upward distortion in effort, temperature, and success probability. The upward distortion result of Propositions 1A and 1B is not driven by the restriction that the effort and temperature action spaces  $(\mathcal{E}, \mathcal{T})$  are discrete. Appendix E provides an example with continuous action spaces in

which decentralization distorts upward the effort, temperature, and success probability.

**REMARK 1.** Proposition 1B provides sufficient conditions under which decentralization distorts effort, temperature, and success probability upward. However, there are also parameter values for which the classical downward-distortion logic prevails, and all three objects move in the opposite direction. For example, suppose effort is evenly spaced with  $e_H = 2$  and the reward  $R = 6.5$ . Let the success probability  $p(e, \tau)$  be given by eqs. (11) and (12) with  $\alpha = 0.25$ ,  $\Delta = 0.45$ , and  $\varepsilon = 0.15$ . Then decentralization distorts effort downward ( $e^* = e_M < e_H = e^{FB}$ ), temperature downward ( $\tau^* = \tau_L < \tau_H = \tau^{FB}$ ), and success probability downward ( $p^* = 0.70 < 0.90 = p^{FB}$ ). Appendix E provides a parallel example with continuous action spaces in which decentralization likewise distorts effort, temperature, and success probability downward. In that example, the principal optimally lowers temperature to raise the success probability, and the resulting reduction in effort–temperature complementarity weakens incentives sufficiently that induced effort falls.

**REMARK 2.** In some sense, the system hyperparameter of temperature can be interpreted as defining the *production technology* the agent uses. We say that a production technology is *inferior* if the probability of success under that technology is lower for every effort level. Formally, the production technology under temperature  $\tau_B$  is inferior to that under  $\tau_A$  if and only if  $p(e, \tau_B) < p(e, \tau_A)$  for all  $e \in \mathcal{E}$ . The centralized system never employs an inferior production technology. In contrast, for some scenarios, it is optimal for the principal to choose an inferior production technology. A higher-temperature configuration can be optimal even when it lowers expected output, because it increases the marginal effect of effort on performance and hence the informativeness of success. This relaxes the incentive constraint: the principal can implement a given effort with a smaller success-contingent bonus, thereby reducing the information rent. (Appendix C provides necessary and sufficient conditions for the principal to optimally choose the inferior production technology in the simple case where there are two effort and temperature choices.) This finding resonates with a broader theme in the organizational economics literature: principals may optimally cede control over aspects of the productive environment to strengthen incentives. Aghion and Tirole (1997), for example, show that delegating decision rights can motivate agent effort by raising the agent’s stake in outcomes; in our setting, the principal “delegates” to an inferior technology that serves a parallel incentive function.

The upward distortion in temperature drives the upward distortions in effort and success probability. We now formalize one aspect of this mechanism.

**LEMMA 1.** *For a fixed effort  $e$ , decentralization distorts temperature upward  $\tau^*(e) \geq \tau^{FB}(e)$ .*

As noted above, the complementarity between temperature and effort, formalized in inequality (3), means that the principal’s cost of inducing any particular effort level  $e > 0$  decreases in the

temperature  $\tau$ . Formally, the principal's optimal bonus  $b^*(e, \tau)$  which induces effort  $e > 0$  under temperature  $\tau$  decreases in  $\tau$ ; this result is formalized by Lemma A2(b) in Appendix A. Because, as noted above, for any effort  $e$ , the centralized system's optimal temperature  $\tau^{FB}(e)$  maximizes the success probability  $p(e, \tau)$ , the principal's upward distortion of temperature (to reduce the cost of inducing effort) comes at the expense of reducing the probability of success  $p(e, \tau^*(e)) \leq p(e, \tau^{FB}(e))$ . We examine the implications of these distortions for the impact of the reward for success on the optimal temperature and effort in the next section.

## 5. How Task Stakes Shape AI System and Incentive Design

How should the principal change its provision of incentives and the configuration of the AI system in response to changes in the external environment? For example, how does a change in the task stakes, which in our setting is captured in the reward for success  $R$ , affect the principal's optimal temperature  $\tau^*$  and induced effort  $e^*$ ? Similarly, how does a change in the costliness of effort  $k$  affect the principal's optimal decisions? To simplify the exposition, for the bulk of this section we focus on the impact of the reward for success  $R$ . At the end of the section, we briefly discuss how analogous reasoning applies to changes in the costliness of effort  $k$ .

In the classical moral hazard setting, higher rewards lead the principal to optimally induce higher effort. When the rewards for success are higher, the principal places more weight on success and implements higher effort via a higher-powered contract. This intuition continues to hold in our model when the system design—specifically, the temperature of the enterprise AI system—is exogenous. However, when the principal can *endogenously* choose the temperature  $\tau$ , the comparative statics with respect to the reward  $R$  become more nuanced. The principal faces a trade-off: increasing the temperature  $\tau$  amplifies the marginal return to effort, but may reduce the success probability. (The success probabilities in Figure 1 illustrate this trade-off: under high temperature, the marginal effect on the success probability of increasing effort (especially from  $e_M$  to  $e_H$ ) is higher, but the success probability is lower when effort is low  $e_L$  or moderate  $e_M$ .) The next two subsections explore the implications of the subtle interaction between temperature and effort for how the reward  $R$  affects the principal's optimal decisions.

We characterize how the reward  $R$  affects the principal's optimal effort  $e^*$  in Section 5.1 and optimal temperature  $\tau^*$  in Section 5.2. In both sections, we contrast the results with those of the benchmark case where decision making is centralized. In Section 5.1, we also contrast the results with those of the benchmark case where the temperature is exogenous.

### 5.1. Impact on the Optimal Effort

When temperature is exogenous, the agent's optimal effort is increasing in the reward  $R$ : a higher  $R$  raises the marginal value of success and therefore strengthens the incentive to exert effort. Part (a)

of Lemma A4 in Appendix A provides a formal statement of this result for the centralized case. A parallel result occurs under decentralization: as  $R$  increases, the principal responds by strengthening incentives, offering a higher success-contingent bonus and thereby inducing higher effort. See part (b) of Lemma A4. Part (c) of Lemma A4 shows that the monotonicity result for the centralized system extends when temperature is endogenous.

However, this monotonicity need not survive in the decentralized setting when temperature is endogenous. The principal then chooses not only the effort level to implement, but also the sensitivity of success to effort. We now show that the implemented effort can vary non-monotonically with the principal's reward for success  $R$ . The mechanism is subtle: when stakes rise, the principal has stronger incentives to increase the probability of success. One way to do so is to reduce the upward distortion in temperature; that is, move  $\tau$  closer to the success-probability-maximizing level  $\tau^{FB}(e)$  for the implemented effort. This can reduce the sensitivity of success to effort and thereby weaken incentives.

**PROPOSITION 2A.** *There exist scenarios in which the optimal effort  $e^*$  is non-monotonic in the reward  $R$ .*

The next proposition provides sufficient conditions under which the ability of the principal to tune the temperature  $\tau$  reverses the intuitive effort-increases-in-the-reward result. As in Section 4, to keep the conditions simple and easy to interpret, Proposition 2B considers the case where temperature and effort are discrete. To aid in interpreting eqs. (16) and (17), it is useful to note that the inequalities in (18) below imply

$$\delta < \Delta < \alpha. \quad (15)$$

The proof establishes that the success probabilities (eqs. (16) and (17)) satisfy the assumptions in Section 3.

**PROPOSITION 2B.** *Suppose the action spaces for temperature and effort are discrete,  $(\mathcal{T}, \mathcal{E}) = (\{\tau_L, \tau_H\}, \{e_L, e_M, e_H\})$ , effort is evenly spaced, and the success probabilities are*

$$p(e_L, \tau_L) = \alpha, \quad p(e_M, \tau_L) = \alpha + \delta, \quad p(e_H, \tau_L) = \alpha + \delta + \varepsilon, \quad (16)$$

$$p(e_L, \tau_H) = 0, \quad p(e_M, \tau_H) = \Delta, \quad p(e_H, \tau_H) = 2\Delta - \varepsilon, \quad (17)$$

where

$$\frac{\alpha}{2} + \frac{\delta^2}{\alpha + \delta} < \Delta < \frac{\alpha + \delta}{2} \quad \text{and} \quad \delta > 0. \quad (18)$$

There exist thresholds  $0 < R_1 < R_2 < R_3 < R_4$  and  $\bar{\varepsilon} > 0$  such that for all  $\varepsilon \in (0, \bar{\varepsilon})$ , it is optimal for the principal to induce low effort  $e^* = e_L$  for  $R \in (0, R_1)$ , high effort  $e^* = e_H$  for  $R \in (R_2, R_3)$ , and moderate effort  $e^* = e_M$  for  $R \in (R_3, R_4)$ . Moreover, the bonus increases in the reward  $R$  across these regions:  $b^*|_{R \in (0, R_1)} < b^*|_{R \in (R_2, R_3)} < b^*|_{R \in (R_3, R_4)}$ .

The sufficient conditions for the principal's optimal effort to be non-monotonic in the reward for success, namely eqs. (16) to (18), can be interpreted as follows: (i) the success probability  $p(e, \tau)$  decreases in the temperature  $\tau$ , (ii) the success probability  $p(e, \tau)$  exhibits complementarity between effort and temperature, (iii) under high temperature, the success probability  $p(e, \tau_H)$  exhibits nearly constant returns, and (iv) under low temperature, the success probability  $p(e, \tau_L)$  exhibits sharply diminishing returns. Below, we explain the role of each condition in contributing to the non-monotonicity result. For expositional convenience, we say that the reward for success is low when  $R \in (0, R_1)$ , is moderate when  $R \in (R_2, R_3)$ , and is high when  $R \in (R_3, R_4)$ .

To understand why the optimal effort  $e^*$  induced by the principal increases and then decreases as the reward for success increases across these regions, it is useful to understand the two considerations that drive the principal's choice of effort and temperature. First, the principal seeks to maximize the probability of receiving a positive payoff, which only occurs when the outcome is successful; that is, the principal seeks to maximize  $p(e, \tau)$ . Second, the principal seeks to minimize the bonus  $b^*(e, \tau)$  paid to the agent. When the principal seeks to induce effort  $e > 0$ , the bonus decreases in the marginal return from effort  $[p(e, \tau) - p(e', \tau)]/(e - e')$ , where  $e' = \max\{x \in \mathcal{E} \mid x < e\}$ , which increases in the temperature  $\tau$ .

When the reward for success  $R$  is small, there is little value in inducing costly effort, so the principal offers bonus  $b^* = 0$  and induces low effort  $e^* = e_L$ . In this case, the principal's temperature choice is wholly driven by the first consideration of maximizing the success probability. Because the success probability decreases in the temperature, the principal chooses low temperature  $\tau^* = \tau_L$ .

When the reward for success increases to a moderate level, there is value in inducing costly effort. Because the success probability  $p(e, \tau)$  exhibits complementarity between effort and temperature, the cost of inducing effort is minimized under the high temperature. Accordingly, the principal switches to  $\tau^* = \tau_H$ . Because under this temperature the returns to effort are nearly constant, it is optimal to induce high effort  $e^* = e_H$ .

When the reward for success increases to a high level, there is value in further increasing the success probability. Because the success probability decreases in the temperature, the principal chooses low temperature  $\tau^* = \tau_L$ . Because the returns to effort sharply diminish under low temperature, it is optimal to induce moderate effort  $e^* = e_M$ .

This explains why, as the reward increases across the three regions, the principal's optimal effort increases from low  $e_L$  to high  $e_H$  and then decreases to moderate  $e_M$ . As one would expect, the optimal success probability and bonus increase as the reward increases across these regions. The more interesting framing is that as the reward increases from moderate to large, the principal induces less effort but increases the bonus to do so. The intuition is that the principal couples the increased bonus

with reduced temperature, and the latter's negative effect on effort outweighs the former's positive effect.

**REMARK 3.** Although Proposition 2B does not pin down the principal's optimal effort when the reward  $R \in [R_1, R_2] \cup [R_4, \infty)$ , numerical examples satisfying the proposition's conditions exhibit the same qualitative pattern: as  $R$  increases from zero, the optimal effort moves from low to high and then to moderate; Appendix F reports a calibrated example that illustrates this behavior and computes the threshold values separating the low-, high-, and moderate-effort regions for a particular parameterization. Moreover, the non-monotonicity result in Propositions 2A and 2B does not rely on discreteness of the action spaces  $(\mathcal{E}, \mathcal{T})$ ; Appendix E provides an example with continuous action spaces in which the principal's optimal effort is nonmonotone in  $R$ .

The managerial implication of Propositions 2A and 2B is that a manager should be wary of applying the intuitive prescription from either the centralized system with endogenous temperature or the decentralized system with exogenous temperature that as the reward increases, the manager should induce higher effort.

## 5.2. Impact on the Optimal Temperature

We now turn to the impact of the reward on the optimal temperature. As noted at the start of this section, it is intuitively appealing that the optimal effort increases in the reward. Because effort and temperature are complements, it is natural to conjecture that the optimal temperature also increases in the reward. Indeed, the first-best temperature  $\tau^{FB}$  increases in the reward  $R$  (see Lemma A5 in Appendix A). Intuitively, because the value of increasing the success probability increases in the reward and because the marginal impact of effort on the success probability increases in the temperature, it is natural that a large reward would be associated with high temperature and effort.

Proposition 3A shows that, in contrast to the result for the centralized setting, the principal's utility-maximizing temperature  $\tau^*$  may be non-monotonic in the reward.

**PROPOSITION 3A.** *There exist parameters under which the optimal temperature  $\tau^*$  is non-monotonic in the reward  $R$ .*

The next proposition provides sufficient conditions under which the temperature-increases-in-the-reward result is reversed. To allow for the development of conditions that are simple and easy to interpret, Proposition 3B considers the case where effort is discrete.

**PROPOSITION 3B.** *Suppose that the action space of temperature is continuous  $\mathcal{T} = [\tau_L, \tau_H]$ , the success probability  $p(e, \tau)$  is concave in the temperature  $\tau$ ,  $(\partial/\partial\tau)p(e, \tau)|_{\tau=\tau_H} < 0 \leq (\partial/\partial\tau)p(e, \tau)|_{\tau=\tau_L}$ , and the action space of effort  $\mathcal{E}$  is discrete. (a) If  $\mathcal{E} = \{e_L, e_H\}$ , then there exists  $\underline{R} \in (0, \infty)$  such that the optimal temperature  $\tau^*$  is invariant to the reward  $R$  on  $R \in (0, \underline{R})$ ,  $\tau^*$  decreases in  $R$  on*

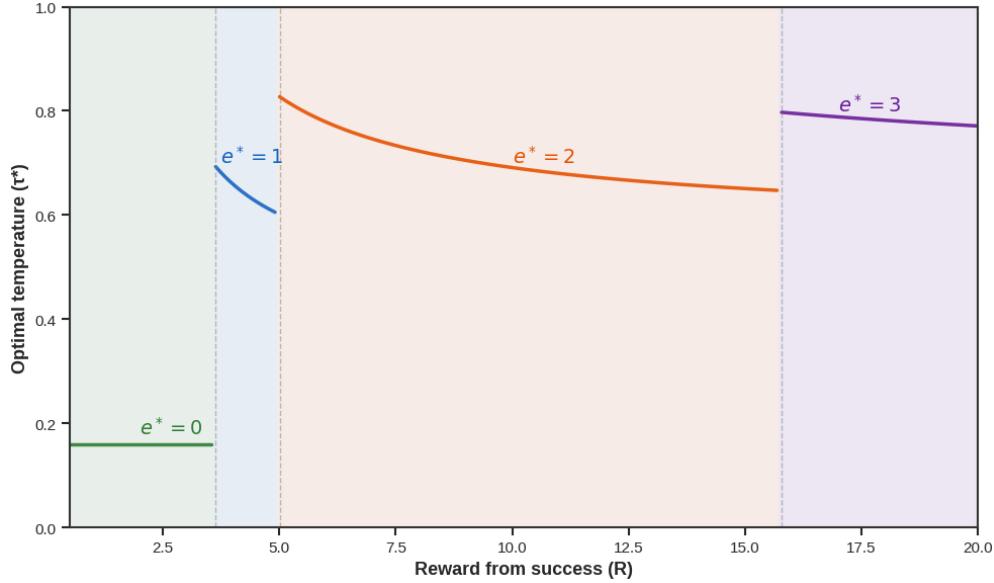
$R \in (\underline{R}, \infty)$ , and  $\lim_{R \uparrow \underline{R}} \tau^*(R) < \lim_{R \downarrow \bar{R}} \tau^*(R)$ . (b) If  $\mathcal{E} = \{0, e_1, e_2, \dots, e_N\}$  and the principal induces effort  $e_j$  for  $R \in (\underline{R}_j, \bar{R}_j)$ , where  $j \in \{1, 2, \dots, N\}$ , then the principal's utility-maximizing temperature  $\tau^*(e_j)$  decreases in the reward  $R$  for  $R \in (\underline{R}_j, \bar{R}_j)$ .

Earlier, in Proposition 2B, the sufficient conditions for the principal's optimal effort  $e^*$  to be non-monotonic in the reward for success  $R$  are rather restrictive in the sense that they require the success probability function  $p(e, \tau)$  to have a specific form. By contrast, Proposition 3B's sufficient conditions for the principal's optimal temperature  $\tau^*$  to be non-monotonic in the reward are relatively mild: the success probability  $p(e, \tau)$  is concave in the temperature, and it is not maximized at the upper limit of the temperature action space. These assumptions are natural in GenAI settings: returns to increasing temperature conceivably diminish as outputs become overly exploratory, and pushing temperature to its maximum can be counterproductive for success (e.g., by making hallucinations overwhelming); intermediate temperatures balance diversity and accuracy more effectively than the extremes (Wolfram 2023).

Proposition 3B's part (a) reveals that, in the discrete effort setting, the principal's optimal temperature  $\tau^*$  is non-monotonic in the reward. Namely, as the reward increases, the optimal temperature is initially flat, then jumps up, and finally decreases. Part (b) reveals that when the principal induces effort  $e > 0$  over some range of reward, the principal's optimal temperature decreases in the reward. Figure 2 illustrates the non-monotonicity of the optimal temperature and the induced “jumps” in effort across stakes.

The temperature choice reflects a trade-off between a *rent-extraction effect* and a *performance-loss effect* once incentives are operative. On the one hand, raising temperature may increase the sensitivity of success to effort, which relaxes the incentive constraint. This allows the principal to implement a given effort with a smaller success-contingent bonus and hence lower information rent. On the other hand, for any implemented effort  $e$ , the centralized benchmark selects  $\tau^{FB}(e) \in \arg \max_{\tau \in \mathcal{T}} p(e, \tau)$ . Thus, distorting temperature away from  $\tau^{FB}(e)$  reduces the attainable success probability. When the optimal contract entails an upward distortion  $\tau^*(e) > \tau^{FB}(e)$ , this loss becomes increasingly costly as the reward  $R$  grows, pushing  $\tau^*(e)$  back toward  $\tau^{FB}(e)$ .

We begin by unpacking the intuition for part (b), where the implemented effort level is fixed at  $e_j$  on  $R \in (\underline{R}_j, \bar{R}_j)$ . As noted in the discussion surrounding Lemma 1, the principal distorts the temperature upward to reduce the cost of inducing effort. Formally, the proof of Proposition 3B shows that under the conditions of the proposition, when the principal induces effort  $e > 0$ , the distortion in temperature is strict  $\tau^*(e) > \tau^{FB}(e)$ . Because  $\tau^{FB}(e)$  maximizes  $p(e, \tau)$ , any deviation from  $\tau^{FB}(e)$  weakly lowers the success probability:  $p(e, \tau^*(e)) \leq p(e, \tau^{FB}(e))$ . Over an interval where effort cannot adjust, the principal mainly trades off the rent-extraction effect of an upward temperature distortion



**Figure 2 Optimal temperature  $\tau^*$  as the reward for success  $R$  increases.** Action spaces are  $\mathcal{T} = [0, 1]$ ,  $\mathcal{E} = \{0, 1, 2, 3\}$ . The success probability is  $p(e, \tau) = a_0 + a_1\tau - a_2\tau^2 + (b_0 + b_1\tau - b_2\tau^2)e - ce^2$  with  $(a_0, a_1, a_2) = (0.17, 0.08, 0.25)$ ,  $(b_0, b_1, b_2) = (0.37, 0.23, 0.11)$ , and  $c = 0.06$ .

against the performance-loss effect of moving away from  $\tau^{FB}(e)$ . As the reward  $R$  increases, the marginal value of raising the probability of success  $p(e, \tau)$  rises, so the optimal temperature decreases toward  $\tau^{FB}(e)$ , as stated in part (b).

To see the intuition for part (a), begin with low stakes,  $R < \underline{R}$ . Inducing  $e_H$  would require a transfer that is not warranted by the additional expected surplus, so the optimal contract is degenerate: the principal sets bonus  $b^* = 0$  and the agent exerts effort  $e^* = e_L = 0$ . With incentives inactive, neither the complementarity effect nor the rent-extraction effect is operative, and the principal chooses the temperature that maximizes success given  $e_L$ , so  $\tau^*(e_L) = \tau^{FB}(e_L)$ . When the reward  $R$  exceeds the threshold  $\underline{R}$ , the principal optimally responds by offering a bonus  $b^* > 0$  that induces the agent to exert effort  $e_H$ . At the point that the reward increases over threshold  $\underline{R}$ , the principal's optimal temperature “jumps up” from  $\tau^{FB}(e_L)$ . Two forces contribute to this jump. First, because effort and temperature are complements, the success-probability-maximizing benchmark shifts upward from  $\tau^{FB}(e_L)$  to  $\tau^{FB}(e_H)$ . Second, conditional on implementing  $e_H$ , the rent-extraction effect motive generates an additional upward distortion, so  $\tau^*(e_H) > \tau^{FB}(e_H)$ . Following the logic from part (a), as the reward  $R$  increases further (while  $e_H$  remains optimal), the performance-loss effect becomes more important, and the optimal temperature declines toward  $\tau^{FB}(e_H)$  even though doing so increases the incentive cost of implementing  $e_H$ .

Figure 2 illustrates how the pattern described in part (a) with a single reward threshold may extend to one with multiple reward thresholds when the action space for effort  $\mathcal{E}$  includes multiple non-zero

elements. In the figure, each time the reward crosses a threshold, the principal's optimal induced effort increases, and the optimal temperature jumps up and then decreases until the reward crosses a subsequent threshold.

**REMARK 4.** The non-monotonicity results in Propositions 3A and 3B are not artifacts of the discrete-effort and continuous-temperature restrictions imposed for tractability. Although Proposition 3B takes  $\mathcal{T}$  to be continuous, the same qualitative pattern obtains when temperature is discrete. In particular, as noted in the discussion following Proposition 2B, under the conditions of that proposition—where the temperature action space is discrete—the principal's optimal temperature is also non-monotonic in the reward for success. Likewise, the non-monotonicity result in Propositions 3A and 3B is not an artifact of assuming a discrete effort action space  $\mathcal{E}$ . Appendix E provides an example with continuous effort and continuous temperature in which the principal's optimal temperature is non-monotonic in the reward.

The managerial implication of Propositions 3A and 3B is that a manager should be wary of applying the intuitive prescription from the centralized system that, as the reward increases, the manager should choose a higher temperature.

We conclude by discussing the impact of the costliness of effort  $k$  on the optimal effort and temperature. It is straightforward to show that increasing the costliness of effort has the same directional effect on the optimal effort and temperature as decreasing the reward for success  $R$ . Specifically, as the costliness of effort  $k$  increases, the centralized system's optimal effort  $e^{FB}$  decreases. Similarly, in the decentralized setting, when the temperature is exogenous, the principal's optimal effort  $e^*$  decreases in the costliness of effort  $k$ . Parallel to Propositions 2A and 2B, one can show that there exist parameters under which the principal's optimal effort is non-monotonic (e.g., first increasing and then decreasing) in the costliness of effort. Thus, the intuitive prescriptions regarding the impact of the costliness of effort on the optimal effort level from the benchmark cases break when temperature is endogenous.

As the costliness of effort  $k$  increases, the centralized system's optimal temperature  $\tau^{FB}$  decreases, and in the setting with fixed effort, the principal's optimal temperature  $\tau^*$  decreases. Parallel to Propositions 3A and 3B, one can show that there exist parameters under which the principal's optimal temperature  $\tau^*$  is non-monotonic (e.g., first increasing and then decreasing) in the costliness of effort  $k$ . The intuition for why the optimal temperature increases in the costliness of effort parallels the intuition above. As noted previously, the principal distorts the temperature upward to reduce the cost of inducing effort  $e$ . When the costliness of effort increases, it becomes attractive to reduce the cost of inducing effort by increasing the distortion (i.e., increasing the temperature  $\tau$ , pushing it farther from

the first-best level  $\tau^{FB}(e)$ ), which reduces the required bonus  $b^*(e, \tau)$  as well as the probability of paying the bonus (i.e., the success probability  $p(e, \tau)$ ).

## 6. Impact of AI System Design on Agent Utility

The principal (weakly) benefits by having discretion over the temperature in designing the GenAI system. Does this discretion come at the expense of the agent? As noted above, by distorting the temperature upward, the principal reduces the agent's rent because it is easier for the principal to infer the agent's effort when the temperature is high. This suggests that the benefit to the principal may come at the expense of the agent.

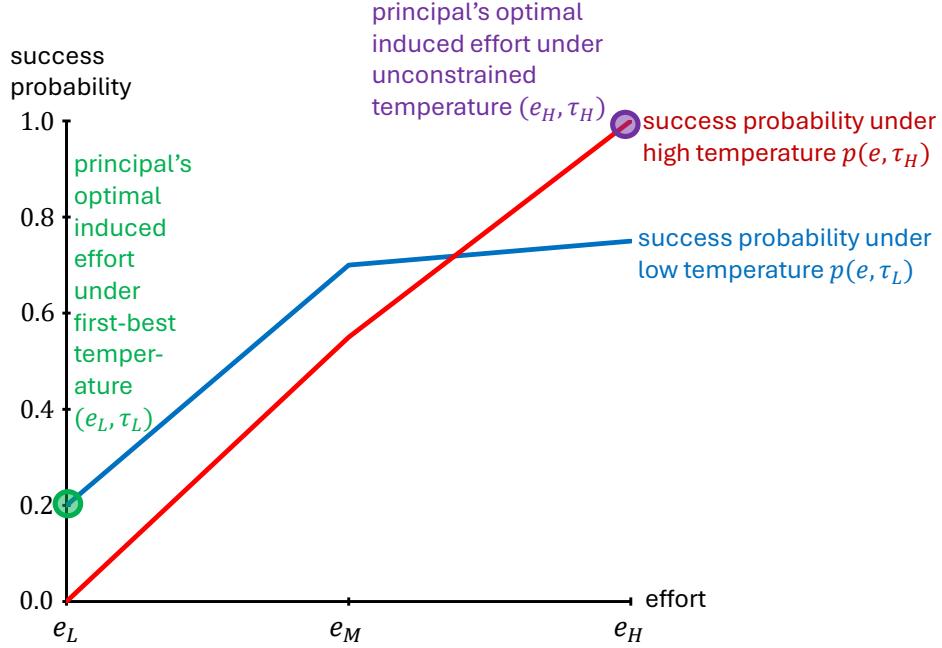
Proposition 4A reveals that the principal's discretion over temperature can be a win-win in that both the principal and the agent benefit. For the benchmark case in which the principal lacks discretion over temperature, we hold temperature fixed at the centralized benchmark choice  $\tau^{FB} = \tau^{FB}(e^{FB})$ , that is, the temperature that maximizes the success probability  $p(e^{FB}, \tau)$ . Let  $U(\tau) = u(e^*(\tau), \tau, b^*(e^*(\tau), \tau))$  denote the agent's expected utility at temperature  $\tau$  when the principal induces the effort level  $e^*(\tau)$  that maximizes the principal's expected profit.

**PROPOSITION 4A.** *Suppose that for any temperature  $\tau$ , the principal induces the effort level  $e^*(\tau)$  that maximizes the principal's expected profit. There exist parameters such that the agent's expected utility under the principal's profit-maximizing temperature is strictly greater than under the first-best temperature  $U(\tau^*) > U(\tau^{FB})$ .*

Next, Proposition 4B provides sufficient conditions under which the agent benefits from the principal's upward distortion of temperature. The conditions mirror those in Proposition 1B, with the difference that the upper bound on the reward for success in (14) is tightened. Figure 3 illustrates Proposition 4B.

**PROPOSITION 4B.** *Suppose that for any temperature  $\tau$ , the principal induces the effort level  $e^*(\tau)$  that maximizes the principal's expected profit. Suppose that the action spaces for temperature and effort are discrete  $(\mathcal{T}, \mathcal{E}) = (\{\tau_L, \tau_H\}, \{e_L, e_M, e_H\})$ , effort is evenly spaced, and the success probability  $p(e, \tau)$  satisfies eqs. (11) and (12), where eqs. (13) and (14) hold and  $R/e_M < (\alpha + \Delta)/\Delta^2$ . Then there exists  $\bar{\varepsilon} > 0$  such that if  $\varepsilon \in (0, \bar{\varepsilon})$ , then the agent's expected utility under the principal's profit-maximizing temperature is strictly greater than under the first-best temperature  $U(\tau^*) > U(\tau^{FB})$  and decentralization distorts the temperature upward  $\tau^* > \tau^{FB}$ .*

For the reasons discussed following Proposition 1B, under the conditions of Proposition 4B, the first-best temperature is low  $\tau^{FB} = \tau_L$ . Because under low temperature the success probability is relatively insensitive to effort, the information rent ceded to the agent to induce moderate or high effort is prohibitively large. Consequently, it is optimal for the principal to induce low effort  $e^*(\tau^{FB}) = e_L$ ,



**Figure 3 Temperature discretion can raise agent surplus (Proposition 4B).** The success probability  $p(e, \tau)$  is plotted against effort levels  $(e_L, e_M, e_H) = (0, 1, 2)$  under low and high temperatures  $\tau_L$  and  $\tau_H$ . Parameters are  $\alpha = 0.20$ ,  $\Delta = 0.50$ ,  $\varepsilon = 0.05$ , and  $R = 2.70$ . When the principal is constrained to operate under the first-best temperature  $\tau^{FB} = \tau_L$ , the principal optimally induces low effort  $e_L$ , yielding success probability  $p = 0.20$  and zero agent utility. With temperature discretion, the principal optimally selects  $(e_H, \tau_H)$ , yielding success probability  $p = 1.0$  and strictly positive agent utility.

which the principal achieves by offering bonus  $b^*(e^*(\tau^{FB}), \tau^{FB}) = 0$ . With no possibility of payment from the principal, the agent's utility is zero.

When the principal has discretion over temperature, the principal distorts the temperature upward  $\tau^* = \tau_H > \tau^{FB}$  because doing so reduces the information rent associated with inducing high effort. Under high temperature, it is attractive for the principal to induce high effort  $e^*(\tau^*) = e_H$ . Inducing the agent to exert costly effort requires the principal to cede rent to the agent. Hence, the agent utility is higher when the principal has discretion over temperature, even when the principal uses that discretion to distort the temperature upward to reduce the agent's rent.

The welfare gain for the agent has a parallel in classic results on monitoring: when outcome signals become more informative, principals can profitably induce higher effort, and agents can benefit if the equilibrium shifts from a lower-effort regime with lower rent to a higher-effort regime with higher rent (see, e.g., Holmstrom 1979, Diamond 1984, Dye 1986, Bester and Krähmer 2008). In our setting, temperature serves a *distinct* role from monitoring. Monitoring improves the *observability* of effort without changing the production technology; temperature changes the *technology* itself, as embodied in the success probability  $p(e, \tau)$ , by influencing the sensitivity of success to effort. Both instruments

can reduce the bonus needed to implement a target effort, but only temperature simultaneously affects the distribution of outcomes that determines the principal's objective. This technological channel is what allows the agent's surplus to rise even when the principal chooses a higher, bonus-dampening temperature.

Our analysis so far assumes that temperature and effort are complements, so that  $p(e, \tau)$  exhibits increasing differences. However, in principle, temperature and effort could be substitutes, meaning that the direction of the inequality in eqs. (1) to (3) is reversed: higher temperature causes incremental effort to be less effective. This alternative formulation alters only one of our central insights. When lower temperature makes success more sensitive to hidden effort (i.e., the success signal becomes more informative about effort), the minimal success-contingent bonus needed to implement a given effort falls as temperature decreases, so the analog of Lemma 1 reverses and the moral-hazard distortion in temperature is toward *lower* temperature  $\tau^*(e) \leq \tau^{FB}(e)$ . Importantly, our remaining findings hold qualitatively under this alternative formulation. Under decreasing differences, one can construct environments in which decentralization distorts effort and success probability upward, the agent benefits from the principal's discretion over temperature, and optimal temperature responds non-monotonically to task stakes. The economic logic is unchanged: the principal uses temperature to sharpen the link between effort and outcomes, reducing the cost of implementing high effort. What differs is which temperature level serves this role.

## 7. Concluding Remarks

The organizational challenge posed by GenAI is both technological and economic: how should firms design systems—comprising both human agents and AI—that function jointly to deliver creativity and reliability? This paper develops a unified framework to analyze the trade-offs embedded in enterprise AI system design, with particular attention to hallucination, creativity, and moral hazard. Whereas much of the public and policy debate treats AI hallucination as a systems engineering failure, we show it can instead be the rational consequence of incentive-aligned organizational choices.

A central message of our analysis is that system design and incentive design are intertwined. Temperature plays a dual role. It governs the system's generative behavior, and it enters the agency problem by changing how responsive performance is to effort. Raising temperature increases the marginal effect of the agent's review and correction on the probability of success, which lowers the incentive payments required to elicit effort. As a result, the principal may optimally choose a higher-temperature design that performs well when effort is high but performs poorly when effort is low, because hallucinations are more likely to go uncorrected.

When temperature is exogenous, the standard moral-hazard logic applies: hidden effort makes incentives costly and weakens performance. When the principal can also choose temperature, this

design choice changes the informativeness of outcomes and hence the cost of providing incentives. As a result, decentralization need not weaken performance; it can lead the principal to induce higher effort and achieve a higher probability of success than under the full-information benchmark.

Temperature discretion also has implications for agent welfare. The principal can use temperature to lower the information rent needed to induce effort. At the same time, when temperature choice is coupled with stronger incentives, it can shift the equilibrium toward higher-powered contracts and higher effort, in which case the agent can be strictly better off than under an exogenous-temperature benchmark.

Endogenous temperature also changes the comparative statics in task stakes. When temperature is exogenous, higher rewards support higher-powered incentives and higher induced effort. Temperature choice shifts both the bonus and the sensitivity of success to effort; hence, the optimal temperature and induced effort can vary non-monotonically with task stakes. As stakes rise, the principal may increase temperature to make effort more effective at the margin and thereby reduce the incentive cost (and information rent) of inducing effort. Alternatively, the principal may reduce temperature to move closer to the reliability-maximizing setting for the implemented effort, thereby raising the probability of success at that effort even though incentives weaken. Because the relative importance of these forces changes with stakes, both optimal temperature and induced effort can vary non-monotonically with task stakes. This stake-dependent temperature choice implies that a “one-size-fits-all” approach to enterprise AI governance, such as a uniform requirement for low-temperature factual outputs, is likely suboptimal. In high-stakes settings where human review is essential, driving temperature too low can make effort too costly to incentivize, yielding a system that is safer in isolation but less effective in equilibrium.

Ultimately, the key question is not whether AI will be used in day-to-day business activities, but how firms will shape its internal architecture. Our framework highlights that these choices are endogenous to economic frictions. The effectiveness of AI in enterprise settings depends not only on its technical sophistication but also on how it endogenizes the information structure of the firm. Designing AI systems that are creative, disciplined, and incentive-aligned is a managerial challenge, but it is also a first-order driver of value.

In this paper, we focus on a single-shot use of AI to keep the analysis transparent; iterative human–AI workflows can be interpreted as raising the agent’s effective effort, which would only strengthen the role of effort–temperature complementarity that drives our results. Relatedly, while some deployments record intermediate inputs (e.g., prompt logs or draft iterations), such data are often non-verifiable or only weakly informative about true effort, so our emphasis on hidden effort remains a natural benchmark. More broadly, the same logic extends to richer environments—such as settings with heterogeneous agents or delegated hyperparameter choices—where temperature would

also interact with screening or reflect private costs, without altering the central message that enterprise GenAI design and incentive provision are jointly determined.

## Acknowledgments

We are grateful for constructive feedback from session and seminar participants at the 2025 INFORMS MSOM Conference (London, U.K.), the 2025 Workshop on Unstructured Data and Language Models (Ross School of Business, University of Michigan), and London Business School. We also thank Volodymyr Babich, Gérard Cachon, Francisco Castro, George Chen, Ilgan Dogan, Martin Lariviere, Anton Ovchinnikov, Erica Plambeck, Nicos Savva, and Mohan Sodhi for helpful comments, as well as colleagues who attended our presentations at INFORMS Annual Meetings in 2023 and 2024.

## References

- Abril D (2023) Is ChatGPT safe for work? Here's what companies should consider. *The Washington Post* URL <https://www.washingtonpost.com/business/2023/07/10/chatgpt-safe-company-work-ban-lawyers-code/>.
- Acemoglu D, Restrepo P (2018) The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review* 108(6):1488–1542.
- Aghion P, Tirole J (1997) Formal and real authority in organizations. *Journal of Political Economy* 105(1):1–29.
- Agrawal A, Gans J, Goldfarb A (2018) Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 89–110.
- Agrawal A, Gans JS, Goldfarb A (2024) Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy* 33(2):327–337.
- Alder S (2024) Is ChatGPT HIPAA compliant? *HIPAA Journal* URL <https://www.hipaajournal.com/is-chatgpt-hipaa-compliant/>.
- Allon G (2024) The impact of Gen AI on service jobs. Gad's Newsletter, URL <https://gadallon.substack.com/p/the-impact-of-gen-ai-on-service-jobs>, accessed: 2025-01-13.
- Alp O, Şen A (2021) Delegation of stocking decisions under asymmetric demand information. *Manufacturing & Service Operations Management* 23(1):55–69.
- Atasu A, Ciocan DF, Désir A (2021) Price delegation with learning agents. Working paper, INSEAD.
- Autor DH, Levy F, Murnane RJ (2003) The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics* 118(4):1279–1333.
- Baiman S, Netessine S, Saouma R (2010) Informativeness, incentive compensation, and the choice of inventory buffer. *Accounting Review* 85(6):1839–1860.

- Balachandran KR, Radhakrishnan S (2005) Quality implications of warranties in a supply chain. *Management Science* 51(8):1266–1277.
- Bastani H, Cachon GP (2025) The human–AI contracting paradox, working paper.
- Bester H, Krähmer D (2008) Delegation and incentives. *RAND Journal of Economics* 39(3):553–572.
- Boyaci T, Canyakmaz C, de Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* 70(2):1258–1275.
- Bringsjord S, Govindarajulu NS (2024) Artificial Intelligence. Zalta EN, Nodelman U, eds., *The Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Stanford University), Fall 2024 edition.
- Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *Quarterly Journal of Economics* 140(2):889–942.
- C3ai (2024) C3.ai Inc. Q3 2024 earnings call transcript. URL <https://seekingalpha.com/article/4674584-c3-ai-inc-ai-q3-2024-earnings-call-transcript>, available at Seeking Alpha.
- Chen N, Hu M, Li W (2023) Algorithmic decision-making safeguarded by human knowledge. Working Paper.
- Chen Y, Kirshner SN, Ovchinnikov A, Andiappan M, Jenkin T (2025) A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do? *Manufacturing & Service Operations Management* 27(2):354–368.
- Cummings M (2025) Prohibiting generative AI in any form of weapon control. NeurIPS 2025, San Diego, Poster, URL <https://neurips.cc/virtual/2025/loc/san-diego/poster/121921>, poster session: Fri, Dec 5, 2025.
- Dai T, Ke R, Ryan CT (2021) Incentive design for operations-marketing multitasking. *Management Science* 67(4):2211–2230.
- Dai T, Singh S (2020) Conspicuous by its absence: Diagnostic expert testing under uncertainty. *Marketing Science* 39(3):540–563.
- Dai T, Singh S (2025) Artificial intelligence on call: The physician’s decision of whether to use AI in clinical practice. *Journal of Marketing Research* 62(5):854–875.
- de Véricourt F, Gromb D (2018) Financing capacity investment under demand uncertainty: An optimal contracting approach. *Manufacturing & Service Operations Management* 20(1):85–96.
- de Véricourt F, Gromb D (2019) Financing capacity with stealing and shirking. *Management Science* 65(11):5128–5141.
- de Véricourt F, Gurkan H (2023) Is your machine better than you? You may never know. *Management Science* (in press).
- Demougin D, Fluet C (2001) Monitoring versus incentives. *European Economic Review* 45(9):1741–1764.
- DeRose A (2023) These companies have banned or limited ChatGPT at work. *HR Brew* URL <https://www.hr-brew.com/stories/2023/05/11/these-companies-have-banned-chatgpt-in-the-office>.

- Diamond DW (1984) Financial intermediation and delegated monitoring. *Review of Economic Studies* 51(3):393–414.
- Dye RA (1986) Optimal monitoring policies in agencies. *RAND Journal of Economics* 17(3):339–350.
- Edlin AS, Shannon C (1998) Strict monotonicity in comparative statics. *Journal of Economic Theory* 81(1):201–219.
- Grand-Clément J, Pauphilet J (2023) The best decisions are not the best advice: Making adherence-aware recommendations. URL <https://arxiv.org/abs/2209.01874>.
- Grossman SJ, Hart OD (1983) An analysis of the principal-agent problem. *Econometrica* 51(1):7–45.
- Gurkan H, de Véricourt F (2022) Contracting, pricing, and data collection under the AI flywheel effect. *Management Science* 68(12):8791–8808.
- Holmstrom B (1979) Moral hazard and observability. *Bell Journal of Economics* 10(1):74–91.
- Jarrett PC, Hill J, Howell M, Moore KG, Thoppil JJ, Vargas Ortiz L, Parnell ST, et al. (2025) Temperature-driven variability in emergency diagnostic accuracy by a leading language model. *medRxiv* doi: 10.1101/2025.06.04.25328288, URL <https://doi.org/10.1101/2025.06.04.25328288>.
- Jensen MC, Meckling WH (1976) Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3(4):305–360.
- Karlinsky-Shichor Y, Netzer O (2024) Automating the B2B salesperson pricing decisions: A human-machine hybrid approach. *Marketing Science* 43(1):138–157.
- Ke R, Ryan CT (2018) Monotonicity of optimal contracts without the first-order approach. *Operations Research* 66(4):1101–1118.
- Kim Y, Knight B, Mitrofanov D, Xu Y (2024) AI and worker learning: Evidence from a large-scale field experiment. Working Paper.
- Kirkegaard R (2022) Endogenous success criteria in moral hazard models. *Theoretical Economics* 17(2):445–476.
- Korn J (2023) How companies are embracing generative AI for employees ... or not. *CNN Business* URL <https://edition.cnn.com/2023/09/22/tech/generative-ai-corporate-policy/index.html>.
- Laffont JJ, Martimort D (2002) *The Theory of Incentives: The Principal-Agent Model* (Princeton, NJ, USA: Princeton University Press).
- Li S, Chen KY, Rong Y (2020) The behavioral promise and pitfalls in compensating store managers. *Management Science* 66(10):4899–4919.
- Long X, Nasiry J (2020) Wage transparency and social comparison in sales force compensation. *Management Science* 66(11):5290–5315.
- Lu T, Tomlin B (2025) Augmenting the operations manager with a prediction machine, working paper.

- Mantel Group (2024) Our top 10 insights from a year of genai implementation: A technical guide. URL <https://mantelgroup.com.au>.
- McKinsey (2023) Technology's generational moment with generative AI: A CIO and CTO guide. URL <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/technologys-generational-moment-with-generative-ai-a-cio-and-cto-guide>.
- McKinsey & Company (2023) "A human in the loop is critical." Mckinsey leaders on generative AI at US media day. *New at McKinsey Blog* URL <https://www.mckinsey.com/about-us/new-at-mckinsey-blog/keep-the-human-in-the-loop>.
- McKinsey & Company (2025) The state of AI in 2025: Agents, innovation, and transformation. *McKinsey & Company* URL <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>.
- Mehta B, Salomon O, Alves M, Barth E (2024) AI in financial services survey shows productivity gains across the board. Bain & Company, URL <https://www.bain.com/insights/ai-in-financial-services-survey-shows-productivity-gains-across-the-board/>, accessed: 2025-01-13.
- Mullainathan S, Obermeyer Z (2021) Diagnosing physician error: A machine learning approach to low-value health care. *Quarterly Journal of Economics* 137(2):679–727.
- Nikoofal ME, Gümüs M (2018) Quality at the source or at the end? Managing supplier quality under information asymmetry. *Manufacturing & Service Operations Management* 20(3):498–516.
- Orfanoudaki A, Saghafian S, Song K, Chakerra HA, Cook C (2022) Algorithm, human, or the centaur: How to enhance clinical care? Working Paper No. RWP22-027, Harvard Kennedy School, Cambridge, MA.
- Plambeck EL, Taylor TA (2006) Partnership in a dynamic production system with unobservable actions and noncontractible output. *Management Science* 52(10):1509–1527.
- Plambeck EL, Zenios SA (2000) Performance-based incentives in a dynamic principal-agent model. *Manufacturing & Service Operations Management* 2(3):240–263.
- Plambeck EL, Zenios SA (2003) Incentive efficient control of a make-to-stock production system. *Operations Research* 51(3):371–386.
- Ray S (2023) Samsung bans chatgpt among employees after sensitive code leak. *Forbes* URL <https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/>.
- Rothman J (2023) Metamorphosis. *The New Yorker* (November 20): 28–39.
- Serpa JC, Krishnan H (2017) The strategic role of business insurance. *Management Science* 63(2):384–404.
- Simon HA (1987) Two heads are better than one: The collaboration between AI and OR. *Interfaces* 17(4):8–15.

- Song H, Lai G, Xiao W (2024) Optimal salesforce compensation with general demand and operational considerations. *Manufacturing & Service Operations Management* 26(6):2274–2283.
- Wolfram S (2023) What is ChatGPT doing . . . and why does it work? *Stephen Wolfram Writings*, URL <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work>, accessed May 5, 2025.
- Yu Y, Kong X (2020) Robust contract designs: Linear contracts and moral hazard. *Operations Research* 68(5):1457–1473.

## Appendices

In these appendices, we provide supporting material for the analysis in the main text. Appendix A collects the benchmark problems used for comparison throughout the paper, including the exogenous-temperature moral-hazard benchmark and the centralized benchmark with endogenous temperature. Appendix B contains proofs of the main results and the auxiliary technical arguments deferred from the text. Appendix C characterizes conditions under which an inferior technology is optimal. Appendix D provides conditions that ensure uniqueness of the optimal temperature choice (for a given implemented effort). Appendix E studies extensions in which effort and temperature are continuous. Appendix F provides a numerical illustration of Proposition 2B with  $R$  varying continuously.

### Appendix A: Benchmark Results with Exogenous and Endogenous Temperature

This appendix presents benchmark results that are used as points of comparison throughout the paper. Lemma A1 formalizes how temperature shapes the marginal return to effort under increasing differences. Lemma A2 characterizes the set of success-contingent bonuses that implement a given effort level  $e$  at a given temperature  $\tau$  and shows that the minimal bonus is decreasing in  $\tau$  for any  $e > 0$ . Lemma A3 analyzes the exogenous-temperature benchmark and shows that centralization distorts effort downward. Lemma A4 derives comparative statics in the reward  $R$  when temperature is exogenous and for the first-best effort. Lemma A5 derives comparative statics in  $R$  for the first-best temperature.

To begin with, Lemma A1 is useful in the proof of Lemma A2. Let  $\beta(e_2, e_1, \tau) = (e_2 - e_1)/[p(e_2, \tau) - p(e_1, \tau)]$ .

**LEMMA A1.** *Consider  $\{e_1, e_2\} \in \mathcal{E}$  where  $e_1 < e_2$ . Then  $\beta(e_2, e_1, \tau)$  increases in  $e_1$  and  $e_2$ .*

*Proof of Lemma A1:* It is sufficient to show that  $f(e_1, e_2)$  decreases in  $e_1$  and  $e_2$ , where  $f(e_1, e_2) = [p(e_2, \tau) - p(e_1, \tau)]/(e_2 - e_1)$ . The proof proceeds in two parts. First, we show that  $f(e_1, e_2)$  decreases in  $e_1$ . Second, we show that  $f(e_1, e_2)$  decreases in  $e_2$ . For the remainder of the proof consider  $\{e_L, e_M, e_H\} \in \mathcal{E}$ , where  $0 = e_L < e_M < e_H$ .

First, we show that  $f(e_1, e_2)$  decreases in  $e_1$ . To do so, it is sufficient to show that for any  $e_M \in (e_L, e_H)$ ,  $f(e_L, e_H) > f(e_M, e_H)$ . Because  $p(e, \tau)$  is concave in  $e$ , it follows that

$$p((1 - \gamma)e_H + \gamma e_L) > (1 - \gamma)p(e_H) + \gamma p(e_L) \quad (\text{A1})$$

for any  $\gamma \in (0, 1)$ . Consider any  $e_M \in (e_L, e_H)$  and let  $\gamma = (e_H - e_M)/(e_H - e_L)$ . It follows that  $e_M = (1 - \gamma)e_H + \gamma e_L$ . It then follows from (A1) that  $f(e_L, e_H) > f(e_M, e_H)$ .

Second, we show that  $f(e_1, e_2)$  decreases in  $e_2$ . To do so, it is sufficient to show that for any  $e_M \in (e_L, e_H)$ ,  $f(e_L, e_M) > f(e_L, e_H)$ . Because  $p(e, \tau)$  is concave in  $e$ , it follows that

$$p(\gamma e_H + (1 - \gamma)e_L) > \gamma p(e_H) + (1 - \gamma)p(e_L) \quad (\text{A2})$$

for any  $\gamma \in (0, 1)$ . Consider any  $e_M \in (e_L, e_H)$  and let  $\gamma = (e_M - e_L)/(e_H - e_L)$ . It follows that  $e_M = \gamma e_H + (1 - \gamma)e_L$ . It then follows from (A2) that  $f(e_L, e_M) > f(e_L, e_H)$ . *Q.E.D.*

Lemma A2(a) is useful in the proofs of Lemmas 1 and A3. Lemma A2(b) is useful in the proof of Lemma 1. Let  $b^*(e, \tau)$  denote the principal's optimal bonus that implements effort  $e$  under temperature  $\tau$ .

LEMMA A2. (a) For any effort and temperature  $(e, \tau) \in (\mathcal{E}, \mathcal{T})$ , there exists a bonus  $b$  such that  $e$  is a best response for the agent. Further, the optimal bonus  $b^*(e, \tau) = \sup_{e' \in \mathcal{E} \cup [0, e)} \beta(e, e', \tau)$  for all  $e \neq 0$ , and  $b^*(0, \tau) = 0$ . If the action space of effort is discrete  $\mathcal{E} = \{0, e_1, e_2, \dots, e_N\}$ , where  $0 < e_1 < e_2 < \dots < e_N$ , then  $b^*(e_i, \tau) = \beta(e_i, e_{i-1}, \tau)$  for  $i \in \{1, 2, \dots, N\}$ , where by convention  $e_0 = 0$ . (b) The optimal bonus  $b^*(e, \tau)$  decreases in  $\tau$  for all  $e \neq 0$ .

*Proof of Lemma A2:* (a) The individual rationality constraint (10) for  $e$  to be a best response for the agent holds if and only if

$$bp(e, \tau) \geq e. \quad (\text{A3})$$

The incentive compatibility constraint (9) holds if and only if

$$b[p(e, \tau) - p(e', \tau)] \geq e - e' \quad (\text{A4})$$

for all  $e' \in \mathcal{E}$ . Let  $\underline{\beta}(0, \tau) = 0$ ; let  $\underline{\beta}(e, \tau) = \sup_{e' \in \mathcal{E} \cup [0, e)} \beta(e, e', \tau)$  if  $e \in \mathcal{E} \setminus 0$ ; and let  $\bar{\beta}(e, \tau) = \inf_{e' \in \mathcal{E} \cup (e, \infty)} \beta(e, e', \tau)$ .

We will show that constraints (A3) and (A4) hold for all  $e' \in \mathcal{E}$  if and only if  $b \in [\underline{\beta}(e, \tau), \bar{\beta}(e, \tau)]$ , where  $\underline{\beta}(e, \tau) \leq \bar{\beta}(e, \tau)$ . If  $e > e'$ , then  $p(e, \tau) > p(e', \tau)$  (because  $p(\cdot, \tau)$  is increasing); therefore, (A4) holds if and only if  $b \geq \beta(e, e', \tau)$ . If  $e < e'$ , then  $p(e, \tau) < p(e', \tau)$ ; therefore, (A4) holds if and only if  $b \leq \beta(e', e, \tau)$ .

If  $e = 0$ , then (A4) holds for all  $e' \in \mathcal{E}$  if and only if  $b \leq \beta(e', e, \tau)$  for all  $e' \in \mathcal{E}$ ; therefore, constraints (A3) and (A4) hold for all  $e' \in \mathcal{E}$  if and only if  $b \in [\underline{\beta}(e, \tau), \bar{\beta}(e, \tau)]$ . Note  $\underline{\beta}(e, \tau) = 0 \leq \bar{\beta}(e, \tau)$ .

If  $e \in \mathcal{E} \setminus 0$ , then (A4) holds for all  $e' \in \mathcal{E}$  if and only if both  $b \leq \beta(e', e, \tau)$  for all  $e' \in \mathcal{E} \cup (e, \infty)$  and  $b \geq \beta(e', e, \tau)$  for all  $e' \in \mathcal{E} \cup [0, e)$ . Further,  $e \in \mathcal{E} \setminus 0$  implies

$$\sup_{e' \in \mathcal{E} \cup [0, e)} \beta(e, e', \tau) \geq \frac{e}{p(e, \tau) - p(0, \tau)} \geq \frac{e}{p(e, \tau)},$$

where the first inequality follows because  $0 \in \mathcal{E} \cup [0, e)$ , and the second inequality follows because  $p(0, \tau) \geq 0$ . Thus, if  $b \geq \sup_{e' \in \mathcal{E} \cup [0, e)} \beta(e, e', \tau)$ , then constraint (A3) holds. We conclude that constraints (A3) and (A4) hold for all  $e' \in \mathcal{E}$  if and only if  $b \in [\underline{\beta}(e, \tau), \bar{\beta}(e, \tau)]$ . It follows from Lemma A1 that  $\underline{\beta}(e, \tau) \leq \bar{\beta}(e, \tau)$ . We conclude that under temperature  $\tau$  and any bonus  $b \in [\underline{\beta}(e, \tau), \bar{\beta}(e, \tau)]$ ,  $e$  is a best response for the agent. Because the principal's utility is decreasing in  $b$ , it follows that  $b^*(e, \tau) = \underline{\beta}(e, \tau)$ .

If  $\mathcal{E} = \{0, e_1, e_2, \dots, e_N\}$ , where  $0 < e_1 < e_2 < \dots < e_N$ , then  $\underline{\beta}(e_i, \tau) = \sup_{e' \in \{0, e_1, e_2, \dots, e_{i-1}\}} \beta(e_i, e', \tau) = \beta(e_i, e_{i-1}, \tau)$  for  $i \in \{1, 2, \dots, N\}$ , where the first equality follows from the definition of  $\underline{\beta}(e, \tau)$  and the second equality follows because  $\beta(e, e', \tau)$  increases in  $e'$  (by Lemma A1).

(b) Suppose  $e \neq \min(\mathcal{E})$ . By inequality (3) with strict supermodularity, for  $e' \in \mathcal{E} \cup [0, e)$ ,  $p(e, \tau) - p(e', \tau)$  increases in  $\tau$ , so  $\beta(e, e', \tau) = \frac{e - e'}{p(e, \tau) - p(e', \tau)}$  decreases in  $\tau$ . It follows that  $\underline{\beta}(e, \tau)$  decreases in  $\tau$ . The result follows because  $b^*(e, \tau) = \sup_{e' \in \mathcal{E} \cup [0, e)} \beta(e, e', \tau)$  (by part (a)). *Q.E.D.*

LEMMA A3. In the benchmark setting where temperature  $\tau$  is exogenous, decentralization distorts downward the effort  $e^*(\tau) \leq e^{FB}(\tau)$  and success probability  $p(e^*(\tau), \tau) \leq p(e^{FB}(\tau), \tau)$ .

*Proof of Lemma A3:* Because the success probability  $p(e, \tau)$  increases in effort  $e$ , it is sufficient to show that  $e^*(\tau) \leq e^{FB}(\tau)$ . By Lemma A2, for any effort  $e \in \mathcal{E}$ , there exists a bonus  $b$  such that  $e$  is a best response for the agent. Recall that  $b^*(e, \tau)$  denotes the principal's optimal bonus that implements effort  $e$  under temperature  $\tau$ . Consider two arbitrary effort levels  $e_M$  and  $e_H$  in  $\mathcal{E}$ , where  $e_M < e_H$ . The proof proceeds in two steps.

First, we establish

$$b^*(e_H, \tau) \geq b^*(e_M, \tau) \quad (\text{A5})$$

$$p(e_H, \tau) \cdot b^*(e_H, \tau) - p(e_M, \tau) \cdot b^*(e_M, \tau) \geq e_H - e_M. \quad (\text{A6})$$

In words, inequality (A5) states the optimal bonus increases in effort, and inequality (A6) states that the increase in expected bonus  $p(e, \tau) \cdot b^*(e, \tau)$  required to induce effort  $e_H$  instead of  $e_M$  exceeds the increase with the agent's cost to exert effort  $e_H$  instead of  $e_M$ . It follows from the incentive compatibility constraint (9) that

$$p(e_H, \tau) \cdot b^*(e_H, \tau) - e_H \geq p(e_M, \tau) \cdot b^*(e_H, \tau) - e_M \quad (\text{A7})$$

$$p(e_M, \tau) \cdot b^*(e_M, \tau) - e_M \geq p(e_H, \tau) \cdot b^*(e_M, \tau) - e_H. \quad (\text{A8})$$

Then

$$[p(e_H, \tau) - p(e_M, \tau)] \cdot b^*(e_H, \tau) \geq e_H - e_M \geq [p(e_H, \tau) - p(e_M, \tau)] \cdot b^*(e_M, \tau), \quad (\text{A9})$$

where the first inequality follows from inequality (A7) and the second follows from inequality (A8). Because  $p(e, \tau)$  increases in  $e$ , it follows that  $p(e_H, \tau) > p(e_M, \tau)$ . These inequalities and inequality (A9) together imply inequality (A5). Note

$$p(e_H, \tau) \cdot b^*(e_H, \tau) - p(e_M, \tau) \cdot b^*(e_M, \tau) \geq \quad (\text{A10})$$

$$p(e_H, \tau) \cdot b^*(e_H, \tau) - p(e_M, \tau) \cdot b^*(e_H, \tau) \geq e_H - e_M, \quad (\text{A11})$$

where the first inequality follows from inequality (A5) and the second inequality follows from inequality (A7). Together, (A10)-(A11) imply (A6).

Second, we show that  $e^*(\tau) \leq e^{FB}(\tau)$ . Suppose to the contrary that  $e^*(\tau) > e^{FB}(\tau)$ . It follows from the definition of  $e^{FB}(\tau)$  that

$$p(e^{FB}(\tau), \tau) \cdot R - e^{FB}(\tau) > p(e^*(\tau), \tau) \cdot R - e^*(\tau),$$

or equivalently

$$e^*(\tau) - e^{FB}(\tau) > [p(e^*(\tau), \tau) - p(e^{FB}(\tau), \tau)] \cdot R. \quad (\text{A12})$$

It follows from the definition of  $e^*(\tau)$  that

$$p(e^*(\tau), \tau) \cdot [R - b^*(e^*(\tau), \tau)] > p(e^{FB}(\tau), \tau) \cdot [R - b^*(e^{FB}(\tau), \tau)],$$

or equivalently

$$[p(e^*(\tau), \tau) - p(e^{FB}(\tau), \tau)] \cdot R > p(e^*(\tau), \tau) \cdot b^*(e^*(\tau), \tau) - p(e^{FB}(\tau), \tau) \cdot b^*(e^{FB}(\tau), \tau). \quad (\text{A13})$$

Together, inequalities (A12) and (A13) imply

$$e^*(\tau) - e^{FB}(\tau) > p(e^*(\tau), \tau) \cdot b^*(e^*(\tau), \tau) - p(e^{FB}(\tau), \tau) \cdot b^*(e^{FB}(\tau), \tau). \quad (\text{A14})$$

With  $e^*(\tau) = e_H$  and  $e^{FB}(\tau) = e_M$ , (A14) contradicts (A6). We conclude that  $e^*(\tau) \leq e^{FB}(\tau)$ . *Q.E.D.*

LEMMA A4. (a) In the benchmark setting where temperature  $\tau$  is exogenous, the optimal effort in the centralized setting  $e^{FB}(\tau)$  weakly increases in the reward  $R$ . (b) In the benchmark setting where temperature  $\tau$  is exogenous, the optimal effort in the decentralized setting  $e^*(\tau)$  weakly increases in  $R$ . (c) The optimal effort in the centralized setting  $e^{FB}$  weakly increases in  $R$ .

*Proof of Lemma A4:* We prove part (c) first. With some abuse of notation, let  $(e^{FB}(R), \tau^{FB}(R))$  denote the optimal effort and temperature in the centralized setting under reward  $R$ ; suppose  $R_L < R_H$ . We will show  $e^{FB}(R_L) \leq e^{FB}(R_H)$ . The proof is by contradiction. Suppose to the contrary that  $e^{FB}(R_L) > e^{FB}(R_H)$ . Note

$$p(e^{FB}(R_L), \tau^{FB}(R_L)) \cdot R_L - e^{FB}(R_L) > p(e^{FB}(R_H), \tau^{FB}(R_H)) \cdot R_L - e^{FB}(R_H) \quad (\text{A15})$$

$$p(e^{FB}(R_H), \tau^{FB}(R_H)) \cdot R_H - e^{FB}(R_H) > p(e^{FB}(R_L), \tau^{FB}(R_L)) \cdot R_L - e^{FB}(R_L), \quad (\text{A16})$$

where (A15) follows from the definition of  $(e^{FB}(R_L), \tau^{FB}(R_L))$  and (A16) follows from the definition of  $(e^{FB}(R_H), \tau^{FB}(R_H))$ . Note

$$e^{FB}(R_H) - e^{FB}(R_L) > [p(e^{FB}(R_H), \tau^{FB}(R_H)) - p(e^{FB}(R_L), \tau^{FB}(R_L))] \cdot R_L \quad (\text{A17})$$

$$[p(e^{FB}(R_H), \tau^{FB}(R_H)) - p(e^{FB}(R_L), \tau^{FB}(R_L))] \cdot R_H > e^{FB}(R_H) - e^{FB}(R_L) \quad (\text{A18})$$

where the first inequality follows from (A15) and the second inequality from (A16). Because  $e^{FB}(R_L) > e^{FB}(R_H)$ , it follows that  $0 > p(e^{FB}(R_H), \tau^{FB}(R_H)) - p(e^{FB}(R_L), \tau^{FB}(R_L)) > 0$ , where the first inequality follows from inequality (A17) and the second inequality follows from (A18). This contradiction implies  $e^{FB}(R_L) \leq e^{FB}(R_H)$ .

Next, we prove part (a). It follows from the argument in part (c), adapted to replace  $\tau^{FB}(R_H)$  and  $\tau^{FB}(R_L)$  with  $\tau$ , that  $e^{FB}$  (weakly) increases in  $R$ .

Finally, we prove part (b). Fix the temperature  $\tau$ . By Lemma A2, for any target effort  $e$  there exists a least bonus  $b^*(e, \tau)$  that makes  $e$  incentive compatible for the agent; importantly,  $b^*(e, \tau)$  depends on the technology  $p(\cdot, \tau)$  and effort costs but does *not* depend on the reward  $R$ . When inducing  $e$ , the principal will never choose a bonus above  $b^*(e, \tau)$ , since any higher  $b$  reduces profit one-for-one. Hence the principal's profit from inducing  $e$  at reward  $R$  equals

$$\Pi(e; R, \tau) = p(e, \tau) R - p(e, \tau) b^*(e, \tau).$$

Let  $R_H > R_L \geq 0$ , and let  $e_H \in \arg \max_e \Pi(e; R_H, \tau)$  and  $e_L \in \arg \max_e \Pi(e; R_L, \tau)$  be optimal induced efforts at  $R_H$  and  $R_L$ , respectively. Optimality gives

$$\Pi(e_H; R_H, \tau) \geq \Pi(e_L; R_H, \tau) \quad \text{and} \quad \Pi(e_L; R_L, \tau) \geq \Pi(e_H; R_L, \tau).$$

Subtract the second inequality from the first to obtain

$$[\Pi(e_H; R_H, \tau) - \Pi(e_H; R_L, \tau)] \geq [\Pi(e_L; R_H, \tau) - \Pi(e_L; R_L, \tau)].$$

Using the profit representation and the fact that  $b^*(e, \tau)$  does not depend on  $R$ ,

$$\Pi(e; R_H, \tau) - \Pi(e; R_L, \tau) = p(e, \tau) (R_H - R_L),$$

so the previous display becomes

$$p(e_H, \tau)(R_H - R_L) \geq p(e_L, \tau)(R_H - R_L).$$

Since  $R_H > R_L$ , we obtain  $p(e_H, \tau) \geq p(e_L, \tau)$ . Because  $p(\cdot, \tau)$  is increasing in effort, this implies  $e_H \geq e_L$ . Thus, the optimal induced effort is (weakly) increasing in  $R$ . *Q.E.D.*

**LEMMA A5.** *The optimal temperature in the centralized setting  $\tau^{FB}$  weakly increases in the reward  $R$ .*

*Proof of Lemma A5:* With some abuse of notation, let  $(e^{FB}(R), \tau^{FB}(R))$  denote the optimal effort and temperature in the centralized setting under reward  $R$ ; suppose  $R_L < R_H$ . We will show that  $\tau^{FB}(R_H) \geq \tau^{FB}(R_L)$ . The proof is by contradiction. Suppose to the contrary that  $\tau^{FB}(R_H) < \tau^{FB}(R_L)$ . The centralized system's expected utility  $\Pi(e, \tau)$  depends on  $\tau$  only through the success probability  $p(e, \tau)$  and is increasing in the success probability. It follows that  $\tau^{FB} = \arg \max_{\tau \in \mathcal{T}} \{p(e, \tau)\}$ . Because  $e^{FB}(R_H) \geq e^{FB}(R_L)$  (by Lemma A4(c)) and  $\tau^{FB}(R_H) < \tau^{FB}(R_L)$  (by assumption), it follows from inequality (3) that

$$p(e^{FB}(R_H), \tau^{FB}(R_L)) - p(e^{FB}(R_L), \tau^{FB}(R_L)) > p(e^{FB}(R_H), \tau^{FB}(R_H)) - p(e^{FB}(R_L), \tau^{FB}(R_H)), \quad (\text{A19})$$

Hence,

$$0 \geq p(e^{FB}(R_L), \tau^{FB}(R_H)) - p(e^{FB}(R_L), \tau^{FB}(R_L)) > p(e^{FB}(R_H), \tau^{FB}(R_H)) - p(e^{FB}(R_H), \tau^{FB}(R_L)) \geq 0,$$

where the first inequality follows because  $\tau^{FB}(R_L) \in \arg \max_{\tau \in \mathcal{T}} p(e^{FB}(R_L), \tau)$ , the second inequality follows from inequality (A19), and the third inequality follows because  $\tau^{FB}(R_H) \in \arg \max_{\tau \in \mathcal{T}} p(e^{FB}(R_H), \tau)$ . The contradiction implies  $\tau^{FB}(R_H) \geq \tau^{FB}(R_L)$ . *Q.E.D.*

## Appendix B: Proofs of Technical Results

Because Proposition 1A follows immediately from Proposition 1B, it suffices to prove Proposition 1B. The proofs rely on a standard reduction: for any target effort  $e$  and temperature  $\tau$ , the principal chooses the minimal bonus  $b^*(e, \tau)$  that satisfies incentive compatibility and participation. The next lemma records the resulting expression for  $b^*(e, \tau)$  when effort is discrete. It is used in the proofs of Propositions 1B to 4B.

**LEMMA B1.** *If the effort action space is discrete  $\mathcal{E} = \{e_L, e_M, e_H\}$ , then  $b^*(e_L, \tau) = 0$ ,  $b^*(e_M, \tau) = \beta(e_M, e_L, \tau)$ , and  $b^*(e_H, \tau) = \beta(e_H, e_M, \tau)$ . If  $\mathcal{E} = \{e_L, e_H\}$ , then  $b^*(e_L, \tau) = 0$  and  $b^*(e_H, \tau) = \beta(e_H, e_L, \tau)$ .*

*Proof of Lemma B1.* Suppose  $\mathcal{E} = \{e_L, e_M, e_H\}$ . Because  $e_L = 0$ , it follows by Lemma A2 that  $b^*(e_L, \tau) = 0$ . From the discrete part of Lemma A2(a),  $b^*(e_M, \tau) = \beta(e_M, e_L, \tau)$  and  $b^*(e_H, \tau) = \max\{\beta(e_H, e_M, \tau), \beta(e_H, e_L, \tau)\}$ . Because  $\beta(e_H, e', \tau)$  increases in  $e'$  (by Lemma A1), it follows that  $b^*(e_H, \tau) = \beta(e_H, e_M, \tau)$ . The proof for the case where  $\mathcal{E} = \{e_L, e_H\}$  follows by parallel argument. *Q.E.D.*

*Proof of Proposition 1B.* Let  $\bar{\varepsilon} = \min_{i \in \{1, 2, 3, 4\}} \{\varepsilon_i\}$ , where  $\varepsilon_i$  is defined later in the proof for  $i \in \{1, 2, 3, 4\}$ . The proof proceeds in four parts. First, we show that (11)-(12) satisfy our assumptions. Second, we show that the conditions in the lemma imply  $(e^{FB}, \tau^{FB}) = (e_M, \tau_L)$ . Third, we show that the conditions in the lemma imply  $(e^*, \tau^*) = (e_H, \tau_H)$ . Fourth, we show that  $p(e^*, \tau^*) > p(e^{FB}, \tau^{FB})$ .

First, we show that (11)-(12) satisfy our assumptions. Let  $\varepsilon_1 = \alpha$  and  $\varepsilon_2 = \Delta/2$ ; note inequality (13) implies  $\varepsilon_2 > 0$ ; and recall  $\varepsilon \in (0, \min\{\varepsilon_1, \varepsilon_2\})$ . Because  $\varepsilon \in (0, \Delta)$ , it follows that  $p(e, \tau)$  is concave, increasing in  $e$  for  $\tau \in \{\tau_H, \tau_L\}$ . Consider inequality (3) where  $(\bar{\tau}, \underline{\tau}) = (\tau_H, \tau_L)$

$$p(\bar{e}, \tau_H) - p(\underline{e}, \tau_H) > p(\bar{e}, \tau_L) - p(\underline{e}, \tau_L). \quad (\text{B1})$$

Note  $\varepsilon > 0$  implies inequality (B1) where  $(\bar{e}, \underline{e}) = (e_M, e_L)$ ;  $\varepsilon < \Delta/2$  implies inequality (B1) where  $(\bar{e}, \underline{e}) = (e_H, e_M)$ . Together, these imply inequality (B1) where  $(\bar{e}, \underline{e}) = (e_H, e_L)$ .

Second, we show that the conditions in the lemma imply  $(e^{FB}, \tau^{FB}) = (e_M, \tau_L)$ . Note  $(e^{FB}, \tau^{FB}) = (e_M, \tau_L)$  if and only if

$$\begin{aligned} p(e_M, \tau_L) \cdot R - e_M &> p(e_H, \tau_L) \cdot R - e_H \\ p(e_M, \tau_L) \cdot R - e_M &> p(e_M, \tau_H) \cdot R - e_M \\ p(e_M, \tau_L) \cdot R - e_M &> p(e_H, \tau_H) \cdot R - e_H \\ p(e_M, \tau_L) \cdot R - e_M &> \max(p(e_L, \tau_L), p(e_L, \tau_H)) \cdot R, \end{aligned}$$

which holds if and only if

$$\min\left(\frac{1}{\Delta - \alpha}, \frac{1}{\varepsilon}\right) > \frac{R}{e_M} > \frac{1}{\Delta} \text{ and } \alpha > 0. \quad (\text{B2})$$

The centralized system's profit is greater under  $(e_M, \tau_L)$  than  $(e_H, \tau_L)$  if and only if  $1/\varepsilon > R/e_M$ , is greater under  $(e_M, \tau_L)$  than  $(e_M, \tau_H)$  if and only if  $\varepsilon < \alpha$ , is greater under  $(e_M, \tau_L)$  than  $(e_H, \tau_H)$  if and only if  $1/(\Delta - \alpha) > R/e_M$ , is greater under  $(e_M, \tau_L)$  than  $(e_L, \tau_L)$  if and only if  $R/e_M > 1/\Delta$ , and is greater under  $(e_M, \tau_L)$  than  $(e_L, \tau_H)$  if and only if  $R/e_M > 1/(\Delta + \alpha)$ . Let  $\varepsilon_3 = \Delta - \alpha$ . Note that inequalities (13)-(14) and  $\varepsilon < \Delta - \alpha$  imply that the inequalities in (B2) hold, which implies  $(e^{FB}, \tau^{FB}) = (e_M, \tau_L)$ .

Third, we show that the conditions in the lemma imply  $(e^*, \tau^*) = (e_H, \tau_H)$ . Because  $b^*(e_H, \tau) = \beta(e_H, e_M, \tau)$ ,  $b^*(e_M, \tau) = \beta(e_M, e_L, \tau)$ , and  $b^*(e_L, \tau) = 0$  (by Lemma B1), it follows that  $(e^*, \tau^*) = (e_H, \tau_H)$  if and only if

$$\begin{aligned} \pi(e_H, \tau_H, \beta(e_H, e_M, \tau_H)) &> \pi(e_H, \tau_L, \beta(e_H, e_M, \tau_L)) \\ \pi(e_H, \tau_H, \beta(e_H, e_M, \tau_H)) &> \pi(e_M, \tau_L, \beta(e_M, e_L, \tau_L)) \\ \pi(e_H, \tau_H, \beta(e_H, e_M, \tau_H)) &> \pi(e_M, \tau_H, \beta(e_M, e_L, \tau_H)) \\ \pi(e_H, \tau_H, \beta(e_H, e_M, \tau_H)) &> \max(\pi(e_L, \tau_L, 0), \pi(e_L, \tau_H, 0)), \end{aligned}$$

which holds if

$$\Delta - \alpha - \varepsilon > 0 \text{ and } \varepsilon < \Delta/2 \quad (\text{B3})$$

$$\frac{R}{e_M} > \frac{2\Delta\varepsilon + \varepsilon^2 - \Delta^2 - \alpha(\Delta - \varepsilon)}{(\Delta - \alpha - \varepsilon)(\Delta - \varepsilon)\varepsilon} \quad (\text{B4})$$

$$\frac{R}{e_M} > \frac{(\Delta - \alpha)(\Delta + \varepsilon) + 2\alpha\varepsilon}{(\Delta - \alpha)(\Delta - \varepsilon)\Delta} \quad (\text{B5})$$

$$\frac{R}{e_M} > \frac{\Delta + \varepsilon}{(\Delta - \varepsilon)^2} \quad (\text{B6})$$

$$\frac{R}{e_M} > \frac{2\Delta}{(2\Delta - \alpha)(\Delta - \varepsilon)}. \quad (\text{B7})$$

In the limit, as  $\varepsilon \rightarrow 0$ , inequalities (B3)-(B7) hold if and only if

$$\Delta > \alpha \text{ and } \frac{R}{e_M} > \frac{2}{2\Delta - \alpha}. \quad (\text{B8})$$

Therefore, there exists  $\varepsilon_4 > 0$  such that if inequality (B8) holds and  $\varepsilon < \varepsilon_4$ , then inequalities (B3)-(B7) hold. Note that inequalities (13)-(14) imply the inequalities in (B8). Therefore, inequalities (13)-(14) and  $\varepsilon < \varepsilon_4$  imply that inequalities (B3)-(B7) hold, which implies  $(e^*, \tau^*) = (e_H, \tau_H)$ .

Fourth, we show that  $p(e^*, \tau^*) > p(e^{FB}, \tau^{FB})$ . Note  $p(e^*, \tau^*) = p(e_H, \tau_H) = 2\Delta > \alpha + \Delta = p(e_M, \tau_L) = p(e^{FB}, \tau^{FB})$ , where the first equality follows from step three, the last equality follows from step two, and the inequality follows from inequality (13).  $Q.E.D.$

*Proof of Lemma 1:* From the centralized system's optimization problem (5), it follows that  $\tau^{FB}(e) \in \arg \max_{\tau \in \mathcal{T}} p(e, \tau)$ . The proof of the inequality  $\tau^*(e) \geq \tau^{FB}(e)$  is by contradiction. Suppose to the contrary that  $\tau^*(e) < \tau^{FB}(e)$ . Under temperature  $\tau \in \{\tau^*(e), \tau^{FB}(e)\}$ , there exists a bonus  $b$  such that  $e$  is a best response (by Lemma A2(a)). Note that the principal's optimal bonus  $b^*(e, \tau) \leq R$ ; further,  $p(e, \tau) > 0$  for all  $e > 0$ . Because  $b^*(e, \tau)$  decreases in  $\tau$  (by Lemma A2(b)) and  $\tau^*(e) < \tau^{FB}(e)$ , it follows that  $b^*(e, \tau^*(e)) > b^*(e, \tau^{FB}(e))$ . Note

$$p(e, \tau^{FB}(e)) \cdot [R - b^*(e, \tau^{FB}(e))] \geq p(e, \tau^*(e)) \cdot [R - b^*(e, \tau^{FB}(e))] \quad (\text{B9})$$

$$> p(e, \tau^*(e)) \cdot [R - b^*(e, \tau^*(e))]. \quad (\text{B10})$$

where inequality (B9) follows because  $\tau^{FB}(e) \in \arg \max_{\tau \in \mathcal{T}} p(e, \tau)$  and inequality (B10) follows because  $b^*(e, \tau^*(e)) > b^*(e, \tau^{FB}(e))$ . Inequalities (B9) and (B10) imply that the principal's expected utility is higher under temperature  $\tau = \tau^{FB}(e)$  than  $\tau = \tau^*(e)$ , a contradiction.  $Q.E.D.$

Because Proposition 2A follows immediately from Proposition 2B, it is sufficient to provide proof of only the latter.

*Proof of Proposition 2B.* We begin by establishing some definitions and preliminaries. Note  $e_H - e_M = e_M - e_L = e_M$ , where the first equality holds because effort is evenly spaced and the second equality holds because  $e_L = 0$ . Let  $R_1 = e_M/\Delta$ ,

$$R_3(\varepsilon) = \frac{\alpha/\delta - \Delta/(\Delta - \varepsilon)}{\alpha + \delta - 2\Delta + \varepsilon} e_M,$$

$R_2(\varepsilon) = R_3(\varepsilon) - \psi(\varepsilon)$  and  $R_4(\varepsilon) = R_3(\varepsilon) + \psi(\varepsilon)$ , where  $\psi(\varepsilon)$  is specified subsequently. Recall  $b^*(e, \tau)$  denotes the principal's optimal bonus under effort  $e$  and temperature  $\tau$ , i.e., the solution to (8)-(10) when  $(e, \tau)$  is exogenous. With some abuse of notation, let  $\pi(e, \tau) = p(e, \tau)[R - b^*(e, \tau)]$  and  $\pi(e) = \pi(e, \tau^*(e))$ . Throughout, we consider  $\varepsilon \in (0, \min(\delta, \Delta/2, \Delta(1 - \delta/\alpha)))$ .

The remainder of the proof proceeds in six parts. First, we establish inequality (15) and  $R_3(\varepsilon) > 0$ . Second, we show that (16)-(17) satisfy our assumptions. Third, we show that  $R = R_3(\varepsilon)$  implies  $\pi(e_M, \tau_L) = \pi(e_H, \tau_H)$ . Fourth, we show that there exist  $\varepsilon \in (0, \min(\delta, \Delta/2, \Delta(1 - \delta/\alpha), \Delta - \delta))$ , and  $\psi(\varepsilon) > 0$  such that for  $R \in (R_2(\varepsilon), R_4(\varepsilon))$ , (i)  $\tau^*(e_M) = \tau_L$ , (ii)  $\tau^*(e_H) = \tau_H$ , (iii)  $p(e_H, \tau^*(e_H)) < p(e_M, \tau^*(e_M))$ , and (iv)  $\max(\pi(e_H), \pi(e_M)) > \pi(e_L)$ . Fifth, we show that there exist  $\varepsilon \in (0, \min(\delta, \Delta/2, \Delta(1 - \delta/\alpha), \Delta - \delta))$  and  $\psi(\varepsilon) > 0$  such that (v)  $e^*|_{R \in (R_2(\varepsilon), R_3(\varepsilon))} = e_H$  and  $e^*|_{R \in (R_3(\varepsilon), R_4(\varepsilon))} = e_M$  and (vi)  $b^*(e^*, \tau^*(e^*))|_{R \in (R_2(\varepsilon), R_3(\varepsilon))} < b^*(e^*, \tau^*(e^*))|_{R \in (R_3(\varepsilon), R_4(\varepsilon))}$ . Sixth, we show that if  $R \in (0, R_1)$ , then  $e^* = e_L$ .

First, we establish inequality (15) and  $R_3(\varepsilon) > 0$ . It follows from  $\alpha/2 + \delta^2/(\alpha + \delta) < (\alpha + \delta)/2$  that  $\delta < \alpha$ . This together with the second inequality in (18) implies  $\Delta < \alpha$ . It remains to show that  $\delta < \Delta$ . It follows from  $\delta < \alpha$  that  $\delta < \alpha/2 + \delta^2/(\alpha + \delta)$ . This together with the first inequality in (15) implies  $\delta < \Delta$ . To see that  $R_3(\varepsilon) > 0$ , note that  $\varepsilon < \Delta(1 - \delta/\alpha)$  implies that the numerator is positive, and the second inequality in (18) and  $\varepsilon > 0$  imply that the denominator of  $R_3(\varepsilon)$  is positive.

Second, we show that (16)-(17) satisfy our assumptions. Because  $0 < \varepsilon < \delta < \alpha$ , it follows that  $p(e, \tau_L)$  is concave, increasing in  $e$ . Because  $0 < \varepsilon < \Delta$ , it follows that  $p(e, \tau_H)$  is concave, increasing in  $e$ . Consider inequality (3) where  $(\bar{\tau}, \underline{\tau}) = (\tau_H, \tau_L)$

$$p(\bar{e}, \tau_H) - p(\underline{e}, \tau_H) > p(\bar{e}, \tau_L) - p(\underline{e}, \tau_L) \quad (\text{B11})$$

Note  $\Delta > \delta$  implies inequality (B11) where  $(\bar{e}, \underline{e}) = (e_M, e_L)$ ;  $\varepsilon < \Delta/2$  implies inequality (B11) where  $(\bar{e}, \underline{e}) = (e_H, e_M)$ . Together, these imply inequality (B11) where  $(\bar{e}, \underline{e}) = (e_H, e_L)$ .

Third, we show that  $R = R_3(\varepsilon)$  implies  $\pi(e_M, \tau_L) = \pi(e_H, \tau_H)$ . By Lemma B1,  $b^*(e_M, \tau_L) = \beta(e_M, e_L, \tau_L) = e_M/\delta$ , and  $b^*(e_H, \tau_H) = \beta(e_H, e_M, \tau_H) = e_M/(\Delta - \varepsilon)$ . It then follows from  $R = R_3(\varepsilon)$  that

$$\pi(e_M, \tau_L) = (\alpha + \delta) \cdot (R_3(\varepsilon) - e_M/\delta) = (2\Delta - \varepsilon) \cdot (R_3(\varepsilon) - e_M/(\Delta - \varepsilon)) = \pi(e_H, \tau_H).$$

Fourth, we show that there exist  $\varepsilon \in (0, \min(\delta, \Delta/2, \Delta(1 - \delta/\alpha), \Delta - \delta))$ , and  $\psi(\varepsilon) > 0$  such that for  $R \in (R_2(\varepsilon), R_4(\varepsilon))$ , (i)  $\tau^*(e_M) = \tau_L$ , (ii)  $\tau^*(e_H) = \tau_H$ , (iii)  $p(e_H, \tau^*(e_H)) < p(e_M, \tau^*(e_M))$ , and (iv)  $\max(\pi(e_H), \pi(e_M)) > \pi(e_L)$ . Because  $p(e_H, \tau^*(e_H))$ ,  $p(e_M, \tau^*(e_M))$ , and  $\pi(e_i)$  for  $i \in \{H, M, L\}$  are continuous in  $\varepsilon$  and  $R$ , it is sufficient to show that for  $R = R_3(\varepsilon)$ , in the limit as  $\varepsilon \rightarrow 0$ , the following hold: (i)  $\pi(e_M, \tau_L) > \pi(e_M, \tau_H)$ , (ii)  $\pi(e_H, \tau_H) > \pi(e_H, \tau_L)$ , (iii)  $2\Delta - \varepsilon < \alpha + \delta$ , and (iv)  $\pi(e_M, \tau_L) = \pi(e_H, \tau_H) > \pi(e_L)$ . The second inequality in (18) implies condition (iii). Condition (i) holds if and only if

$$(\alpha + \delta) \cdot (R_3(0) - e_M/\delta) > \Delta \cdot (R_3(0) - e_M/\Delta),$$

which holds because  $\Delta > \delta$ . By Lemma B1,  $b^*(e_H, \tau_L) = \beta(e_H, e_M, \tau_L) = e_M/\varepsilon$ . It follows that  $\lim_{\varepsilon \rightarrow 0} \pi(e_H, \tau_L) = -\infty$ . Therefore, to establish condition (ii) it is sufficient to show that  $\pi(e_H, \tau_H) > 0$ . This follows because  $\Delta > \delta$  implies  $R_3(0) > e_M/\Delta$ . The equality in condition (iv) was established in Step 3. Note  $b^*(e_L, \tau) = 0$  for  $\tau \in (\tau_L, \tau_H)$  (by Lemma B1), so  $\pi(e_L) = \alpha R_3(0)$ . Therefore, the inequality in condition (iv) holds because

$$2\Delta \cdot (R_3(0) - e_M/\Delta) > \alpha R_3(0),$$

which follows from the first inequality in (18).

Fifth, we show that there exist  $\varepsilon \in (0, \min(\delta, \Delta/2, \Delta(1 - \delta/\alpha), \Delta - \delta))$  and  $\psi(\varepsilon) > 0$  such that (v)  $e^*|_{R \in (R_2(\varepsilon), R_3(\varepsilon))} = e_H$  and  $e^*|_{R \in (R_3(\varepsilon), R_4(\varepsilon))} = e_M$  and (vi)  $b^*(e^*, \tau^*(e^*))|_{R \in (R_2(\varepsilon), R_3(\varepsilon))} < b^*(e^*, \tau^*(e^*))|_{R \in (R_3(\varepsilon), R_4(\varepsilon))}$ . It follows from Step 4 that there exist  $\varepsilon \in (0, \min(\delta, \Delta/2, \Delta(1 - \delta/\alpha), \Delta - \delta))$ , and  $\psi(\varepsilon) > 0$  such that for  $R \in (R_2(\varepsilon), R_4(\varepsilon))$ ,  $(\partial/\partial R)[\pi(e_M, \tau_L) - \pi(e_H, \tau_H)] = p(e_M, \tau^*(e_M)) - p(e_H, \tau^*(e_H)) > 0$ . Result (v) then follows from Step 3. Note  $b^*(e^*, \tau^*(e^*))|_{R \in (R_2(\varepsilon), R_3(\varepsilon))} = b^*(e_H, \tau^*(e_H)) = b^*(e_H, \tau_H) = \beta(e_H, e_M, \tau_H) = e_M/(\Delta - \varepsilon)$ , where the first equality follows by result (v), the second equality follows by Step 4's condition (ii), and the third equality follows by Lemma B1. Similarly,  $b^*(e^*, \tau^*(e^*))|_{R \in (R_3(\varepsilon), R_4(\varepsilon))} =$

$b^*(e_M, \tau^*(e_M)) = b^*(e_M, \tau_L) = \beta(e_M, e_L, \tau_L) = e_M/\delta$ , where the first equality follows by result (v), the second equality follows by Step 4's condition (i), and the third equality follows by Lemma B1. Thus,  $b^*(e^*, \tau^*(e^*))|_{R \in (R_2(\varepsilon), R_3(\varepsilon))} < b^*(e^*, \tau^*(e^*))|_{R \in (R_3(\varepsilon), R_4(\varepsilon))}$  if and only if  $\varepsilon < \Delta - \delta$ .

Sixth, we show that if  $R \in (0, R_1)$ , then  $e^* = e_L$ . By Lemma B1,  $b^*(e_M, \tau_L) = \beta(e_M, e_L, \tau_L) = e_M/\delta$ ,  $b^*(e_H, \tau_L) = \beta(e_H, e_M, \tau_L) = e_M/\varepsilon$ ,  $b^*(e_M, \tau_H) = \beta(e_M, e_L, \tau_H) = e_M/\Delta$ , and  $b^*(e_H, \tau_H) = \beta(e_H, e_M, \tau_H) = e_M/(\Delta - \varepsilon)$ . It follows from  $0 < \varepsilon < \delta < \Delta$  (where the last inequality follows by Step 1), that  $\min_{e \in \{e_M, e_H\}, \tau \in \{\tau_L, \tau_H\}} b^*(e, \tau) = e_M/\Delta$ . Recall  $R_1 = e_M/\Delta$ . It follows that if  $R \in (0, R_1)$ , then  $\max_{e \in \{e_M, e_H\}, \tau \in \{\tau_L, \tau_H\}} \pi^*(e, \tau) < 0$ . Because  $\pi^*(e_L, \tau_L) = \alpha R > 0 = \pi^*(e_L, \tau_H)$ , it follows that  $(e^*, \tau^*) = (e_L, \tau_L)$ . *Q.E.D.*

Because Proposition 3A follows immediately from Proposition 3B, it is sufficient to provide proof of only the latter.

*Proof of Proposition 3B.* We prove the two parts in reverse order.

(b) The proof proceeds in two steps. First, we show that  $\tau^*(e_j) > \tau^{FB}(e_j)$  and  $(\partial/\partial\tau)p(e_j, \tau)|_{\tau=\tau^*(e_j)} < 0$  for  $j \in \{1, 2, \dots, N\}$ . Note  $\tau^{FB}(e_j) \in \arg\max_{\tau \in [\tau_L, \tau_H]} \{p(e_j, \tau) \cdot R - e_j\}$ , which implies  $\tau^{FB}(e_j) \in \arg\max_{\tau \in [\tau_L, \tau_H]} p(e_j, \tau)$ . Because  $p(e, \tau)$  is concave in  $\tau$  and  $(\partial/\partial\tau)p(e, \tau)|_{\tau=\tau_H} < 0 \leq (\partial/\partial\tau)p(e, \tau)|_{\tau=\tau_L}$ , it follows that  $(\partial/\partial\tau)p(e_j, \tau^{FB}(e_j)) = 0$ . Therefore, to establish that  $\tau^*(e_j) > \tau^{FB}(e_j)$ , it is sufficient to show that  $(\partial/\partial\tau)p(e_j, \tau^*(e_j)) < 0$ . If  $\tau^*(e_j) = \tau_H$ , the result is immediate. Suppose for the remainder that  $\tau^*(e_j) < \tau_H$ . Because  $b^*(e_i, \tau) = \beta(e_i, e_{i-1}, \tau)$  for  $i \in \{1, 2, \dots, N\}$  (by Lemma A2(a)), it follows that  $\tau^*(e_j) \in \arg\max_{\tau \in [\tau_L, \tau_H]} \pi(e_j, \tau, \beta(e_j, e_{j-1}, \tau))$ , where, by convention,  $e_0 = 0$ . Note  $(\partial/\partial\tau)\pi(e_j, \tau, \beta(e_j, e_{j-1}, \tau))|_{\tau=\tau^*(e_j)} = [R - \beta(e_j, e_{j-1}, \tau^*(e_j))](\partial/\partial\tau)p(e_j, \tau^*(e_j)) - p(e_j, \tau^*(e_j)) \cdot (\partial/\partial\tau)\beta(e_j, e_{j-1}, \tau^*(e_j)) = 0$ . Because  $(\partial/\partial\tau)\beta(e_j, e_{j-1}, \tau) < 0$  (by inequality (2)), it follows that  $(\partial/\partial\tau)p(e_j, \tau^*(e_j)) < 0$ .

Second, we show that  $(\partial/\partial R)\tau^*(e_j; R) < 0$  for  $j \in \{1, 2, \dots, N\}$ , where we generalize the notation for  $\tau^*$  to denote the dependence on  $R$ . With some abuse of notation let  $\pi(e_j, \tau; R) = p(e_j, \tau) \cdot [R - b^*(e_j, \tau)]$ . Restrict attention to  $\tau \in (\tau^{FB}(e_j), \tau_H]$ , which contains every optimizer  $\tau^*(e_j; R)$  by Step 1. Because  $p(e, \tau)$  is concave in  $\tau$ , it follows that on this interval

$$\frac{\partial p(e_j, \tau)}{\partial \tau} < 0.$$

Further,  $b^*(e_j, \tau)$  is independent of  $R$ . Hence,

$$\frac{\partial^2}{\partial \tau \partial R} \pi(e_j, \tau; R) = \frac{\partial p(e_j, \tau)}{\partial \tau} < 0.$$

Thus, the objective function  $\pi$  satisfies strict decreasing differences in  $(\tau, R)$  on the relevant domain. It follows from the Strict Monotonicity Theorem (Edlin and Shannon 1998) that the optimal selection  $\tau^*(e_j; R)$  decreases in  $R$  on  $(R_j, \bar{R}_j)$ .

(a) The proof proceeds in five parts.

First, we define some expressions that are useful in subsequent steps. We generalize the expression for the principal's expected utility in equation (6) to  $\pi(e, \tau, b; R)$  to denote the dependence on the reward  $R$ . Let  $\tau^*(e; R)$  denote the principal's expected-utility-maximizing temperature under effort  $e$ , reward  $R$ , and bonus

$b^*(e, \tau^*(e; R))$ , where recall that  $b^*(e, \tau)$  is the optimal bonus which induces effort  $e$  under temperature  $\tau$ . Let  $\underline{R} = \inf\{R : \pi(e_H, \tau^*(e_H; R), b^*(e_H, \tau^*(e_H; R)); R) > \pi(e_L, \tau^*(e_L; R), b^*(e_L, \tau^*(e_L; R)); R)\}$ .

Second, we show that  $b^*(e_H, \tau^*(e_H; R)) > 0$  and  $\tau^*(e_L; R) = \tau^{FB}(e_L)$  for all  $R$ ,

$$\underline{R} = \inf\{R : \pi(e_H, \tau^*(e_H; R), b^*(e_H, \tau^*(e_H; R)); R) > \pi(e_L, \tau^{FB}(e_L), 0; R)\}, \quad (\text{B12})$$

and  $\underline{R} \in (0, \infty)$ . Because  $e_H > e_L = 0$ , it follows that under bonus  $b = 0$  and any temperature  $\tau \in [\tau_L, \tau_H]$ , the agent's best response effort  $e = e_L$ ; it follows that  $b^*(e_H, \tau^*(e_H; R)) > 0$ . Because  $b^*(e_L, \tau) = 0$  for all  $\tau \in \mathcal{T}$  (by Lemma B1) and  $\pi(e_L, \tau, 0) = \Pi(e_L, \tau)$ , it follows that  $\tau^*(e_L; R) = \tau^{FB}(e_L)$  and  $b^*(e_L, \tau^*(e_L; R)) = 0$  for all  $R$ . Thus,

$$\pi(e_L, \tau^*(e_L; R), b^*(e_L, \tau^*(e_L; R)); R) = \pi(e_L, \tau^{FB}(e_L), 0; R). \quad (\text{B13})$$

Inequality (B12) follows. It remains to show  $\underline{R} \in (0, \infty)$ . Let  $\bar{R} = (e_H - e_L)/[p(e_H, \tau^{FB}(e_L)) - p(e_L, \tau^{FB}(e_L))]^2$ . Because  $p(e, \tau)$  strictly increases in  $e$ , it follows that  $\bar{R} < \infty$ . To establish that  $\underline{R} < \infty$  it is sufficient to show that

$$\pi(e_H, \tau^*(e_H; R), b^*(e_H, \tau^*(e_H; R)); R) > \pi(e_L, \tau^{FB}(e_L), 0; R) \quad (\text{B14})$$

for  $R > \bar{R}$ . It follows from the definition of  $\tau^*(e_H; R)$  that  $\pi(e_H, \tau^*(e_H; R), b^*(e_H, \tau^*(e_H; R)); R) \geq \pi(e_H, \tau^{FB}(e_L), b^*(e_H, \tau^{FB}(e_L)); R)$ . Therefore,

$$\begin{aligned} & \pi(e_H, \tau^*(e_H; R), b^*(e_H, \tau^*(e_H; R)); R) - \pi(e_L, \tau^{FB}(e_L), 0; R) \\ & \geq \pi(e_H, \tau^{FB}(e_L), b^*(e_H, \tau^{FB}(e_L)); R) - \pi(e_L, \tau^{FB}(e_L), 0; R) \\ & = [p(e_H, \tau^{FB}(e_L)) - p(e_L, \tau^{FB}(e_L))]R - \beta(e_H, e_L, \tau^{FB}(e_L)), \end{aligned} \quad (\text{B15})$$

where equality (B15) follows because  $b^*(e_H, \tau^{FB}(e_L)) = \beta(e_H, e_L, \tau^{FB}(e_L))$  (by Lemma B1) and  $b^*(e_L, \tau) = 0$  (by Lemma B1). Because  $\beta(e_H, e_L, \tau^{FB}(e_L)) < \infty$  and  $p(e, \tau)$  strictly increases in  $e$ , it follows that the quantity on the right hand side of (B15) is strictly positive for  $R > \bar{R}$ . This implies that inequality (B14) holds for  $R > \bar{R}$ . We conclude  $\underline{R} < \infty$ . Because  $\pi(e_H, \tau^*(e_H; 0)), b^*(e_H, \tau^*(e_H; 0); 0) = -b^*(e_H, \tau^*(e_H; 0); 0) < 0 = \pi(e_L, \tau^{FB}(e_L), 0; 0)$ , it follows that  $\underline{R} > 0$ .

Third, we show the following: for  $R < \underline{R}$ ,  $\tau^* = \tau^{FB}(e_L)$ ; for  $R > \underline{R}$ ,  $\tau^* = \tau^*(e_H; R)$ ;  $\tau^*$  is invariant to  $R$  for  $R < \underline{R}$ ; and  $\tau^*$  decreases in  $R$  on  $R > \underline{R}$ . It follows from the definition of  $\underline{R}$  that if  $R < \underline{R}$ , then  $(e^*, \tau^*) = (e_L, \tau^*(e_L; R))$ ; because  $\tau^*(e_L; R) = \tau^{FB}(e_L)$  (by Step 2) and  $\tau^{FB}(e_L) \in \arg \max_{\tau \in [\tau_L, \tau_H]} p(e_L, \tau)$  (as noted in Step 1 of the proof of part (b)), it follows that  $\tau^{FB}(e_L)$  is invariant to  $R$ . From part (b),  $\tau^*(e_H; R)$  decreases in  $R$ .

It remains to show that if  $R > \underline{R}$ , then  $\tau^* = \tau^*(e_H; R)$ . Suppose to the contrary that there exists  $R_H > \underline{R}$  such that  $\tau^* \neq \tau^*(e_H; R_H)$ . It follows that under reward  $R_H$ , the optimal effort  $e^* = e_L$  and the principal's expected profit is given by equality (B13). Thus,

$$\pi(e_L, \tau^{FB}(e_L), 0; R_H) > \pi(e_H, \tau^*(e_H; R_H), b^*(e_H, \tau^*(e_H; R_H)); R_H). \quad (\text{B16})$$

It follows from equality (B12) that there exists  $R_L \in [\underline{R}, R_H)$  such that

$$\pi(e_H, \tau^*(e_H; R_L), b^*(e_H, \tau^*(e_H; R_L)); R_L) > \pi(e_L, \tau^{FB}(e_L), 0; R_L). \quad (\text{B17})$$

Further,

$$b^*(e_H, \tau^*(e_H; R_L)) < [p(e_H, \tau^*(e_H; R_L)) - p(e_L, \tau^{FB}(e_L))]R_L \quad (\text{B18})$$

$$< [p(e_H, \tau^*(e_H; R_L)) - p(e_L, \tau^{FB}(e_L))]R_H, \quad (\text{B19})$$

where inequality (B18) follows from inequality (B17) and where inequality (B19) follows because  $R_L < R_H$ . Further, inequality (B19) can be rewritten as

$$\pi(e_H, \tau^*(e_H; R_L), b^*(e_H, \tau^*(e_H; R_L)); R_H) > \pi(e_L, \tau^{FB}(e_L), 0; R_H). \quad (\text{B20})$$

It follows from the definition of  $(\tau^*(e; R), b^*(e, \tau^*(e; R)))$  that

$$\pi(e_H, \tau^*(e_H; R_H), b^*(e_H, \tau^*(e_H; R_H)); R_H) \geq \pi(e_H, \tau^*(e_H; R_L), b^*(e_H, \tau^*(e_H; R_L)); R_H). \quad (\text{B21})$$

Together inequalities (B20) and (B21) imply

$$\pi(e_H, \tau^*(e_H; R_H), b^*(e_H, \tau^*(e_H; R_H)); R_H) > \pi(e_L, \tau^{FB}(e_L), 0; R_H), \quad (\text{B22})$$

which contradicts (B16). We conclude for  $R > R_L$ ,  $\tau^* = \tau^*(e_H; R)$ .

Fourth, we show that  $\tau^{FB}(e_L) < \tau^{FB}(e_H)$ . As noted in the proof of Step 1 of part (b),  $\tau^{FB}(e_i) \in \arg \max_{\tau \in [\tau_L, \tau_H]} p(e_i, \tau)$ , where  $i \in \{L, H\}$ . Because  $p(e, \tau)$  is concave in  $\tau$  and  $(\partial/\partial\tau)p(e, \tau)|_{\tau=\tau_H} < 0 \leq (\partial/\partial\tau)p(e, \tau)|_{\tau=\tau_L}$ , it follows that  $(\partial/\partial\tau)p(e_i, \tau^{FB}(e_i)) = 0$  for  $i$ . The proof is by contradiction; suppose  $\tau^{FB}(e_L) \geq \tau^{FB}(e_H)$ . Note

$$0 = (\partial/\partial\tau)p(e_L, \tau^{FB}(e_L)) \leq (\partial/\partial\tau)p(e_L, \tau^{FB}(e_H)) < (\partial/\partial\tau)p(e_H, \tau^{FB}(e_H)) = 0, \quad (\text{B23})$$

where the weak inequality holds because  $p(e, \tau)$  is concave in  $\tau$  and  $\tau^{FB}(e_L) \geq \tau^{FB}(e_H)$  and the strict inequality follows from inequality (1). It follows from the contradiction in (B23) that  $\tau^{FB}(e_L) < \tau^{FB}(e_H)$ .

Fifth, we show that  $\lim_{R \uparrow R_L} \tau^* < \lim_{R \downarrow R_L} \tau^*$ . Note  $\lim_{R \uparrow R_L} \tau^* = \tau^{FB}(e_L) < \tau^{FB}(e_H) < \tau^*(e_H; R) = \lim_{R \downarrow R_L} \tau^*$ , where the equalities follow by Step 3, the first inequality follows by Step 4, and the second inequality follows by Step 1 of part (b). *Q.E.D.*

Because Proposition 4A follows immediately from Proposition 4B, it is sufficient to provide proof of only the latter.

*Proof of Proposition 4B.* The proof proceeds in three steps. First, we show that  $U(\tau^{FB})$  is zero. Second, we show that  $U(\tau^*)$  is strictly positive. Third, we show that  $\tau^* > \tau^{FB}$ . For the purposes of this proof, we refer to the threshold  $\bar{\varepsilon}$  in the statement of Proposition 1B as  $\hat{\varepsilon}$ ; we define  $\bar{\varepsilon}$  in the statement of Proposition 4B as  $\bar{\varepsilon} = \min(\hat{\varepsilon}, \Delta, \tilde{\varepsilon})$ , where  $\tilde{\varepsilon} = (\alpha/2 + \Delta)\Delta/(\alpha + \Delta)$ .

First, we show that  $U(\tau^{FB})$  is zero. Because  $b(e_L, \tau_L) = 0$  (by Lemma B1), it is sufficient to show that  $e^*(\tau^{FB}), \tau^{FB}) = (e_L, \tau_L)$ . The proof of Proposition 1B establishes that  $\varepsilon < \hat{\varepsilon}$  implies  $\tau^{FB} = \tau_L$ . It remains to show that  $\varepsilon < \bar{\varepsilon}$  implies  $e^*(\tau_L) = e_L$ . Note  $e^*(\tau_L) = e_L$  if and only if

$$p(e_L, \tau_L) \cdot R > p(e_M, \tau_L) \cdot (R - b(e_M, \tau_L)) \quad (\text{B24})$$

$$p(e_L, \tau_L) \cdot R > p(e_H, \tau_L) \cdot (R - b(e_H, \tau_L)). \quad (\text{B25})$$

Note  $b(e_M, \tau_L) = \beta(e_M, e_L, \tau_L)$  and  $b(e_H, \tau_L) = \beta(e_H, e_M, \tau_L)$  (by Lemma B1). Thus, inequality (B24) holds if and only if  $R/e_M < (\alpha + \Delta)/\Delta^2$ , which holds by assumption. Further, inequality (B25) holds if and only if  $R/e_M < g(\varepsilon)$ , where  $g(\varepsilon) = (\alpha + \Delta + \varepsilon)/[(\Delta + \varepsilon)\varepsilon]$ . Because  $g(\cdot)$  is decreasing and  $R/e_M < (\alpha + \Delta)/\Delta^2 < g(\bar{\varepsilon})$ , it follows that  $R/e_M < g(\varepsilon)$  for  $\varepsilon < \bar{\varepsilon}$ . This establishes that  $e^*(\tau_L) = e_L$ .

Second, we show that  $U(\tau^*)$  is strictly positive. Note  $U(\tau^*) = p(e_H, \tau_H) \cdot b(e_H, \tau_H) - e_H = 2\varepsilon e_M / (\Delta - \varepsilon) > 0$ , where the first equality follows because  $(e^*(\tau^*), \tau^*) = (e_H, \tau_H)$  (as established in the proof of Proposition 1B), the second equality follows because  $b(e_H, \tau_H) = \beta(e_H, e_M, \tau_H)$  (by Lemma B1), and the inequality follows because  $\varepsilon \in (0, \Delta)$ .

Third, we show that  $\tau^* > \tau^{FB}$ . This follows because  $\tau^* = \tau_H$  (as established in Step 2) and  $\tau^{FB} = \tau_L$  (as established in Step 1).  $\text{Q.E.D.}$

## Appendix C: Characterizing Inferior Technology Choice

We characterize the necessary and sufficient conditions for the principal to optimally choose the inferior production technology in the setting where the action space consists of two effort levels and two temperatures. Without loss of generality, let  $\mathcal{E} = \{e_H, e_L\}$ ,  $\mathcal{T} = \{\tau_H, \tau_L\}$ ,  $p(e_L, \tau_L) = \Gamma$ ,  $p(e_H, \tau_L) = \Gamma + \Delta$ ,  $p(e_L, \tau_H) = \gamma$ , and  $p(e_H, \tau_H) = \gamma + \delta$ , where  $\Gamma \geq 0$ ,  $\gamma \geq 0$ ,  $\Delta > 0$ , and  $\delta > 0$ . It is straightforward to verify that if the low temperature technology is inferior, then it is never optimal for the principal to choose the inferior production technology. Suppose instead that the high temperature technology is inferior  $p(e, \tau_H) < p(e, \tau_L)$  for  $e \in \mathcal{E}$ , which holds if and only if  $\Gamma > \gamma$  and  $\Gamma + \Delta > \gamma + \delta$ . It is optimal for the principal to choose the inferior production technology  $\tau^* = \tau_H$  if and only if

$$\max\left(\frac{1}{\delta}, \frac{\gamma + \delta}{\delta(\gamma + \delta - \Gamma)}\right) < \frac{R}{e_H} < \frac{\delta\Gamma - \Delta\gamma}{\delta\Delta(\Gamma + \Delta - \gamma - \delta)}. \quad (\text{C1})$$

In the special case where  $\gamma = 0$ , the condition (C1) simplifies to

$$\frac{1}{\delta - \Gamma} < \frac{R}{e_H} < \frac{\Gamma}{\Delta(\Gamma + \Delta - \delta)}. \quad (\text{C2})$$

As  $\Delta \rightarrow 0$ , the rightmost term in (C2) approaches infinity. Thus, we conclude that condition (C1) tends to hold when  $\Delta$  and  $\gamma$  are small and  $\delta$  is large, meaning that (i) the marginal effect of effort on the success probability is large when the temperature is high and is small when the temperature is low and (ii) under high temperature, failure occurs with high probability if the agent exerts low effort.

The argument below establishes that condition (C1) is necessary and sufficient for  $\tau^* = \tau_H$ . First, note that if the principal's optimal effort  $e^* = e_L$ , then  $\gamma < \Gamma$  implies  $\tau^* = \tau_L$ . Therefore,  $\tau^* = \tau_H$  requires that  $e^* = e_H$ . It remains to show that the principal's expected profit is maximized under effort and temperature  $(e_H, \tau_H)$ . It follows from  $1/\delta < R/e_H$  that the principal's expected profit is greater under  $(e_H, \tau_H)$  than  $(e_L, \tau_H)$ . It follows from  $(\gamma + \delta)/[\delta(\gamma + \delta - \Gamma)] < R/e_H$  that the principal's expected profit is greater under  $(e_H, \tau_H)$  than  $(e_L, \tau_H)$ . It follows from the last inequality in (C1) that the principal's expected profit is greater under  $(e_H, \tau_H)$  than  $(e_H, \tau_L)$ .

## Appendix D: Uniqueness Conditions for the Optimal Temperature $\tau^*(e)$

We consider how the principal can implement a given effort level  $e$ . The principal selects temperature  $\tau$  and bonus  $b \geq 0$  to maximize expected utility:

$$\Pi(e, \tau; b) = p(e, \tau)[R - b],$$

subject to the agent's incentive compatibility (IC) and individual rationality (IR) constraints. Given  $e$  and  $\tau$ , the principal selects the minimal feasible bonus  $b^*(e, \tau)$ , which simplifies the optimization problem to:

$$\max_{\tau} \Pi(e, \tau) = p(e, \tau)[R - b^*(e, \tau)].$$

Previous analysis establishes that, for any fixed  $e > 0$ , the minimal bonus  $b^*(e, \tau)$  required to implement effort  $e$  is decreasing in  $\tau$ ; see Lemma A2(b) in Appendix A.

We examine two scenarios:

**Case 1:** If  $p(e, \tau)$  is monotonically increasing in  $\tau$ , the optimal temperature is the boundary solution:

$$\tau^*(e) = \tau_{\max},$$

which is clearly unique.

**Case 2:** When  $p(e, \tau)$  is concave in  $\tau$ , inducing effort  $e$  in the continuous environment exactly satisfies the agent's IC constraint:

$$\frac{\partial p(e, \tau)}{\partial e} b^*(e, \tau) = 1.$$

Thus, the principal's objective simplifies to:

$$\max_{\tau \in \mathcal{T}} \Pi(e, \tau) = R p(e, \tau) - \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e}.$$

**LEMMA D1.** *A sufficient condition for the strict negativity of the second-order condition, ensuring uniqueness of the optimal temperature  $\tau^*(e)$ , is:*

$$\frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} \right) \geq 0.$$

*Proof of Lemma D1.* Differentiating the principal's objective function twice with respect to temperature yields:

$$\frac{\partial^2 \Pi(e, \tau)}{\partial \tau^2} = R \frac{\partial^2 p(e, \tau)}{\partial \tau^2} - \frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} \right).$$

By assumption, strict concavity of  $p(e, \tau)$  in  $\tau$  implies

$$\frac{\partial^2 p(e, \tau)}{\partial \tau^2} < 0,$$

making the first term strictly negative.

Given the condition stated in the lemma:

$$\frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} \right) \geq 0,$$

the second term is non-negative. This ensures strict negativity of the second-order condition, guaranteeing uniqueness of the optimal temperature  $\tau^*(e)$ . *Q.E.D.*

This condition implies that the ratio of success probability to marginal productivity of effort

$$\frac{p(e, \tau)}{\partial p(e, \tau) / \partial e}$$

is (weakly) convex in temperature. Intuitively, this convexity captures the deterioration of incentive-provision efficiency at an accelerating rate as temperature rises, ensuring uniqueness and stability of the optimal temperature in the principal's problem.

To illustrate this condition, consider the following example:

**EXAMPLE D1.** Suppose  $p(e, \tau) = e^\alpha \tau^\beta$  with  $\alpha > 0$  and  $0 < \beta \leq 1$ .

1.  $\frac{\partial^2 p(e, \tau)}{\partial \tau^2} = \beta(\beta - 1)e^\alpha \tau^{\beta-2}$ , which is negative if  $0 < \beta < 1$  and zero if  $\beta = 1$ .
2.  $\frac{\partial^2 p(e, \tau)}{\partial e \partial \tau} = \alpha \beta e^{\alpha-1} \tau^{\beta-1} \geq 0$ , which is clearly satisfied.
3.  $\frac{\partial p(e, \tau)}{\partial \tau} \frac{\partial p(e, \tau)}{\partial e} - p(e, \tau) \frac{\partial^2 p(e, \tau)}{\partial e \partial \tau} = 0$ , marginally satisfying the condition.

Notably,

$$\frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} = \frac{1}{\alpha},$$

which is independent of  $\tau$ , so the sufficient condition holds.

When  $\beta = 1$ ,  $p(e, \tau) = e^\alpha \tau$ , so  $\frac{\partial^2 p(e, \tau)}{\partial \tau^2} \equiv 0$  and

$$\frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} \right) \equiv 0.$$

Hence the principal's objective is linear in  $\tau$ . The maximizer is therefore attained at a boundary point of  $\mathcal{T}$ , so  $\tau^*(e)$  is unique even though the second-order condition is weak rather than strictly negative.

**COROLLARY D1.** *For separable probability functions where  $p(e, \tau) = f(e) \cdot g(\tau)$ , the sufficient condition always holds.*

This follows because  $\frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} = \frac{f(e)}{f'(e)}$ , which is independent of  $\tau$ .

**PROPOSITION D1.** *For the non-separable probability function:*

$$p(e, \tau) = \omega \cdot \tau + (1 - \omega) \cdot f(e) \cdot g(\tau), \quad \omega \in (0, 1),$$

a sufficient condition for the uniqueness of the optimal temperature is that  $\frac{\tau}{g(\tau)}$  is convex in  $\tau$ .

*Proof of Proposition D1.* From Corollary D1, we know that separable probability functions guarantee uniqueness. For this non-separable case, we must verify the condition:

$$\frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} \right) \geq 0.$$

Computing  $p_e(e, \tau) = (1 - \omega)f'(e)g(\tau)$ , we can express the ratio as:

$$\begin{aligned} \frac{p(e, \tau)}{\partial p(e, \tau) / \partial e} &= \frac{\omega \cdot \tau + (1 - \omega) \cdot f(e) \cdot g(\tau)}{(1 - \omega)f'(e)g(\tau)} \\ &= \frac{\omega}{(1 - \omega)f'(e)} \cdot \frac{\tau}{g(\tau)} + \frac{f(e)}{f'(e)}. \end{aligned}$$

Taking the second derivative with respect to  $\tau$ :

$$\begin{aligned}\frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau)/\partial e} \right) &= \frac{\partial^2}{\partial \tau^2} \left[ \frac{\omega}{(1-\omega)f'(e)} \cdot \frac{\tau}{g(\tau)} + \frac{f(e)}{f'(e)} \right] \\ &= \frac{\omega}{(1-\omega)f'(e)} \cdot \frac{\partial^2}{\partial \tau^2} \left( \frac{\tau}{g(\tau)} \right).\end{aligned}$$

Since  $\frac{\omega}{(1-\omega)f'(e)} > 0$  for  $\omega \in (0, 1)$  and  $f'(e) > 0$ , the sign of the second derivative depends solely on  $\frac{\partial^2}{\partial \tau^2} \left( \frac{\tau}{g(\tau)} \right)$ .

For the uniqueness condition to hold, we require:

$$\frac{\partial^2}{\partial \tau^2} \left( \frac{p(e, \tau)}{\partial p(e, \tau)/\partial e} \right) \geq 0 \iff \frac{\partial^2}{\partial \tau^2} \left( \frac{\tau}{g(\tau)} \right) \geq 0.$$

This is equivalent to requiring that  $\frac{\tau}{g(\tau)}$  be convex in  $\tau$ . Q.E.D.

The convexity of  $\frac{\tau}{g(\tau)}$  implies that the marginal effectiveness of temperature in providing incentives decreases at an increasing rate. Below are two examples:

EXAMPLE D2. Consider the linear probability function:

$$g(\tau) = a\tau + b(1 - \tau),$$

where  $a < b$  and  $\tau \in [0, 1]$ .

The ratio  $\frac{\tau}{g(\tau)} = \frac{\tau}{(a-b)\tau+b}$  has the following derivatives:

$$\begin{aligned}\frac{d}{d\tau} \left[ \frac{\tau}{(a-b)\tau+b} \right] &= \frac{b}{[(a-b)\tau+b]^2} \\ \frac{d^2}{d\tau^2} \left[ \frac{\tau}{(a-b)\tau+b} \right] &= \frac{-2b(a-b)}{[(a-b)\tau+b]^3}\end{aligned}$$

Since  $a < b$ , we have  $(a-b) < 0$ , making the second derivative strictly positive throughout the domain  $\tau \in [0, 1]$ .

EXAMPLE D3. Consider the rational probability function

$$g(\tau) = \frac{1}{1+k\tau},$$

where  $k > 0$  and  $\tau \in [0, 1]$ .

For this function, the ratio  $\frac{\tau}{g(\tau)} = \tau \cdot (1+k\tau) = \tau + k\tau^2$  has the second derivative

$$\frac{d^2}{d\tau^2} [\tau + k\tau^2] = 2k.$$

Since  $k > 0$ , the second derivative is constant and strictly positive for all  $\tau \in [0, 1]$ .

EXAMPLE D4. Let  $p(e, \tau) = \omega\tau + (1-\omega)f(e)g(\tau)$  where  $\omega \in (0, 1)$ ,  $f$  is differentiable with  $f'(e) > 0$ , and

$$g(\tau) = \tau^\alpha(1-\tau)^\beta \quad \text{with } \alpha \geq 1, \beta \in (0, 1).$$

Then  $g(\tau) \in (0, 1)$  for all  $\tau \in (0, 1)$ , and  $g(0) = g(1) = 0$ . In particular,  $p(e, \tau) \in [0, 1]$  for all  $(e, \tau) \in [0, 1]^2$  (under the maintained bounds on  $f(e)$  used in this appendix). Moreover, for any fixed  $e$  and  $\tau \in (0, 1)$ ,

$$\frac{p(e, \tau)}{\partial p(e, \tau)/\partial e} = \frac{\omega}{1-\omega} \cdot \frac{\tau}{g(\tau)} \cdot \frac{1}{f'(e)} + \frac{f(e)}{f'(e)}.$$

Hence convexity of  $\tau/g(\tau)$  in  $\tau$  implies convexity of  $p(e, \tau)/(\partial p(e, \tau)/\partial e)$  in  $\tau$ . In the benchmark case  $\alpha = 1$ ,

$$\frac{\tau}{g(\tau)} = (1-\tau)^{-\beta} \quad \text{and} \quad \frac{d^2}{d\tau^2} \left( \frac{\tau}{g(\tau)} \right) = \beta(\beta+1)(1-\tau)^{-\beta-2} > 0 \quad \text{for all } \tau \in (0, 1),$$

so  $\tau/g(\tau)$  is strictly convex on  $(0, 1)$ .

We have established sufficient conditions for the uniqueness of the optimal temperature in our principal–agent framework. The key insight is that the convexity of the ratio  $\frac{p(e, \tau)}{\partial p(e, \tau)/\partial e}$  with respect to temperature ensures a unique solution to the principal’s optimization problem. In the special case of separable probability functions, uniqueness is automatically guaranteed. For non-separable cases, we require the convexity of  $\frac{\tau}{g(\tau)}$ , which has been verified for certain functional forms through analytical and numerical methods.

## Appendix E: Continuous Effort and Temperature

This appendix provides numerical examples with continuous effort and continuous temperature that parallel the discrete results in the main body. Throughout we use the quadratic class

$$p(e, \tau) = (1 - \tau)(c_L + a_L e - b_L e^2) + \tau(c_H + a_H e - b_H e^2), \quad e, \tau \in [0, 1], \quad (\text{E1})$$

which is concave in effort when  $b_L, b_H > 0$  and strictly supermodular when  $(a_H - a_L) + 2(b_L - b_H)e > 0$  for all  $e \in [0, 1]$ . For each example we report  $(e^{FB}, \tau^{FB}; p^{FB})$  and  $(e^*, \tau^*, b^*; p^*)$  and verify  $0 \leq p(e, \tau) \leq 1$  on  $[0, 1]^2$ .

EXAMPLE E1 (UPWARD DISTORTION). Let

$$(c_L, a_L, b_L; c_H, a_H, b_H) = (0.2980, 0.2601, 0.3924; 0.0747, 0.3949, 0.1129)$$

and take  $R = 38.95$ . These parameters satisfy strict supermodularity ( $a_H - a_L = 0.1348 > 0$  and  $b_L > b_H$ ). Solving the centralized and decentralized problems yields

$$(e^{FB}, \tau^{FB}; p^{FB}) = (0.300, 0.000; 0.341), \quad (e^*, \tau^*, b^*; p^*) = (1.000, 1.000, 5.909; 0.357).$$

Thus decentralization raises both effort and temperature, and increases success probability. Moreover  $0.075 \leq p(e, \tau) \leq 0.357$  on  $[0, 1]^2$ .

EXAMPLE E2 (UPWARD DISTORTION). Let

$$(c_L, a_L, b_L; c_H, a_H, b_H) = (0.2960, 0.6510, 0.7193; 0.0571, 0.7361, 0.1134)$$

and take  $R = 3$ . Strict supermodularity again holds ( $a_H - a_L = 0.0850 > 0$  and  $b_L > b_H$ ). The computed solutions are

$$(e^{FB}, \tau^{FB}; p^{FB}) = (1.000, 1.000; 0.680), \quad (e^*, \tau^*, b^*; p^*) = (0.000, 0.000, 0.000; 0.296).$$

Hence decentralization reduces effort, temperature, and success probability; here  $0.057 \leq p(e, \tau) \leq 0.680$  on  $[0, 1]^2$ .

EXAMPLE E3 (NONMONOTONE  $\tau^*(R)$ ). Let  $p(e, \tau) = (1 - \tau)(c_L + a_L e - b_L e^2) + \tau(c_H + a_H e - b_H e^2)$  for  $(e, \tau) \in [0, 1]^2$ , and take

$$(c_L, a_L, b_L; c_H, a_H, b_H) = (0.5246, 1.0971, 0.6478; 0.0196, 1.8994, 0.9473).$$

Then  $p(e, \tau) \in [0, 1]$  for all  $(e, \tau) \in [0, 1]^2$ ,

$$\frac{\partial^2 p(e, \tau)}{\partial e^2} = -2((1 - \tau)b_L + \tau b_H) < 0, \quad \frac{\partial^2 p(e, \tau)}{\partial e \partial \tau} = (a_H - a_L) + 2(b_L - b_H)e > 0 \text{ for all } e \in [0, 1],$$

so  $p$  is strictly supermodular in  $(e, \tau)$ . For  $e > 0$ , an interior contract that implements effort  $e$  uses the bonus

$$b(e, \tau) = \frac{1}{\partial p(e, \tau)/\partial e}$$

and yields principal payoff

$$\Phi(e, \tau; R) \equiv p(e, \tau) \left( R - \frac{1}{\partial p(e, \tau)/\partial e} \right).$$

At the boundary  $e = 0$  the interior formula does not apply: the principal can always set  $b = 0$ , which induces  $e = 0$  and yields the shutdown payoff  $\Pi^0(\tau; R) \equiv p(0, \tau)R$ . Hence, for each  $\tau$  we must compare  $\max_{e>0} \Phi(e, \tau; R)$  to  $\Pi^0(\tau; R)$ . Restricting attention to  $\tau \in \{0, 1\}$ , the resulting maximized payoffs cross twice. Numerically, there exist cutoffs  $\underline{R} \approx 3.584$  and  $R^\times \approx 29.389$  such that

$$\tau^*(R) = \begin{cases} 0, & R \leq \underline{R}, \\ 1, & \underline{R} < R < R^\times, \\ 0, & R \geq R^\times. \end{cases}$$

Thus  $\tau^*(R)$  is nonmonotone in  $R$  once the shutdown option  $(b, e) = (0, 0)$  is included in the principal's problem. Representative optima are: (i) at  $R = 1$ ,  $(\tau^*, e^*, b^*; p^*) = (0, 0, 0; 0.525)$ ; (ii) at  $R = 10$ ,  $(\tau^*, e^*, b^*; p^*) \approx (1, 0.693, 1.706; 0.881)$ ; (iii) at  $R = 50$ ,  $(\tau^*, e^*, b^*; p^*) \approx (0, 0.617, 3.353; 0.955)$ . The first-best (full-information) choices satisfy  $\tau^{FB} = 0$  and are  $(e^{FB}; p^{FB}) \approx (0.075; 0.603)$  at  $R = 1$ ,  $(0.770; 0.985)$  at  $R = 10$ , and  $(0.831; 0.989)$  at  $R = 50$ .

**EXAMPLE E4 (UPWARD DISTORTION WITH AGENT BENEFIT).** Consider

$$(c_L, a_L, b_L; c_H, a_H, b_H) = (0.40, 0.50, 1.00; 0.00, 1.00, 0.60).$$

The centralized solution satisfies  $6p_e(e, 0) = 1$ , yielding  $(e^{FB}, \tau^{FB}) = (1/6, 0)$  and  $p^{FB} \approx 0.456$ . Under decentralization,  $(e, \tau) = (0.40, 1)$  gives  $p = 0.304$  and  $p_e = 0.520$ , so the least implementing bonus is  $b^* = 1/p_e \approx 1.923$  and the agent's rent is  $U = pb^* - e \approx 0.185 > 0$ . The principal's payoff at  $(0.40, 1)$  is  $\Phi \approx 1.24$ , which exceeds the value at  $(e^{FB}, 0)$ ; thus the profit-maximizing choice has  $\tau^* = 1 > \tau^{FB}$ , and the agent strictly benefits.

## Appendix F: Numerical Illustration of Proposition 2B

This appendix provides a numerical illustration of Proposition 2B, where the reward  $R$  interval is extended to be continuous. Fix parameters

$$\alpha = 0.60, \quad \delta = 0.20, \quad \Delta = 0.38, \quad \varepsilon = 0.05,$$

and effort levels  $e_L = 0$ ,  $e_M = 1$ , and  $e_H = 2$ . Substituting into the success probabilities in Proposition 2B yields

$$p(e_L, \tau_L) = 0.60, \quad p(e_M, \tau_L) = 0.80, \quad p(e_H, \tau_L) = 0.85,$$

$$p(e_L, \tau_H) = 0, \quad p(e_M, \tau_H) = 0.38, \quad p(e_H, \tau_H) = 0.71.$$

For any  $(e, \tau)$ , the minimal bonus that implements effort  $e$  satisfies

$$b^*(e, \tau) = \max_{e' < e} \frac{e - e'}{p(e, \tau) - p(e', \tau)}. \tag{F1}$$

Hence,

$$\begin{aligned} b^*(e_L, \tau_L) &= b^*(e_L, \tau_H) = 0, & b^*(e_M, \tau_L) &= \frac{1}{0.20} = 5, & b^*(e_M, \tau_H) &= \frac{1}{0.38} \approx 2.63, \\ b^*(e_H, \tau_H) &= \max\left\{\frac{2}{0.71}, \frac{1}{0.71 - 0.38}\right\} = \frac{1}{0.33} \approx 3.03, & b^*(e_H, \tau_L) &= \max\left\{\frac{2}{0.85 - 0.60}, \frac{1}{0.85 - 0.80}\right\} = 20. \end{aligned}$$

Given  $(e, \tau)$ , the principal's payoff from paying the minimal bonus is

$$\pi(e, \tau; R) = p(e, \tau)[R - b^*(e, \tau)]. \quad (\text{F2})$$

Therefore,

$$\begin{aligned} \pi(e_L, \tau_L; R) &= 0.60R, & \pi(e_L, \tau_H; R) &= 0, \\ \pi(e_M, \tau_L; R) &= 0.80(R - 5) = 0.80R - 4, & \pi(e_M, \tau_H; R) &\approx 0.38(R - 2.63) \approx 0.38R - 1.00, \\ \pi(e_H, \tau_H; R) &\approx 0.71(R - 3.03) \approx 0.71R - 2.15, & \pi(e_H, \tau_L; R) &= 0.85(R - 20) = 0.85R - 17. \end{aligned}$$

$R_1$  is the smallest bonus required to induce positive effort:  $R_1 = \frac{1}{0.38} \approx 2.63 = b^*(e_M, \tau_H)$ . Hence, for  $R < R_1$ , every positive-effort policy satisfies  $R - b^*(e, \tau) < 0$  and yields negative payoff, so the principal chooses  $(e_L, \tau_L)$ . The next threshold is defined by indifference between  $(e_L, \tau_L)$  and  $(e_H, \tau_H)$ :

$$0.60R = 0.71(R - 3.03) \Rightarrow R_2 = \frac{2.15}{0.11} \approx 19.56.$$

Thus,  $(e_H, \tau_H)$  dominates  $(e_L, \tau_L)$  for  $R > R_2$ . Indifference between  $(e_H, \tau_H)$  and  $(e_M, \tau_L)$  yields

$$0.71(R - 3.03) = 0.80(R - 5) \Rightarrow R_3 = \frac{1.85}{0.09} \approx 20.54,$$

so  $(e_H, \tau_H)$  is optimal on  $(R_2, R_3)$  among these candidates, and  $(e_M, \tau_L)$  dominates  $(e_H, \tau_H)$  for  $R > R_3$ .

Finally, indifference between  $(e_M, \tau_L)$  and  $(e_H, \tau_L)$  solves

$$0.80(R - 5) = 0.85(R - 20) \Rightarrow R_4 = 260.$$

Although Proposition 2B does not characterize the optimum on  $(R_1, R_2)$  in general, in this numerical instance  $(e_L, \tau_L)$  remains optimal throughout  $(R_1, R_2)$ . Combining the comparisons above, the optimal implemented effort satisfies

$$e^*(R) = \begin{cases} e_L, & 0 < R < R_2, \\ e_H, & R_2 < R < R_3, \\ e_M, & R_3 < R < R_4, \\ e_H, & R > R_4, \end{cases}$$

so effort is nonmonotone in  $R$  (low–high–moderate–high). In the nonmonotone region  $(R_2, R_3)$  the optimal temperature is high, whereas outside this region the optimal temperature is low.