Managerial Blame and Defensive Adherence: How Organizations Create AI Over-Reliance.

Hossein Nikpayam, Mirko Kremer
Frankfurt School of Finance and Management, 60322 Frankfurt am Main, Germany
h.nikpayam@fs.de, m.kremer@fs.de

Francis de Véricourt
ESMT Berlin
francis.devericourt@esmt.org

We empirically test a theory of AI over/under-reliance in a delegated and AI-assisted decision making environment that endogenizes institutional incentives for decision makers (DMs) to rely on algorithmic recommendations - because algorithms produce case-specific and auditable reference points for managerial judgments of decisions and outcomes, they provide the conditions for managers to blame DMs for algorithm overrides, and DMs to exercise their discretion (to override algorithms) defensively.

We develop a model that captures these dynamics, and provide empirical support for its key behavioral mechanisms and predictions. We find that DMs are able to reap (some) benefits from their private knowledge in the absence of organizational context, but when decisions are judged by a manager, they often fail to override an algorithm when they have more precise evidence to the contrary, i.e., they lower performance in the very informational state that in many settings provides the most compelling reason for keeping the human in the loop in the first place.

Key words: human-machine; over-reliance; delegation; blame; procedural fairness; distributional fairness

1. Introduction

Despite rapid technological advancements and substantial investments, integrating artificial intelligence (AI) into business processes often yields disappointing results (McKinsey & Company 2023). Implementation failures frequently stem from organizational and behavioral issues, as AI is unlikely to fully replace human decision-makers (DMs) but rather redefine their roles. While AI excels at quickly and accurately extracting relevant predictions from large datasets (e.g., medical diagnoses), humans bring cognitive flexibility and contextual knowledge that algorithms lack (Boyacı et al. 2024).

DMs thus need to find the right balance in how much they rely on algorithmic input and may occasionally need to override an algorithmic recommendation. This is not easy because machine-assisted tasks are rarely made in an organizational vacuum, but instead are typically delegated by an authority (henceforth, the manager) who may hold the DM accountable for decision process and outcome. For example, 24% of nurses in a recent survey said they had been prompted by a clinical

algorithm to make choices they believed "*were not in the best interest of patient*", citing fear of disciplinary action for bad clinical outcomes as a reason for their willingness to follow a machine recommendation against their own clinical judgment (Bannon 2023). For obvious reasons, we need to interpret such anecdotes and self-reports with caution. On the one hand, managerial pressure on employees to follow machines might be justified, as a necessary corrective for DMs that would otherwise under-rely on algorithmic advice. On the other hand, anecdotal evidence is consistent with managers who pressure DMs to follow algorithmic recommendations while reserving the option to blame the DM for bad outcomes nevertheless. Such managerial behavior, besides being inconsistent with fundamental principles of fairness, then expands the purpose of DMs relying on algorithmic advice from improving decisions to sheltering from blame. As a result, DMs may follow an algorithm against their better judgment, and even when that choice is not in the manager's best interest.

What then can explain, if not rationalize, such behavioral dynamics between managers and employees, and the role of machines in it? A possible answer can be found from within a standard agency framework, with a slight expansion towards algorithms that not only improve information (and thus decision quality, theoretically) but also subtly impact the informational asymmetry that characterizes agency settings: While neither manager nor DM can know or understand at reasonable cognitive cost how the machine arrives at a recommendation, both parties can observe or verify what the recommendation was. Unlike static rules, algorithmic recommendations provide desirable flexibility that allows for the leverage of human tacit knowledge, but they produce salient, case-specific and auditable reference points for managerial judgment. This can have detrimental consequences when the manager who lacks further decision-relevant information over-weighs machine recommendations in their evaluation of decision process and outcome, providing compelling theoretical rationale for the idea that institutional context introduces a systematic bias toward higher algorithmic adherence. Our study presents a test of this prediction. To properly label decisions as algorithmic *under*- or *over*-reliance (vis-a-vis a theoretical benchmark), and study the mechanisms underlying such miscalibrated behavior, we turn to controlled laboratory conditions that allow for a systematic and controlled variation of institutional context, information structure, and available channels for blame.

We study delegated and algorithm-assisted decision making in an environment that endogenizes institutional incentives for DMs to rely on algorithmic recommendations. Specifically, in our setting a manager delegates to a DM a risky diagnostic decision that is made prior to observing one of two possible states of the world. To improve their belief about the state, the DM applies their expertise and elicit an imperfect "human" signal. The DM is further assisted by a machine which makes an imperfect recommendation, generating a second "machine" signal. Importantly, these signals may align or conflict, and the DM *privately* knows whether their own judgment is more or less reliable

than the machine's recommendation. The manager incurs a loss if the DM's decision does not match the realized state and pays the DM a discretionary bonus *after* observing the outcome.[1]

Notably, to focus on behavioral mechanisms that drive algorithm adherence, our setting removes any incentive misalignment that may arise between the two parties (e.g., due to costly effort) and hence the need for formal contracting solutions to rectify such misalignment. Hence, because delegation in this setting does not affect the DM's behavior under full rationality assumptions, we would need to go beyond standard economic theory if we were to observe adherence- and outcome-dependent manager judgments and suboptimal DM algorithm adherence.

We derive testable predictions from a simple informal contracting model with boundedly rational DMs and managers that care for distributional and procedural fairness. Our key predictions derive from the informational asymmetry that characterizes delegated human-machine decision making: the DM possesses decision-relevant private knowledge, which is neither codifiable as input to a machine algorithm nor reliably observable for (or easy to communicate to) the manager, leaving the manager with the daunting task of inferring the quality of the DM's decision process from *ex-post* observables: the outcome (which depends on decision and luck) and whether the DM followed or overrode the machine signal. As a result of this belief updating process, the manager pays a lower bonus when the DM overrode the machine signal after a bad outcome. By contrast, when the outcome is good, overriding the machine yields little or no additional bonus. The key implication is that, in equilibrium, DMs are prone to aligning decisions with machine recommendations even when they should not.

We design laboratory experiments to test these predictions under controlled conditions. Because our main prediction concerns the effect of delegation on algorithm adherence, we implement a *NoDelegation* baseline in which the manager makes the decision after observing both diagnostic signals, and a *Delegation* treatment in which the manager delegates the decision to the DM. Further, our design includes several features that allow us to provide behavioral mechanisms evidence.

We find that *NoDelegation* outperforms a machine-only benchmark that removes human bias from the decision process entirely, showing that human DMs partially capture the expected decision value from the human signal that (sometimes) allows them to override a less precise machine signal. In contrast, *Delegation* of machine-assisted decisions to a human DM hurts performance, and our data provides mechanisms evidence that is consistent with the behavioral assumptions of our theoretical analyses. Bonus payments are generally outcome- and adherence dependent, and they are consistent with blame - after a bad decision outcome, managers pay lower bonuses when they observe that the DM overrode the machine, even though they know that the DM most likely followed more accurate

[1] In our study, we use financial bonus payments to operationalize a more general construct: manager judgments that are directed specifically at, *and* have real consequences for, the DM. In practice, these judgments and consequences might be foregone promotions, legal or social consequences, or financial penalties.

private information (that the manager cannot observe). As a result, DMs fail to override a machine signal when they have more precise evidence to the contrary, i.e., they lower performance in the very informational state that provides the most compelling reason for keeping the human in the loop!

The main contribution of our study is a novel behavioral mechanism with implications for firms that wish to get the most from human-AI collaboration. Specifically, while algorithmic over- and under-reliance have been linked to cognitive biases (e.g., Logg et al. 2019) we study and characterize bias that emerges from the organizational context. The central results of our study is that, in large parts due to blame-like manager judgments that we predict and find in our data, DM adhere to algorithmic recommendations in ways that ultimately weakens the value of introducing AI. We discuss possible remedies to this issue in Section 8.

2. Literature review

Our study relates and contributes to a thriving behavioral literature that studies human-AI interaction, including the question of when and why DMs override machine recommendations, and when and why such overrides help or hurt decision performance. Given the overwhelming evidence of the conditions under which DMs systematically under-rely or over-rely on machine input, the important question for managers is how to design organizational structures and processes in order to mitigate bias in algorithm-assisted decision making (when it exists). Many studies provide managerial insights on how firms can improve the efficacy of machine-assisted decision making, by changing organizational design variables such as *governance structure* and *decision processes* (Dietvorst et al. 2018, Athey et al. 2020, Ibrahim et al. 2021, Beer et al. 2026), *education* and *training* (Dietvorst et al. 2015, Castelo et al. 2019, Buçinca et al. 2021, Caro and de Tejada Cuenca 2023, Banker and Khetani 2019, Yeomans et al. 2019, Hou and Jung 2021, Lehmann et al. 2022, Snyder et al. 2026, Balakrishnan et al. 2026), the *technology* itself (Grand-Clément and Pauphilet 2024, Fügener et al. 2021, Vasconcelos et al. 2023, Buçinca et al. 2021, Bauer et al. 2023, Bansal et al. 2021), and *incentives* (Castelo et al. 2019, Luong et al. 2020, Feier et al. 2022, Vasconcelos et al. 2023, Albright 2024).

Like this prior research we are interested in how organizational context (in particular, incentives) shapes the adherence behavior of a machine-assisted DM. Our main departure is that we endogenize organizational context: rather than just studying how machine-assisted DMs respond to socio-economic incentives, we study how such incentives arise in the first place, endogenously in the interaction between a DM and the organizational context that shapes the DM's incentives to follow machine recommendations or not.

Related to our study is a very recent surge of theoretical research on the agency problems that arise when AI usage is delegated within organizations: how competence signaling or agent's hidden type can distort adoption and reliance decisions (Weitzner 2024, McLaughlin and Spiess 2025), how

misaligned objectives, for instance tensions between clinical outcomes and private payoffs, generate agency costs (Dai and Singh 2025), and how hidden action, such as unobserved effort, exacerbates incentive misalignment (Dai and Taylor 2025, Guan et al. 2025, Bastani and Cachon 2025). Unlike these theoretical studies (which assume fully rational agents), although aligned with the goal to build a better understanding of performance drivers of AI-assisted decisions in organizations, we study algorithm adherence through a behavioral lens *and* provide an empirical test.

To focus on behavioral mechanisms, we create a setting that eliminates incentive conflict between rational managers and DMs. In particular, eliciting a signal is costless for the DM, and the manager sets the bonus ex post, implying a zero bonus in the rational benchmark. Under these conditions, the manager's adherence-dependent judgments of the DM (operationalized via bonus payments) and the DM's adherence-behavior that we observe in the data is not explicable from the standard principal-agent contract framework with rational actors. Rather, it is driven by behavioral mechanisms and decision biases. To the best of our knowledge, we are the first to study, both theoretically and empirically, how algorithm adherence can arise endogenously from behavioral factors (e.g., blame avoidance), within an organizational context.

Our decision setting—where the DM has real discretion to either follow or override the machine signal, and adherence emerges from organizational context rather than a hard-coded mandate—connects naturally to established research on rules, policies, and routines. In this literature, rules are typically not "mechanically binding" but are followed (or violated) under discretion (Morrison 2006, Gill 2026). Similarly, qualitative work on routine dynamics emphasizes that even seemingly "static" SOPs are enacted interpretively: routines can be sources of flexibility and change, not just inertia, precisely because their performance requires situated judgment (Feldman and Pentland 2003, Salvato and Rerup 2018, Feldman et al. 2021). Together, these streams suggest that the empirically relevant question is often not whether discretion exists (it does), but how organizations create conditions under which discretion is exercised versus suppressed.

Where our study differs from generic policy/routine adherence is in the type of benchmark that an organization introduces with AI, and how it reshapes the evaluation of decisions. What managers can reward/sanction depends on what they can observe and verify (Spencer and Rerup 2024), and they will often rely on observable proxies when process quality is otherwise opaque (Ouchi 1979, Eisenhardt 1989). In turn, accountability and anticipated evaluation can shift DMs toward "defensible" choices that are easier to justify (Tetlock 1985, Lerner and Tetlock 1999) and represent a personally safer option rather than the organization's best option (Artinger et al. 2019, 2025). Our contribution is to show how algorithmic recommendations can become exactly such a defensible, de facto standard: unlike most rules and routines (which are either broad or require interpretive enactment), an AI system produces a prediction that can act as a salient, case-specific, auditable reference point for each

decision task. When the DM's tacit judgment is neither observable nor easily communicable, managers might condition their judgments of the DM on what is legible (the machine's recommendation and outcomes), which may endogenously creates pressure to adhere to the machine's prediction, and collapse intended "human-in-the-loop flexibility" into de facto rigidity.

3. Theory

3.1. Decision making environment

We study human-machine interaction in a setting where a manager delegates to a decision maker (DM) a diagnostic decision $a \in \{R, B\}$. The decision is made under uncertainty about the true state of the world, represented by the random variable Y with possible realizations $y \in \{R, B\}$, and prior probability $\mu_0 \equiv \Pr(Y = R)$. To decide, the DM relies on two imperfect signals about the true state, one from their own expertise and one from the machine. Whether human expertise outperforms the machine depends on the task and the DM's skills, information known only to the DM.

Formally, the DM receives first information $\mathbf{s} = (s_M, s_H, \theta)$, where $s_M \in \{R, B\}$ denotes the machine prediction of precision q_{s_M}, $s_H \in \{R, B\}$ the human expertise signal of precision $q^{\theta}_{s_H}$, and $\theta \in \{hi, lo\}$ indicates whether the DM's signal is of high or low precision. The signals are conditionally independent and informative, and while s_M is public information, s_H is private to the DM. The DM's precision depends on random variable Θ with distribution $\mu_\theta \equiv \Pr(\Theta = hi) = 1 - \Pr(\Theta = lo)$, such that the DM's signal is more precise ($q^{hi}_{s_H} > q_{s_M}$) if $\theta = hi$, and less precise otherwise.

Given $\mathbf{s}$, the DM updates their belief about the true state, $\mu_{\mathbf{s}} \equiv \Pr(Y = R|\mathbf{s})$, using Bayes' rule (see Appendix A.1.1) and chooses action a that maximizes their utility $\mathcal{U}^{dm}$. The manager then receives payoff $r(a, y)$, where $r(a, y) = \overline{r}$ if the DM's decision matches the true state, $a = y$, and $r(a, y) = \underline{r} < \overline{r}$ otherwise. We let $\overline{a}$ (resp. $\underline{a}$) define the decision that maximizes (resp. minimizes) expected payoff $\mathbb{E}_Y[r(a, y)|\mathbf{s}]$, such that $\overline{a} = R$ if $\mu_{\mathbf{s}} \geq 0.5$ and $\overline{a} = B$ otherwise. After observing decision a, true state y, and machine signal s_M, the manager pays the DM a bonus b that maximizes their utility $\mathcal{U}^{ma}$.

To simplify the technical exposition and create an environment best suited to provide sharp and empirically testable predictions, we fix $\mu_0 = 0.5$. In this case, when $\theta = hi$ ($q^{hi}_{s_H} > q_{s_M}$), the DM maximizes expected payoff by following their signal, i.e., $\overline{a} = s_H$. Conversely, when $\theta = lo$, the payoff maximizing decision is to follow the machine signal, $\overline{a} = s_M$. We further assume the machine signal remains valuable and informative.[2] Figure 1 illustrates the timeline of events.

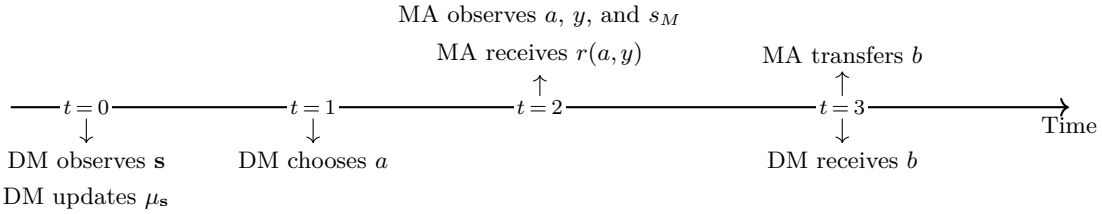
3.2. DM: Stochastic choice.

We assume that the DM has limited cognitive capacity (Luce 2005). Specifically, define A as the random choice variable, and a as its realization. Given signal $\mathbf{s}$, the optimal choice probability is

$$\mathcal{P}(\mathbf{s}) \equiv \Pr(A = \overline{a}|\mathbf{s}) = 1 - \Pr(A = \underline{a}|\mathbf{s}) = \frac{\mathbb{E}_Y[\mathcal{U}^{dm}(\overline{a}, y)|\mathbf{s}]^{1/\beta}}{\mathbb{E}_Y[\mathcal{U}^{dm}(\overline{a}, y)|\mathbf{s}]^{1/\beta} + \mathbb{E}_Y[\mathcal{U}^{dm}(\underline{a}, y)|\mathbf{s}]^{1/\beta}} \tag{1}$$

[2] Equivalently, the human is not much more precise than the machine: $q^{hi}_{s_H} - q_{s_M} \leq \bar{q} < 0.5$.

Figure 1 Timeline of events (MA: manager, DM: decision maker)



where $\mathcal{U}^{dm}$ corresponds to bonus b paid by the manager, and $\beta > 0$ is inversely related to the DM's rationality. As $\beta \to 0$, the model approaches a rational DM; as $\beta \to \infty$, choices become uniformly random. This stochastic framework has strong descriptive appeal (Hey and Orme 1994, Loomes et al. 2002) and enables a precise definition of delegation-driven algorithm reliance (Section 3.6).

3.3. A model of informal contracting under social preferences.

We adopt the Quantal Response Equilibrium (QRE) framework formulated by McKelvey and Palfrey (1995), as the solution concept.[3] The manager chooses b to maximize utility $\mathcal{U}^{ma} \equiv r(a,y) - b$, taking into account that the agent noisily optimizes $\mathcal{U}^{dm} = b$ as per Eq. 1. For the one-shot interaction implemented in our experiments, standard economic theory predicts that self-interested managers will not pay any bonus to the DM, i.e., $b^* = 0$. A large body of fairness literature, however, suggests that this is not a reasonable expectation (Fehr and Schmidt 1999, Bolton et al. 2005). We define the manager's utility using a social preference model, similar to Chassang and Zehnder (2016),[4]

$$\mathcal{U}^{ma} = \underbrace{r(a,y) - b}_{monetary} - |\underbrace{\alpha\big(r(a,y) - 2b\big)}_{distributional} + \underbrace{\gamma\big(\tilde{r}(s_M,a,y) - 2b\big)}_{procedural}|. \tag{2}$$

First, the model accounts for the manager's *distributional* preferences regarding their own realized *ex-post* payoff $r(a,y)$ versus the DM's bonus b: the term $r(a,y) - 2b$ measures the deviation from an equal split of the realized payoff, and $\alpha > 0$ (which is common knowledge, as are all other behavioral parameters) reflects the strength of this preference. Second, the model accounts for *procedural* fairness: the manager's preference to judge the DM not just by the decision outcome $r(a,y)$, but by whether that outcome can be attributed to an *ex-ante* good or bad choice, rather than random luck for which the DM is not accountable. Crucially, procedural fairness requires an assessment of ex-ante decision quality, $\tilde{r}(s_M,a,y)$ (formally defined in the next section), so that the analogous term $\tilde{r}(s_M,a,y) - 2b$ measures the deviation from an equal split based on perceived decision quality rather than realized outcomes; the strength of this preference is governed by $\gamma > 0$. As we discuss in Section 3.4, this assessment depends on whether the DM followed the publicly observable machine signal s_M.[5]

[3] To ensure the existence of stable and well-behaved equilibria, we impose the assumption that $\beta > \underline{\beta}$ (see also the related discussion in Brock and Durlauf 2001).

[4] For expositional ease, we do not model the DM as fairness oriented (see Appendix A.1.4 for this case).

[5] The piecewise-linear specification provides a parsimonious representation of the manager's fairness concerns. The qualitative insights extend to more general fairness-loss functions (Appendix A.1.2).

We impose mild assumptions ($\alpha+\gamma > \frac{1}{2}$ and $\alpha/\gamma \geq \overline{\kappa}$) to align the players' incentives (see Appendix A.1.3 for the unrestricted analysis). The next proposition characterizes the equilibrium decisions of the DM and the manager.

PROPOSITION 1. *In equilibrium, the manager prefers $a = \overline{a}$, pays a bonus $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right)$, and the DM chooses $\overline{a}$ with probability $\mathcal{P}(\boldsymbol{s}) \geq \frac{1}{2}$.*

Proposition 1 shows that the bonus b^* is a convex combination of two yardsticks, the realized outcome $r(a,y)$ and the perceived ex-ante quality $\tilde{r}(s_M,a,y)$, with weights given by the relative strength of the manager's distributional and procedural concerns. The second term is what carries machine adherence into compensation: as $\gamma \to 0$ the bonus responds to outcomes alone, and the adherence effects we characterize below vanish.

3.4. Manager: Beliefs.

In our set-up, we model the manager's assessment of the *ex-ante* quality of the DM's decision as

$$\tilde{r}(s_M,a,y) \equiv \omega(s_M,a,y) \cdot \underbrace{\mathbb{E}_Y[r(a,y) \mid s_M]}_{\text{Exp. payoff given } s_M} = \omega(s_M,a,y) \cdot \sum_{s_H,\theta} \underbrace{\Pr(s_H,\theta \mid s_M,a,y)}_{\substack{\text{MA's belief about} \\ \text{DM's private info}}} \cdot \underbrace{\mathbb{E}_Y[r(a,y) \mid \mathbf{s}]}_{\substack{\text{Exp. payoff given} \\ \text{DM's private info}}}. \quad (3)$$

Here, $\omega(s_M,a,y)$ is a behavioral term that captures how this decision assessment is moderated by what the manager can observe, while the expected payoff term $\mathbb{E}_Y[r(a,y) \mid s_M]$ reflects a rational and fairness-oriented manager's attempt to separate outcomes driven by good decision making from those driven by luck. Importantly, the DM's choice to follow or override the machine signal has two opposing effects on the manager's assessment of the DM's decision, $\mathbb{E}_Y[r(a,y) \mid s_M]$. On the one hand, overriding the machine signal increases the manager's belief that the DM correctly followed a high precision human signal, captured by $\Pr(s_H,\theta \mid s_M,a,y)$.[6] On the other hand, following the machine signal increases the expected payoff $\mathbb{E}_Y[r(a,y) \mid \mathbf{s}]$. The following inequality indicates that the latter effect is stronger (see Proposition EC.1 in Appendix A.1.6 for a formal proof and the intuition driving this result).

$$\mathbb{E}_Y[r(a,y) \mid \underbrace{s_M = a}_{follow}] > \mathbb{E}_Y[r(a,y) \mid \underbrace{s_M \neq a}_{override}]. \quad (4)$$

A rational manager thus assesses the DM's decision more positively when the DM chose to follow the machine signal. Yet, there are plausible behavioral reasons to refine this prediction, based on a rich social psychology literature on *outcome bias* (Baron and Hershey 1988) and *omission-commission effects* (Spranca et al. 1991, Cox et al. 2017). This research provides insights on how people judge

[6] This mirrors the logic of models in career-concerns settings, where high-ability DMs receive more precise signals than the machine and low-ability DMs less precise ones (Weitzner 2024). In such settings, overriding the machine always leads to a higher bonus, as pay depends on the posterior probability of being high-ability.

decisions of others, depending on the outcome as well as (regardless of outcome) on the interpretation of the decision as an *omission* that maintains the status quo or default option or a *commission* that actively changes it (see, e.g., Anderson et al. 2024). In our setting, because the manager can only observe the machine signal and because following this signal is the optimal decision most often ($\Pr(\overline{a} = s_M) > \Pr(\overline{a} = s_H)$), it is natural to interpret that following the machine signal ($a = s_M$) is perceived as an omission, while overriding it ($a \neq s_M$) is perceived as a commission. This interpretation is consistent with prior work suggesting that algorithmic recommendations serve as a reference point for decision making (Froomkin et al. 2019, McLaughlin and Spiess 2025, Dai and Singh 2025).

A consistent finding in the social judgment literature is that acts of commission, i.e., overriding the machine signal, are evaluated more favorably than omissions after good outcomes, but *substantially* more harshly after bad outcomes (Bostyn and Roets 2016, Guglielmo and Malle 2019).[7] Relatedly, negative events trigger stronger sense-making and procedural fairness concerns (Brockner and Wiesenfeld 1996). To formally accommodate such effects, Eq. 3 includes a term $\omega(s_M, a, y) > 0$ that scales the expected payoff accordingly. Without loss of generality, we normalize $\omega(s_M, a = s_M, y) = 1$, set $\omega(s_M, a \neq s_M, y = a) \equiv \omega_g > 1$, and $\omega(s_M, a \neq s_M, y \neq a) \equiv \omega_b < 1$. Thus, $\omega_g > 1$ means that after a good outcome an override is evaluated more favorably than following the machine signal, whereas $\omega_b < 1$ captures the harsher evaluation of overrides after bad outcomes (Bostyn and Roets 2016). Moreover, we assume $(\omega_g - 1)/(1 - \omega_b) < \overline{\omega}$, where $\overline{\omega} < 1$ ensures that the effect for bad outcomes is sufficiently stronger than for good outcomes. This restriction ensures that the harsher evaluation of overrides after bad outcomes dominates the more favorable evaluation of successful overrides in expectation (Guglielmo and Malle 2019).

Overall, relative to the predictions for a fully rational manager as in Eq. 4,[8] omission-commission effects amplify the predictions for bad outcomes: the manager perceives decision quality as (even) higher after observing that the DM followed the machine. For good outcomes, the effects either weaken the prediction (but still favoring "follow"), and might in fact reverse (favoring "override"), which depends on the effect's strength (whether $\omega_g > \omega^*$). The next section formalizes this logic.

3.5. Manager: Bonus Payments.

The next theorem formalizes how bonus payments depend on what the fairness-minded manager (Eq. 2) observes: choice a, state y, and machine signal s_M. Parts A-B show that bonus payments depend

[7] Contributing to this asymmetry, outcomes we label "good" are better understood as baseline or expected (i.e., "nothing went wrong"), and thus need not warrant praise or higher bonuses in many relevant decision settings—for example, a patient survives, a flight lands safely, or a product/service simply works as intended.

[8] These predictions are formalized in Proposition EC.1 (manager's bonus payments) and Lemma EC.9 (DM's machine reliance). Without omission-commission bias (i.e., $\omega \equiv 1$), the manager pays a higher bonus when the DM follows rather than overrides the machine signal, thereby inducing greater overreliance on the machine by the DM than in the absence of delegation.

on machine adherence, as this drives the manager's perception of the ex-ante quality of the DM's decision (Prop. EC.1). Part A provides the key prediction: after a bad outcome, despite knowing that the DM likely chose optimally (Prop. 1), the manager pays a lower bonus when the DM overrode the machine signal, and may even reward an ex-ante mistake (such as following a machine signal that contradicts a higher-precision human signal) more generously than an optimal decision. In contrast, after a good outcome, whether the manager pays a lower or a higher bonus when the DM overrides the machine depends on the strength of omission–commission effects, as characterized in Part B. Part C shows that bonus payments are higher for good outcomes.

THEOREM 1 **(Payment Structure)**. *At equilibrium, bonus payments satisfy:*

A. Suppose a bad outcome ($y \neq a$). Then $b(s_M, \underbrace{a = s_M}_{follow}, y \neq a) > b(s_M, \underbrace{a \neq s_M}_{override}, y \neq a)$.

B. Suppose a good outcome ($y = a$). Then $b(s_M, \underbrace{a = s_M}_{follow}, y = a) > b(s_M, \underbrace{a \neq s_M}_{override}, y = a)$ *if* $\omega_g < \omega^*$, *and* $b(s_M, \underbrace{a = s_M}_{follow}, y = a) \leq b(s_M, \underbrace{a \neq s_M}_{override}, y = a)$ *if* $\omega_g \geq \omega^*$.

C. $b(s_M, a, \underbrace{y = a}_{good}) > b(s_M, a, \underbrace{y \neq a}_{bad})$,

We note that Theorem 1 rests on the DM being noisy (Section 3.2). As $\beta \to 0$ the DM never errs, so an override would perfectly reveal a more precise human signal and the manager would have nothing to penalize; it is because $\beta > 0$ that machine adherence becomes an informative, if imperfect, cue to ex-ante decision quality (Lemmas EC.5 and EC.6).

3.6. DM: Machine reliance and decision quality.

We can now turn to our central question: how does the delegation of machine-assisted decisions affect the quality of these decisions? Our primary interest is on informational states with a conflict between the human and the machine signal, which we denote by $\mathbf{s}'_{hi} = (s_M, s_H \neq s_M, \theta = hi)$ and $\mathbf{s}'_{lo} = (s_M, s_H \neq s_M, \theta = lo)$.[9] Here, the DM *under-relies* on the machine by using s_H when they should follow s_M (if $\theta = lo$), which occurs for a bounded rational DM with probability $1 - \mathcal{P}_i(\mathbf{s}'_{lo}) \equiv \Pr(A = \underline{a} | \mathbf{s}'_{lo})$ as per Eq. (1). Similarly, the DM *over-relies* on the machine by following s_M when they should use s_H (if $\theta = hi$), which occurs with probability $1 - \mathcal{P}_i(\mathbf{s}'_{hi})$.

The bonus-payment structure in Theorem 1 has immediate implications for the DM's propensity to over- or under-rely on the machine. The next result summarizes these implications for conflicting signals,[10] contrasting behavior when decisions are delegated ($i = d$) versus not delegated ($i = n$).

[9] Note that $s_M \neq s_H^{hi}$ is the *only* state where the DM can add value for the parameters implemented in our study (specifically, $\mu_0 = 0.5$ and $\underline{r} = \overline{r}$), as s_H^{θ} would not change decisions in the remaining three informational states in $\mathbf{s}$.

[10] The less interesting case of aligning signals requires additional assumptions on ω_g and ω_b; see Appendix A.1.7.

THEOREM 2 **(Machine Reliance.)**. *When signals conflict, i.e., $s_H \neq s_M$:*

A. ***Overreliance**: When $\theta = hi$ (i.e., $\bar{a} = s_H$): $0.5 < \mathcal{P}_d(\mathbf{s}'_{hi}) < \underbrace{\mathcal{P}_n(\mathbf{s}'_{hi}) < 1}_{over-reliance}$.*

B. ***Underreliance**: When $\theta = lo$ (i.e., $\bar{a} = s_M$): $0.5 < \mathcal{P}_n(\mathbf{s}'_{lo}) < \underbrace{\mathcal{P}_d(\mathbf{s}'_{lo}) < 1}_{under-reliance}$.*

The key result of Theorem 2 is that, due to bonus payments in equilibrium, DMs (when facing conflicting evidence) rely more on the machine signal than they would in the absence of a manager. In other words, delegation increases *over-reliance* on algorithmic input when the DM's private information s_H is highly informative (case A), whereas delegation lowers *under-reliance* when s_H is less informative (case B).

4. Hypotheses & Roadmap to Studies

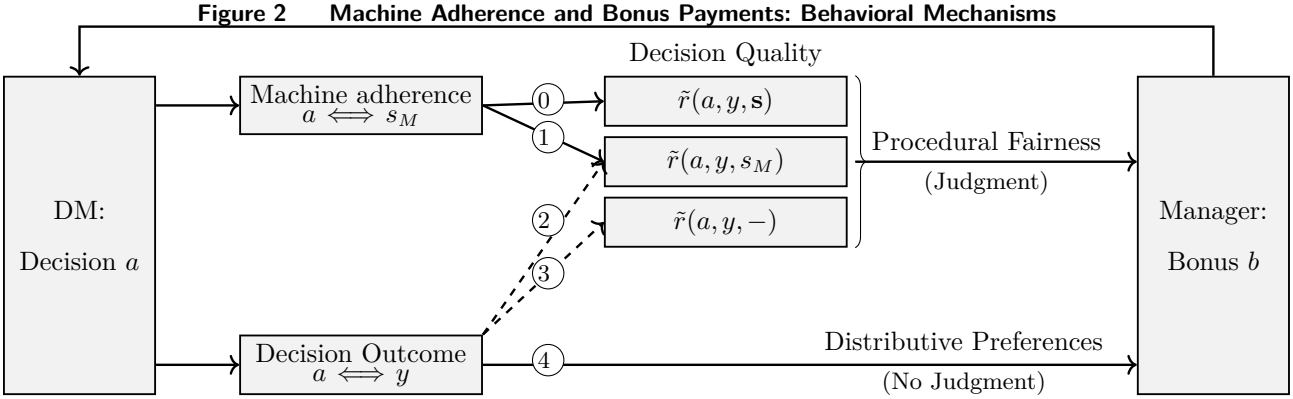
To test the key predictions of our theoretical analyses, as well as the underlying behavioral mechanisms, we design a series of controlled laboratory experiments that implement several variations of the task setting described in Section 3. In each round, the true state $y \in \{R, B\}$ is drawn from the prior $\mu_0 = 0.5$, participants (DMs) receive two conditionally independent informative signals s_M and s_H, choose an action a, the true state y is revealed, the manager's payoff $r(a, y)$ realizes, and the DM receives a bonus b. The basic task environment is identical across our four studies (summarized in Table 1), but we vary the structural parameters $(q_{s_M}, q^{hi}_{s_H}, q^{lo}_{s_H}, \mu_\theta, \overline{r}, \underline{r})$.

Table 1 Overview of Experimental Studies and Treatments

				Structural parameters									
Study	Treatment	N	T	q_{s_M}	$q^{hi}_{s_H}$	$q^{lo}_{s_H}$	μ_θ	$\overline{r}$	$\underline{r}$	Tasks	View	Bonus Source	Public s_H
1	*NoDelegation*	25	40	0.7	0.8	0.6	0.5	\$100	-\$50	1	–	–	–
	Delegation	80	40	0.7	0.8	0.6	0.5	\$100	-\$50	1	Yes	Outside budget	No
2	*NoDelegation*	62	40	0.6	0.7	–	1	\$100	\$30	1	–	–	–
	Delegation	124	40	0.6	0.7	–	1	\$100	\$30	1	No	Manager payoff	No
	DelegationObs	60	40	0.6	0.7	–	1	\$100	\$30	1	No	Manager payoff	Yes
	Delegation_3P	54	40	0.6	0.7	–	1	\$100	\$30	1	No	Outside budget	No
3	*NoDelegation*	28	30	0.7	0.8	0.6	0.5	\$100	-\$50	5	–	–	–
	Delegation	88	30	0.7	0.8	0.6	0.5	\$100	-\$50	5	Yes	Manager payoff	No

In this environment, our theoretical analyses make two main predictions about the interaction between managers and DMs. First, restating Theorem 1, we hypothesize that manager's assessment of the DM's choices (and, as a result, how they punish or reward the DM via their bonus payments) depend systematically on the DM's algorithm adherence (A-B) as well as on the outcome (C).

Figure 2 Machine Adherence and Bonus Payments: Behavioral Mechanisms



HYPOTHESIS 1. Managers (A) strongly punish algorithm overrides after bad outcomes; (B) weakly punish (or even reward) algorithm overrides after good outcomes; (C) pay less for bad outcomes.

Second, focusing on states in which the DM's private signal s_H conflicts with the machine signal s_M (formally, states $\mathbf{s}'_\theta$), we predict that DMs rely on algorithms because algorithm adherence shapes how managers evaluate their decisions (Hypothesis 1). Specifically, Theorem 2 predicts that DMs over-rely on less precise machine signals (i.e., $\theta = hi$) and under-rely on more precise machine signals (i.e., $\theta = lo$). The empirical challenge we face is that we could observe this pattern even without delegation, i.e., in the absence of adherence-dependent bonus payments, e.g., simply due to bounded rationality. To properly test the hypothesis that DM's rely on algorithm specifically because of the incentives that arise endogenously in a delegated setting, in addition to our main treatment *Delegation*, we implement a baseline treatment *NoDelegation*. In this treatment, a single DM receives all information $\mathbf{s}$, chooses a, and earns $r(a, y)$. Restating Theorem 2, we hypothesize that *Delegation* increases algorithm reliance relative to a setting with *NoDelegation*.

HYPOTHESIS 2. Delegation (A) increases over-reliance on a less precise algorithm, and (B) decreases under-reliance on a more precise algorithm.

4.1. Behavioral mechanisms

If we find support for Hypotheses 1 and 2, is it for the reasons that we believe drives algorithm reliance? Our model analysis in Section 3 makes predictions based on specific mechanisms from a behavioral framework illustrated in Figure 2. Across our studies, and with reference to Table 1, we implement several design features that allow us to shed light on the behavioral mechanisms underlying our main predictions. We briefly preview these here, and provide additional details in Appendix B.

4.1.1. View Feature. Although adherence-dependent bonus payments in of themselves are indicative of the judgment process underlying procedural fairness concerns, we attempt to gather more direct evidence in Studies 1 and 3 that implement a "view" feature: Rather than displaying the

public machine signal s_M as a default, the user interface allows managers to access s_M by clicking a "view" button. The feature allows us to cleanly distinguish bonus payments based only on outcomes (paths ③ and ④ in Figure 2) from those based also on machine adherence (paths ① and ②); see the formal analysis in EC.6. Hence, if a manager decides to actively view the machine signal s_M prior to their bonus decision, they reveal their principle motivation to be procedurally fair and let machine adherence guide their assessment of the DM's decision quality. And only if a manager views the machine signal can they label a decision "omission" or "commission". Viewing is thus necessary for adherence-based judgment, but not sufficient: our inference rests on whether bonus payments subsequently vary with machine adherence, not on the click itself.

4.1.2. Outside Budget & 3^{rd} Party Feature. To tease apart the behavioral mechanisms underlying the predicted bonus payments (Figure 2), including mechanisms that our model is silent on, we implement in Studies 1 and 2 two features borrowed from Gurdal et al. (2013). First, managers can make bonus payments between \$0 and \$B from an outside account owned by the experimenter. Because the manager cannot keep any unused bonus budget, they have no direct financial incentive in the bonus decision. Hence, if we observe variations in managers' bonus payments to the DM, they are not for personal financial gain, the manager's (in)ability to pay, or for risk-sharing motives.

Second, managers decide how much money to transfer to the DM as well as to a random third party selected from among the other DMs in the session. The manager thus allocates the bonus budget between two other players that differ only by whether they are associated with the decision and the outcome (i.e., the DM) or not (i.e., the random third party). This feature allows us to assess if payments to the DM reflect a judgment that is directed specifically at the DM (paths ② and ③ in Figure 2). If the manager only made bonus payments to the DM, then observing the manager pay less to the DM after a bad outcome would be consistent with other mechanisms, such as distributive preferences (path ④) or the desire to share disappointment with the outcome with all others. With the third-party feature, however, if we observe that payments to the third party are substantially less sensitive (or even directionally opposed) to decision outcomes and algorithmic adherence than payments to the DM, this would provide evidence that managers specifically blame their DMs.

4.1.3. *DelegationObs* Treatment. Our theory predicts that machine-adherence emerges endogenously from adherence-dependent (mis)judgments that are specifically driven by managers' uncertainty about the DM's information at the time of choice (i.e., $\mathbf{s}$). To test this idea, we implement in Study 2 a *DelegationObs* treatment that provides the manager full information about $\mathbf{s}$, effectively switching off (or weakening) any behavioral mechanism that rely on the manager's ambiguity about $\mathbf{s}$. This treatment allows for sharp adherence-dependent inferences about ex-ante decision quality (path ⓪ in Figure 2) - if the manager observes that the DM overrode the machine signal, they

do not have to guess whether the DM had conflicting evidence from a more precise human signal (which would justify the override) or not; see the formal analysis in EC.5. This further leaves less moral wiggle room that otherwise (via paths ① and ②) might allow managers to pay lower bonus payments without guilt.

4.1.4. Belief Elicitation. At the end of the sessions for Studies 1 and 3, we administered a short incentivized survey that elicited from participants their beliefs about key objects central to our model. First, from the DM's perspective, we elicited beliefs about the posterior probability of the true state after observing signals $\mathbf{s}$, namely $\mu_{\mathbf{s}}$. Second, from the manager's perspective, we elicited beliefs about the DM's private information at the time of the decision, namely $\Pr(s_H, \theta \mid s_M, a, y)$ (paths ① and ② in Figure 2). Taken together, these surveys provide supporting evidence for the mechanism driving our results. We report the details in Appendix C.5.

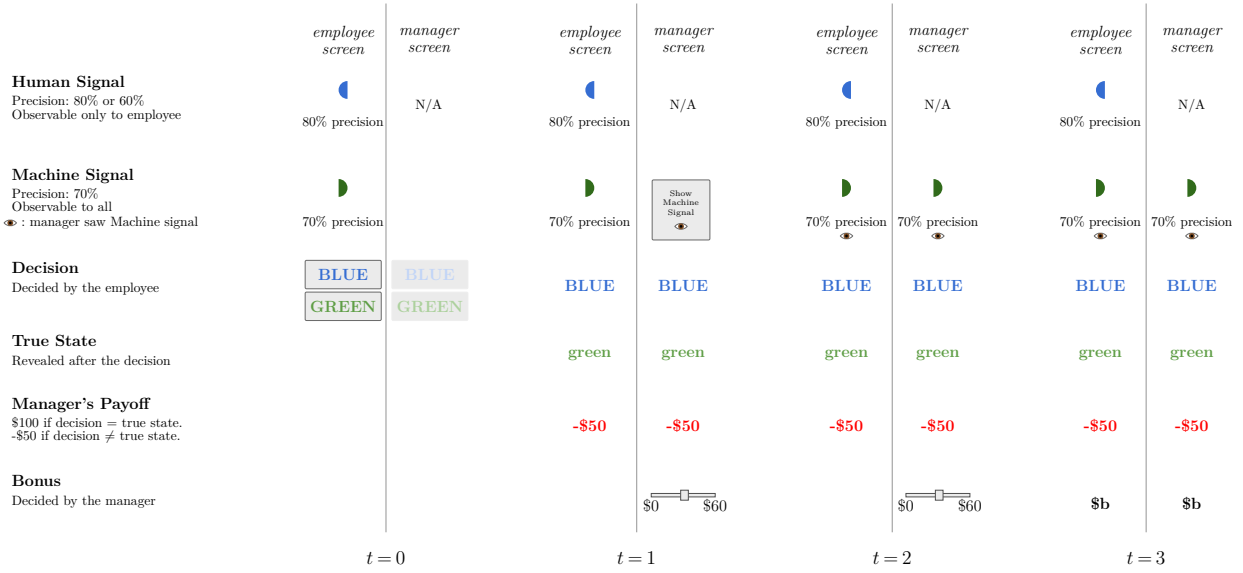
4.1.5. Algorithm Design & Presentation. Our setting captures the value of a machine in the decision making process via q_{s_M} (vs. $q^{\theta}_{s_H}$). Because it is theoretically irrelevant how a signal s_M is generated as long as it has the properties of q_{s_M}, the presentation of the algorithm to participants becomes a matter of framing only. Our design aims to provide meaningful context to discriminate between the two signals beyond the mere semantics of the terms "human" and "machine", while retaining our control over q_{s_M} and $q^{\theta}_{s_H}$. We do *not* frame s_M explicitly as a decision *recommendation*, to avoid demand effects that could possibly arise simply from the term "recommendation". Participants are told that machine signals come from a statistical *algorithm* that analyzes high-dimensional *data*, with a new case (new variable values) in each round. We provide more detail in Appendix B.2.

4.2. Recruitment, Software, Payments.

In total, 521 participants were included in our studies, and each human subject participated in one treatment only. We recruited from a participant pool associated with the experimental laboratory at a large public university. The experiment was implemented in oTree (Chen et al. 2016). Figure 3 provides a stylized illustration of the decision screens faced by the DM and the manager in Study 1; the actual screens for all studies and treatments are reproduced in Appendix B.3.

In each session, after arriving at the laboratory, participants were randomly assigned to isolated cubicles. Participants then read instructions which were presented on-screen in an easy to navigate format. After reading the instructions, participants were randomly assigned to the manager or the DM role, which they then kept for the duration of the experiment. At the beginning of each round, each manager (responsible for bonus decisions) was randomly matched with one DM (responsible for the diagnostic decision). The one-shot and anonymous nature of relationships in our experiments effectively eliminates reputational concerns. Information on all experimental sessions is provided in Appendix Table EC.1.

Figure 3 Stylized decision interface across a round (Study 1)



Note. The stages shown in this illustration correspond directly to the timeline of events in Figure 1.

We computed cash earnings for each participant by multiplying the total earnings from four randomly selected rounds by a predetermined exchange rate. Participants were paid their earnings in private and in cash at the end of the session. Earnings, including the participation fee, amounted to €18.71 on average, with a maximum of €53. For all studies, prior to collecting data, we pre-registered the main research hypotheses, key dependent and independent variables, analysis plans, target sample sizes, and exclusion criteria.[11]

4.3. Data and Analysis.

At the most granular level, in each period $t \in \{1, ..., T\}$ and for each pair $i \in \{1, ..., I\}$, we observe the DM's decision a_{it} as well as the manager's bonus decision b_{it}. In Studies 1 and 2, we also observe the bonus b_{it}^{3p} paid to a random third party, and thus the sum $\Sigma_{it} = b_{it} + b_{it}^{3p}$. Our data also comprises signals $\mathbf{s}_{it} \doteq \{\theta_{it}, s_{H,it}, s_{M,it}\}$ and true states y_{it}, which are pre-generated outcomes of random variables (see Appendix B.2 for details).

To formally test our main hypotheses, we rely on a set of regression models tailored to the different outcome variables of interest. To study DMs' machine-adherence choices—central to Hypothesis 2—we estimate linear probability models, with $OptDecision_{it}$ (1 if $a_{it} = \bar{a}|\mathbf{s}_{it}$, and 0 otherwise) as DV, and several IVs for what the DM can observe, $PreciseH_{it}$ (1 if $\theta_{it} = hi$, and 0 otherwise) and $Conflict_{it}$ (1 if $s_{H,it} \neq s_{M,it}$, and 0 otherwise). To study managers' bonus decisions—central to Hypothesis 1—we estimate OLS models with $Bonus_{it}^{x}$ ($x \in \{dm, 3\text{p}, \Sigma\}$) as DV, and several IVs

[11] The full pre-registration documents are available at `https://aspredicted.org/pp26rx.pdf` (Study 1); `https://aspredicted.org/fytj-63h2.pdf`, `https://aspredicted.org/3zsx-gz4m.pdf`, `https://aspredicted.org/rm7k5g.pdf` and `https://aspredicted.org/su82hz.pdf` (Study 2); `https://aspredicted.org/qp7k65.pdf` (Study 3).

for what the manager can observe, $BadOutcome_{it}$ (1 if $a_{it} = y_{it}$, 0 otherwise) and $FollowM_{it}$ (1 if $a_{it} = s_{M,it}$, 0 otherwise). To control for possible learning pattern, we include in our estimations $Round_{it}$, mean-centered so that intercepts (when applicable) can be interpreted as outcomes at the average round. For all model specifications, standard errors are computed using two-way cluster-robust methods, clustered by participant i and by matching-group/session g.

5. Study 1.

5.1. Design and Implementation

We implement two treatments that correspond to the settings studied in Section 3.6. In *Delegation*, DM can observe the human signal, its precision level, *and* the machine signal. Whereas, the manager can only observe s_M, by clicking a button after the true state is realized (see Section 4.1.1). From an outside budget B, the manager pays a bonus to the DM and another bonus to a random third party who, by design, is unrelated to the decision and its outcome (see Section 4.1.2). In *NoDelegation*, a single DM observes both signals and their respective precision levels, chooses a, and earns $r(a, y)$.

Table 1 summarizes Study 1's parameters. We set $\overline{r} = \$100$ and $\underline{r} = -\$50$ to create high stakes for incorrect decisions. In particular, $\underline{r} < 0$ makes bad outcomes salient and ensures that ex-post mistakes by the DM reduce the manager's earnings significantly. The manager pays bonuses from an outside budget of $B = \$60$ provided by the experimenter. We set the machine signal precision at $q_{s_M} = 0.7$, while the human signal is high-precision ($q^{hi}_{s_H} = 0.8$) or low-precision ($q^{lo}_{s_H} = 0.6$) with equal probability ($\mu_\theta = 0.5$). The manager cannot tell whether the DM followed or overrode a machine signal relative to the DM's private human signal of precision ($\theta = lo, hi$), and has no ex-ante reason to view either as superior: our parameters imply equal ex-ante expected precision, $q^0_{s_H} \equiv \mathbb{E}_\theta \left[q^\theta_{s_H} \right] = 0.7 = q_{s_M}$.

5.2. Results

5.2.1. DM: Algorithm Adherence and Delegation. We first study the DM's tendency to follow the machine signal, correctly or not, and how this algorithm adherence depends on delegation. Not surprisingly, the fraction of correct choices is strictly lower than 1 in both treatments, and for all combinations of signal precision and signal alignment. Table 2 presents the results from a regression that, with *OptDecision* as DV, includes a treatment dummy *Delegation*, the signal precision dummy variable *PreciseH* (1 if $\theta_{it} = hi$, and 0 otherwise), and their interaction. Because our main focus is on decision making when the DM has conflicting signals, and to avoid presenting three-way interaction effects, we fit the regression separate for the instances in our data when signals are in conflict, $Conflict_{it} = 1$, and aligned, $Conflict_{it} = 0$.

When the DM faces a conflict between the machine signal and a high-precision human signal (i.e., the DM should override s_M), the results show that delegation increases algorithmic over-reliance ($\beta_1 + \beta_3 = -0.180, p = 0.03$) providing support for Hypothesis 2A. In contrast, when the machine

Table 2 Drivers of Algorithm Reliance.

DV: *OptDecision*		$Conflict = 1$	$Conflict = 0$
β_0	Constant	0.843***	0.941***
		(0.043)	(0.020)
β_1	*Delegation*	0.067	0.015
		(0.059)	(0.035)
β_2	*PreciseH*	0.071	0.052**
		(0.046)	(0.023)
β_3	*Delegation* × *PreciseH*	-0.248**	-0.060**
		(0.089)	(0.024)
β_4	Round	-0.001	-0.001
		(0.001)	(0.001)
Observations		1,114	1,486
Participants		65	65
Sessions		14	14

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

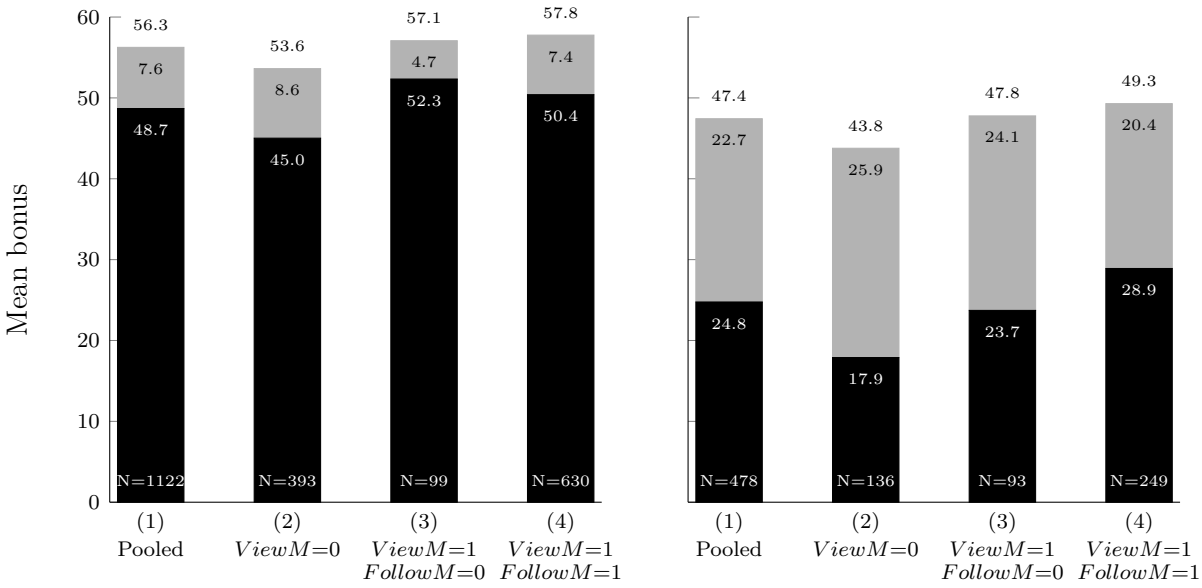
signal conflicts with a low-precision human signal (i.e., the DM should follow s_M), the results show that delegation decreases algorithmic under-reliance, but the effect is weaker and not statistically significant ($\beta_1 = 0.067, p = 0.27$), providing no support for Hypothesis 2B. The implication of these results is that delegation increases algorithm reliance when signals conflict.[12] For the less interesting cases where the two signals align, delegation weakly increases the number of blatantly incorrect choices (i.e., overriding *both* signals) when the human signal is highly precise ($\beta_1 + \beta_3 = -0.045, p = 0.20$), with no treatment effect when it is of low precision ($\beta_1 = 0.015, p = 0.67$).

Table 2 yields another important observation: in *NoDelegation*, the DM is more likely to correctly follow the more precise human signal ($\theta = hi$) than the more precise machine signal ($\theta = lo$) ($\beta_2 = 0.071, p = 0.15$). In *Delegation*, the reverse holds: the DM is less likely to correctly follow the more precise human signal than the more precise machine signal ($\beta_2 + \beta_3 = -0.177, p = 0.04$). While this effect of delegation is not directly predicted by our theory (absent further assumptions on the strength of behavioral effects), it illustrates how organizational context can push DM to ignore superior private information and thus make worse decisions precisely when they should make better ones.

5.2.2. Managers: Bonus Payments. Our data is consistent with Hypothesis 2: when facing conflicting evidence, DMs follow the machine more in delegation than in no delegation. Our theory posits that delegation affects algorithm adherence endogenously, through how managers assess decisions and set bonus payments. Figures 4a (for good outcomes) and 4b (for bad outcomes) display bonus payments, for (1) the pooled data, as well as broken down by manager decisions made (2) without viewing the machine signal s_M, and after viewing s_M and observing either (3) machine

[12] For a direct test of this statement, we fit a simple model, $FollowM = \beta_0 + \beta_1 Delegation + \beta_2 Round$. For $Conflict = 1$, the treatment effect of *Delegation* is $\beta_1 = 0.123$ ($p = 0.01$). See Table EC.2 for details.

Figure 4 Bonus payments to DM (■) and to Third Party (■)



(a) $BadOutcome = 0$ (b) $BadOutcome = 1$

adherence or (4) a machine override. To test our main observations formally, we estimate several regression models (Table 3), separate for the two levels of $ViewM$ to simplify the exposition, and we present several robustness checks in Appendix C.1.2.[13]

To study if bonus payments are **outcome-dependent**, we turn to the data from $ViewM = 0$, i.e., where machine adherence cannot affect the bonus as the manager has chosen *not* to view the machine signal.[14] Consistent with Hypothesis 1C, we observe that managers pay their DMs higher bonuses after a good outcome than after a bad outcome ($\beta_1 = -27.120, p < 0.01$).

To study if bonus payments are **adherence-dependent**, we turn to the data from $ViewM = 1$, i.e., where the manager's choice to view s_M indicates they want to "use" the DM's machine adherence to assess the DM's decision quality. After observing a bad outcome, the manager pays a higher bonus if the DM followed the machine ($\beta_2 + \beta_3 = 5.365, p = 0.05$), providing support for Hypothesis 1A which predicts that managers punish machine overrides. For good outcomes, we find no evidence that overrides are punished ($\beta_2 = -1.918, p = 0.47$), which is in line with Hypothesis 1B.[15]

Figure 4 reveals another noteworthy observation that leverages the **"view-dependent"** feature of our study: bonus payments appear to be lower when the manager choose to *not* view the machine

[13] Specifically, we address possible endogeneity issues that may in principle arise from the fact that both *Bonus* and $ViewM$ are manager decision variables (Table EC.3). Our main results are robust.

[14] For $ViewM = 0$, outcomes can affect bonus payments only via channels ① and ② in Figure 2. In contrast, for $ViewM = 1$, outcomes can also affect payments via ③.

[15] Recall that our theory is silent on the directional effect of machine reliance after good outcomes, and predicts only that its effect on bonus payments is weaker after good than after bad outcomes.

Table 3 Bonus Payments

DV:		$Bonus^{DM}$		$Bonus^{3p}$		Total Bonus	
		$ViewM=0$	$ViewM=1$	$ViewM=0$	$ViewM=1$	$ViewM=0$	$ViewM=1$
β_0	Constant	45.007***	52.429***	8.514***	4.704**	53.521***	57.133***
		(4.220)	(3.097)	(2.263)	(1.626)	(3.634)	(2.760)
β_1	*BadOutcome*	-27.120***	-28.933***	17.351**	19.427***	-9.769**	-9.506*
		(6.139)	(4.222)	(7.524)	(4.607)	(3.792)	(4.236)
β_2	*FollowM*		-1.918		2.621		0.703
			(2.544)		(1.585)		(1.658)
β_3	*BadOutcome* × *FollowM*		7.283		-6.352		0.931
			(4.281)		(3.560)		(2.804)
β_4	*Round*	0.009	0.071	0.053	-0.025	0.061	0.047
		(0.116)	(0.050)	(0.053)	(0.033)	(0.085)	(0.057)
Observations		529	1,071	529	1,071	529	1,071
Participants		30	39	30	39	30	39
Sessions		10	10	10	10	10	10

Notes: Standard errors (in parentheses) are two-way cluster-robust, clustered by participant and session. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

signal. For example, for bad outcomes (Figure 4b), managers pay less when they observe the DM overrode the machine (Hypothesis 1A), but they pay even less when they chose to not observe the DM's machine adherence at all![16] This pattern is not immediately explicable from our model which predicts that, for a given outcome, adherence-blind bonuses should lie strictly between those for observable machine overrides and machine reliance (Proposition EC.6). Instead, because such blindness is caused by the manager's own choice, our data is at least suggestive of managers' strategic ignorance about the DM's state of information at the time of the decision (Dana et al. 2007, Grossman 2014). Although creating such "moral wiggle room" for lower bonus payments does not directly benefit managers financially due to the outside budget feature, it does allow for a relatively guilt-free way to simply (and possibly unfairly) blame the DM for bad outcomes. Appendix C.1.2 provides further evidence consistent with this interpretation: managers who view less frequently pay lower bonuses to their own DMs, but not to third parties.

5.2.3. Managers: Mechanism Evidence (from $Bonus^{3p}$ data). To assess if variations in bonus payments reflect manager judgment directed specifically at the DM, we study bonus payments to the third party (in Figure 4, and $Bonus^{3p}$ in Table 3). Using the data from $ViewM=0$, and thus controlling for adherence-dependence, we make a key observation: while paying a lower bonus to the DM after a bad outcome, managers pay *more* to the random third party ($\beta_1=17.351, p=0.05$). This

[16] We provide the statistical evidence for this claim in Appendix C.1.2, from an integrated framework on the pooled data that allows us (unlike the separate regressions in Table 3) to compare bonus decisions between $View=0$ and $View=1$, while properly addressing endogeneity and manager heterogeneity.

pattern is not consistent with distributive preferences, but an indication that the manager holds the DM individually responsible for the outcome.[17] We observe a similar pattern in the $ViewM = 1$ data: while paying the DM more for following the machine after a bad outcome, managers pay the third party less ($\beta_2 + \beta_3 = -3.731, p = 0.25$), providing corroborating evidence that adherence-dependent bonus payments to the DM reflect judgments that are directed specifically at the DM.

5.2.4. Managers: Mechanism Evidence (from $ViewM$ data). To validate some of our key modeling assumptions in Section 3.4, we next study in more detail the managers' decisions to view the machine signals s_M. First, managers view s_M prior to 67% of all bonus decisions in our data, supporting the key modeling assumption that procedurally fairness-minded managers want to take the machine signal into account when evaluating the DM's decision. Second, our theoretical analyses predict (and our data shows) an asymmetric effect of outcome on bonus payments *because* outcome moderates the thought process underlying the manager's bonus decision, captured in the outcome-dependent factor $\omega(s_M, a, y)$ in our model analysis. We fit a linear probability model, $ViewM = \beta_1 BadOutcome + \beta_2 Round + \beta_3(BadOutcome \times Round)$, and find that managers are on average more likely to view the machine signal after a bad outcome ($\beta_1 = 0.067, p = 0.12$). Further, while the tendency to view decreases after good outcomes ($\beta_2 = -0.005, p = 0.04$), it remains at a high level after bad outcomes ($\beta_2 + \beta_3 = 0.002, p = 0.62$); see Table EC.4 for details. Managers thus are more likely to view the machine signal after a *b*ad outcome than after a *g*ood outcome, which is consistent with our assumptions about the outcome-dependent weights ω_b and ω_g in Section 3.4.

5.2.5. Does Algorithm Adherence "Pay Off"? To properly link bonus payments with decisions, we return to the perspective of the DM who needs to consider the impact of decisions on expected bonus payments *before* the outcome materializes. Given that managers' bonus payments appear to be tilted in favor of machine adherence, it is tempting to ask if DMs are better off (incorrectly) adhering to s_M even when they have more precise, and conflicting, private information? This is not the case. Although machine adherence shelters the DM from blame after a bad outcome, it simultaneously increases the probability of a bad outcome in the first place, which in turn reduces the expected bonus because managers reward good outcomes more than bad outcomes. Thus, the DM trades off the higher probability of a good outcome from overriding against the manager's expected penalty for overriding after a bad outcome. This, in fact, is the prediction implied by Proposition 1 - despite anticipating blame when chance yields a bad outcome, in equilibrium, the DM still is more likely to (correctly) override the machine signal. Our data agrees with this prediction: when the DM

[17] Note that this does not rule out concerns for distributive fairness in our data - on the contrary, total bonus is lower after a bad outcome (this is true for both models we present in Table 3: $ViewM = 0$ and $ViewM = 1$).

has conflicting but superior private information, machine adherence reduces the DM's average bonus by 9.24 points.[18]

6. Study 2: Reducing Ambiguity.

Study 1 provides evidence of outcome- and adherence-dependent bonus payments that reflect managers' judgment of DMs' choices. In principle, these judgments can be rationalized if the manager believes that the DM makes mistakes (an assumption justified by our data) and uses available information to form a belief about the probability that the DM indeed made a mistake. By inducing significant uncertainty about the DM's information at the time of the DM's decision, our decision environment makes such belief updating cognitively challenging, creating the moral wiggle room for asymmetric, and possibly unfair (to the DM) judgments. When the manager observes a costly machine override, it may be psychologically easier to believe that the DM made an *ex-ante* poor decision when the machine signal could have been more precise than the DM's private signal (with $1-\mu_\theta = 0.5$ in Study 1). It thus stands to reason that an environment that reduces ambiguity (about the DM's information) surrounding the manager's judgments would likely mitigate the key driver of algorithm adherence: managers blaming DMs for (correctly!) overriding the machine signal.

6.1. Design and Implementation

We design Study 2 to test this idea. First, we set parameters such that $q_{s_H} > q_{s_M}$ (with, $\mu_\theta = 1$, $q_{s_H} = 0.7$, and $q_{s_M} = 0.6$), leaving no room for the manager to believe that the DM has overridden a more precise machine: a machine override strongly indicates that the DM had conflicting evidence and chose to follow the human signal that the manager knows (but cannot see) is more precise.

Second, besides *NoDelegation* and *Delegation*, we implement a treatment *DelegationObs* that allows the manager to observe the machine signal s_M *and* the DM's signal s_H. The treatment provides a direct test of our theoretical prediction that algorithm reliance arises specifically from the key characteristic of our decision environment, namely, that the DM might hold valuable information that neither the algorithm nor the manager can access. Effectively eliminating the belief updating process that tilts the manager's assessment towards machine-adherence, *DelegationObs* mutes adherence-dependent bonus payments that provide the DM the incentives to override their own (more precise!) signal. As a result, we predict that *DelegationObs* has a weaker effect on machine adherence than *Delegation*, and possibly none at all. We provide the formal analysis behind this claim in Proposition EC.4 (Appendix A.1.8).

[18] This is based on a simple regression $BonusReceive_{it} = \beta_0 + \beta_1 FollowM_{it} + \varepsilon_{it}$ with standard errors clustered at the DM and session levels. We find that $\beta_1 = -9.240$ ($p = 0.09$). See Table EC.5 for details.

Table 4 Drivers of Algorithm Reliance

DV: *OptDecision*		(1) $Conflict = 1$	(2) $Conflict = 0$
β_0	Constant	0.844***	0.956***
		(0.013)	(0.015)
β_1	*DelegationObs*	0.025	0.010
		(0.050)	(0.031)
β_2	*Delegation*	-0.151***	-0.066
		(0.036)	(0.045)
β_3	Round	-0.002**	-0.000
		(0.001)	(0.000)
Observations		2,862	3,298
Participants		154	154
Sessions		10	10

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

6.2. Results

6.2.1. DM: Algorithm Adherence and Delegation. Table 4 presents the results from a regression that, with *OptDecision* as DV, includes dummies for the delegation treatments. Contrasting decisions in *Delegation* and *NoDelegation*, we again find support for Hypothesis 2A: when the DM faces a conflict between the machine signal and the more precise human signal (i.e., the DM should override s_M), *Delegation* increases algorithmic over-reliance ($\beta_2 = -0.151, p < .01$). For the less interesting case of aligning signals, *Delegation* weakly increases the number of blatantly incorrect choices (i.e., overriding *both* signals; $\beta_2 = -0.066, p = .18$).

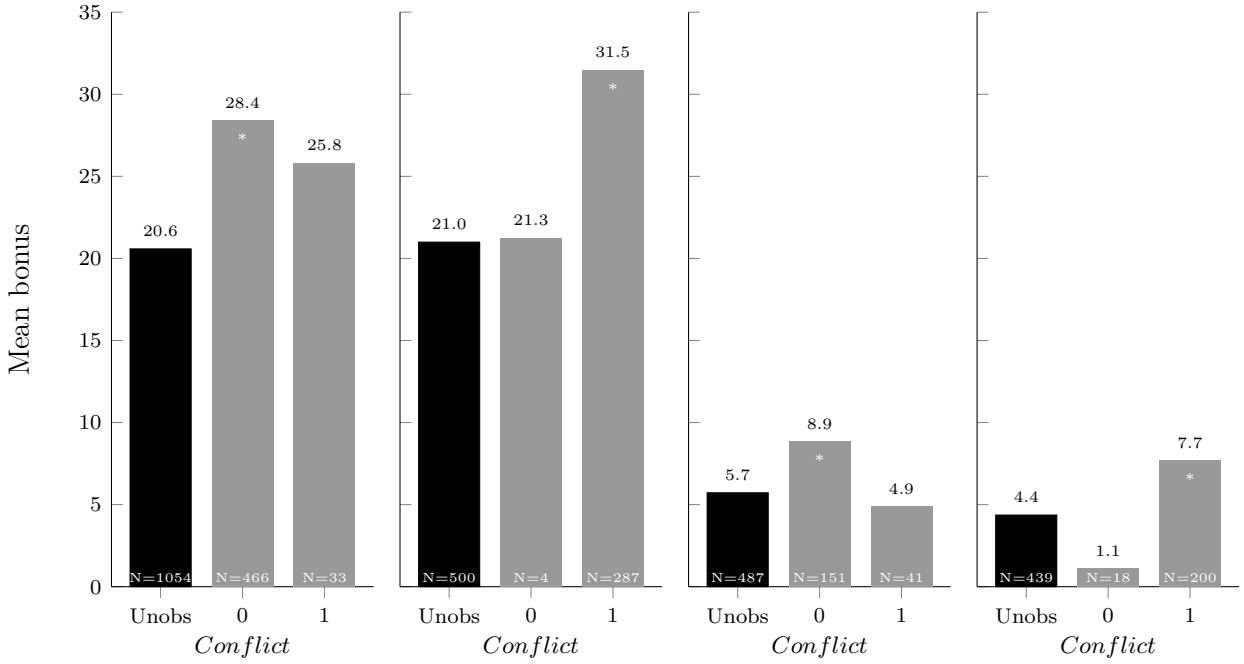
Our theoretical analyses predict delegation-induced machine adherence in large parts because the information asymmetry in *Delegation* leaves the manager with a challenging judgment problem as well as moral wiggle room. To directly test this idea, we designed treatment *DelegationObs* where the DM faces a manager who can observe the DM's information and thus can perfectly assess the ex-ante quality of the DM's choices. As predicted, the data shows that adherence is lower in *DelegationObs* than in *Delegation*, both when signals conflict ($\beta_1 - \beta_2 = 0.175, p = .02$) and when signals align ($\beta_1 - \beta_2 = 0.075, p = 0.17$). In fact, we observe no difference in machine adherence in *DelegationObs* and the baseline *NoDelegation* treatment, both when the DM faces a signal conflict ($\beta_1 = 0.025, p = 0.63$) and when signals align ($\beta_1 = 0.010, p = 0.77$).

To examine whether learning mitigates these results, we estimate models that accommodate treatment-specific effects of round.[19] The data exhibits no apparent trend in *DelegationObs* and *NoDelegation*. But algorithmic adherence (over-reliance for the parameters of this study) in fact grows with experience in *Delegation*, suggesting that the pattern is unlikely to represent a bias that DMs learn to avoid over time, but DMs' response to bonus payments that incentivize algorithm adherence.

[19] We present detailed results in Table EC.6. The specification in Table 4 captures *average* learning across treatments.

6.2.2. Managers: Bonus Payments. Figure 5 displays bonus payments for the two treatments and Table 5 shows the regression results to formally test the key patterns. In terms of **outcome-dependence**, managers in *Delegation* pay the DMs less after a bad outcome, whether the DM followed the machine ($\beta_1 + \beta_3 = -14.799$, $p < 0.01$) or not ($\beta_1 = -16.510$, $p < 0.01$), consistent with Hypothesis 1C. Similarly, managers in *DelegationObs* pay less after a bad outcome, whether the DM follows the machine ($\beta_1 + \beta_3 = -19.475$, $p < 0.01$) or not ($\beta_1 = -19.679, p < 0.01$).

Figure 5 Bonus Payments to DM: *Delegation* (■) and *DelegationObs* (■). *(Conditionally) Optimal Choice.



(a) $BadOutcome = 0$, $FollowM = 1$ (b) $BadOutcome = 0$, $FollowM = 0$ (c) $BadOutcome = 1$, $FollowM = 1$ (d) $BadOutcome = 1$, $FollowM = 0$

In terms of **adherence-dependence**, we find no effect of machine adherence on bonus payments after good outcomes ($\beta_2 = -0.377, p = 0.73$), but again find support for Hypothesis 1A: the manager in *Delegation* punishes machine overrides after bad outcomes ($\beta_2 + \beta_3 = 1.334, p < 0.01$), despite the fact that the parameters of Study 3 should make it abundantly clear that the DM most likely just followed the human signal that the manager *knows* is more precise. We designed *DelegationObs* to test whether incentives to adhere to the machine are mitigated, and thus better aligned with payoff maximization, when the manager can perfectly discriminate between ex-ante good and bad decisions. Our data shows that this clearly is the case. After bad outcomes, managers reward overrides of machine signals that are in conflict with more precise human signals ($\beta_2 + \beta_3 + \beta_6 + \beta_7 = -2.844$, $p = 0.02$), and they strongly punish obviously costly overrides of machine signals that align with the human signal ($\beta_2 + \beta_3 = 7.625$, $p < 0.01$). We observe a similar pattern after good outcomes. The

Table 5 Bonus Payments to DM

DV: $Bonus^{DM}$		*Delegation*	*DelegationObs*
β_0	Constant	20.937***	20.952***
		(2.361)	(4.135)
β_1	*BadOutcome*	-16.510***	-19.679***
		(1.890)	(4.206)
β_2	*FollowM*	-0.377	7.421
		(1.086)	(5.517)
β_3	*BadOutcome* × *FollowM*	1.711	0.205
		(1.127)	(4.911)
β_4	*Conflict*		10.510*
			(5.513)
β_5	*BadOutcome* × *Conflict*		-4.076
			(5.047)
β_6	*Conflict* × *FollowM*		-12.973*
			(7.592)
β_7	*BadOutcome* × *Conflict* × *FollowM*		2.505
			(6.487)
β_8	*Round*	-0.054**	-0.040
		(0.026)	(0.028)
Observations		2,480	1,200
Participants		62	30
Sessions		4	2

Notes: Standard errors clustered by participant. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

bonus payments in *DelegationObs* indicate that managers are in general motivated to adjust bonus payments based on the ex-ante expected payoff of the DM's choice, which is consistent with the procedural fairness mechanism of our theoretical model, and they provide clear incentives for the DM to choose optimally. In contrast, bonus payments in *Delegation* often reward the DM for suboptimal choices, such as following a machine signal that conflicts with a more precise human signal.

6.2.3. Does Algorithm Adherence "Pay Off"? To properly link bonus payments with decisions, we return to the perspective of the DM. For *Delegation*, our data again confirms the prediction implied by Proposition 1: despite anticipating possible blame when chance yields a bad outcome, in equilibrium, the DM is more likely to (correctly) override the machine. Specifically, when signals conflict, incorrectly following the less precise machine signals reduces the DM's average bonus by 3.905 points.[20] Importantly, *DelegationObs* reduces the DM's average bonus for incorrect machine-adherence more strongly (by 7.444 points) because it eliminates the driver for algorithm adherence in *Delegation*: after a bad outcome, managers blame the DM for overriding the machine signal.

6.2.4. Managers: Mechanism Evidence (from $Bonus^{3p}$ data). We implemented an additional treatment, *Delegation_3P*, with the outside budget and 3rd party feature from Study 1. We observe that DMs incorrectly follow the machine signal in 23% of cases when they have a more precise human signal pointing in the opposite direction. And, similar to Study 1, this over-reliance increases

[20] This is based on a simple regression $BonusReceive_{it} = \beta_0 + \beta_1 \cdot DelegationObs_{it} + \beta_2 \cdot FollowM_{it} + \beta_3 \cdot DelegationObs_{it} \cdot FollowM_{it} + \beta_4 \cdot Round + \epsilon_{it}$. See Table EC.7 in Appendix C.2.2 for details.

Table 6 Bonus Payments in *Delegation_3P*

DV:		$Bonus^{DM}$	$Bonus^{3p}$	Total Bonus
β_0	Constant	46.458***	3.041***	49.499***
		(0.975)	(0.807)	(0.287)
β_1	*BadOutcome*	-19.476***	14.441***	-5.036**
		(2.532)	(2.589)	(2.174)
β_2	*FollowM*	-0.290	0.001	-0.290*
		(0.397)	(0.321)	(0.167)
β_3	*BadOutcome* × *FollowM*	0.868	-1.227	-0.359
		(1.576)	(1.491)	(0.766)
β_4	*Round* ($round_c$)	-0.004	0.010	0.006
		(0.034)	(0.032)	(0.028)
Observations		1,080	1,080	1,080
Participants		27	27	27
Sessions		3	3	3

Notes: Standard errors clustered by participant.*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

($t = 1 - 20 : 18.4\%, t = 21 - 40 : 27.6\%, p = 0.09$) with time, which suggests that delegation-driven over-reliance is not a behavior that can be debiased with increased task experience.

To assess whether outcome- and adherence-dependent bonus payments reflect judgments that are directed specifically at the DM, Table 6 presents regression results for bonus payments to the DM and the third party. While paying a lower bonus to the DM after a bad outcome, whether the DM followed the machine signal ($\beta_1 + \beta_3 = -18.608$, $p < 0.01$) or not ($\beta_1 = -19.476$, $p < 0.01$), managers pay more to the random third party ($\beta_1 + \beta_3 = 13.214$, $p < .01$; $\beta_1 = 14.441$, $p < .01$). This pattern is not consistent with distributive preferences alone, but an indication that the manager holds the DM individually responsible for the outcome (via paths ② and ③ in Figure 2).

7. Study 3: Task Aggregation and Feedback Frequency.

Our main theoretical predictions, and most noteworthy results, relate to manager judgments after a DM's choice (and luck) led to a bad outcome. In many natural environments, such bad decision outcomes are embedded in possibly long sequences of decisions that lead to good outcomes, which raises a natural question: Would delegation-driven machine-reliance, and adherence-dependent managerial judgments, survive in such an environment where managers would not necessarily judge their DMs based on every single decision outcome? To address these questions, and assess boundary conditions to our previous results, Study 3 implements a setting in which DMs make several decisions before the manager evaluates their performance and assigns a bonus.

7.1. Design and Implementation

Specifically, in each round, the DM makes five decisions. For each task, one at a time, the DM first observes both signals, then makes a decision, after which the outcome materializes. The manager observes DM's decisions and realized outcomes as they occur. Only after all five tasks are completed,

Table 7 Drivers of Algorithm Reliance.

DV: *OptDecision*		$Conflict = 1$	$Conflict = 0$
β_0	Constant	0.903***	0.995***
		(0.018)	(0.004)
β_1	*Delegation*	-0.0733**	-0.093***
		(0.034)	(0.026)
β_2	*PreciseH*	-0.025	0.002
		(0.038)	(0.004)
β_3	*Delegation* × *PreciseH*	0.006	0.026*
		(0.051)	(0.014)
β_4	Round	0.000	-0.002*
		(0.001)	(0.001)
Observations		4,493	6,307
Participants		72	72
Sessions		16	16

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

the manager can choose to view the machine signal for each task (Section 4.1.1). Finally, the manager pays a bonus to the DM. Managers make bonus payments out of their own payoff, which is the average of the payoffs of the five tasks. We again implement two treatments, *Delegation* and *NoDelegation*, and we return to the parameters from Study 1, i.e., where the manager faces uncertainty over the relative precision of the DM's signals.

The implementation of a setting with less frequent managerial judgment of decisions and outcomes creates some experimental trade-offs. By definition, this setup attenuates the link between a bonus payment and the outcome and adherence of each individual task, complicating our inferences about what drives bonus payments. Further, our interest in the effect of "salient" bad decision outcomes on manager's judgments, requires that bad outcomes should have a strong effect on the manager's payoff, but also that they do not happen so frequently to become the default. We chose the same parameters of Study 1 which, under optimal decision making (which we do not expect), provide a reasonable balance between the probability (%) of bad outcomes and the corresponding effect on the manager's round payoff $\$\bar{r} = r(a, y)/5$: one bad outcome (40%, \$70), two bad outcomes (26%, \$40), three bad outcomes (9%, \$10), four bad outcomes (1%, -\$20), and five bad outcomes (0.1%, -\$50).[21]

7.2. Results

7.2.1. DM: Algorithm Adherence and Delegation. We first study the DM's tendency to follow the machine signal, correctly or not, and how this algorithm adherence depends on delegation, using the same regression framework as in Study 1 (Table 7).

[21] We could instead have chosen parameters that result in predominantly good outcomes, and very rare bad outcomes of significant consequence (such that a bad outcome that "wipes out" several good outcomes). However, this choice would quickly leave us with too small a sample size of (managers' judgments of) bad outcomes.

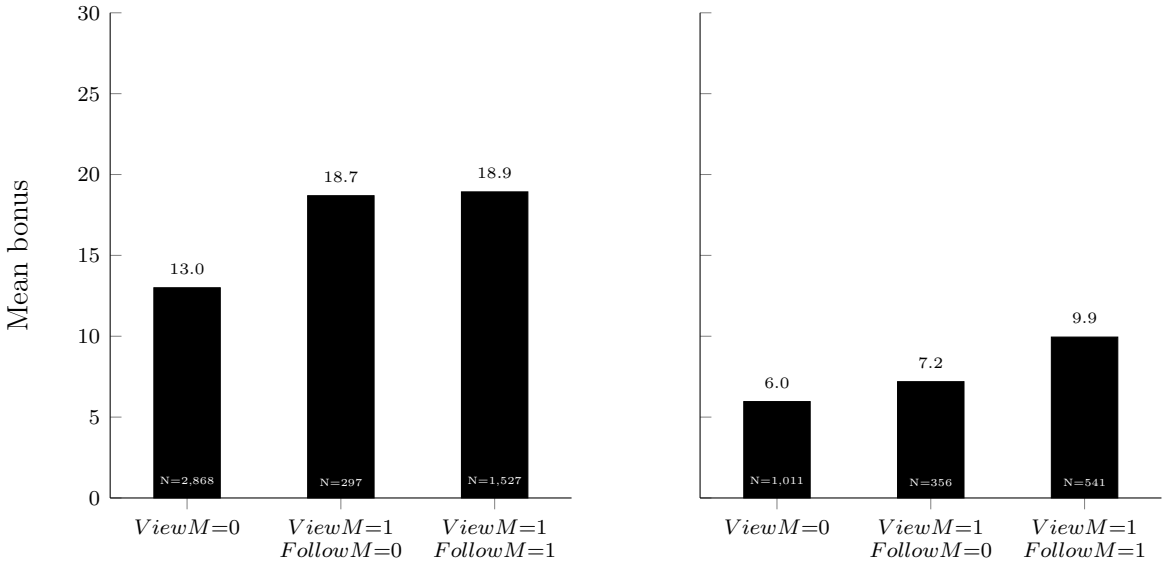
When the DM faces a conflict between the machine signal and a high-precision human signal (i.e., the DM should override s_M), the results show that *Delegation* increases algorithmic over-reliance ($\beta_1 + \beta_3 = -0.067, p = 0.08$), providing support for Hypothesis 2A. However, when the machine signal conflicts with a low-precision human signal (i.e., the DM should follow s_M), *Delegation* decreases algorithm reliance ($\beta_1 = -0.073, p = 0.05$), providing no support for Hypothesis 2B. For the less interesting cases in which the two signals align, *Delegation* increases the number of blatantly incorrect choices (i.e., overriding both signals) when the human signal is highly precise ($\beta_1 + \beta_3 = -0.067, p = 0.02$), and when it is of low precision ($\beta_1 = -0.093, p < 0.01$).

Overall, we find that *Delegation* lowers decision quality in all informational states $\mathbf{s}$, even when the DM faces a more precise and conflicting machine signal, i.e., when our theory predicts that delegation increases machine adherence and thus decision quality (Hypothesis 2B). In an attempt to reconcile the data with our theory, recall that procedurally fairness-oriented managers seek to distinguish ex-ante good and bad decisions, but cannot perfectly do so because they cannot observe the DM's private signal. As a result, managers reward (punish) some ex-ante poor (good) decisions, which effectively shrinks the strength of incentives for the DM in *Delegation*: for any given $\mathbf{s}$, the expected payoff of the correct choice is weaker than in the absence of delegation. This *incentive-contraction* mechanism of delegation lowers the probability that a noisy DM makes the correct decision, and it might mute the main mechanism of our study, i.e, *adherence-dependent* bonus payments that increase the incentives for the DM to follow the machine signal *if* they are sufficiently salient to the DM. As the environment of Study 3 blurs the link between machine adherence and bonuses, DMs may not respond to adherence-dependent incentives, such that the incentive-contraction effect dominates and results in worse decisions in all informational states (which our data suggests). It is also possible that managers, facing a cognitively more challenging task,[22] do not condition their bonus payments on the DM's machine-adherence as strongly (or at all). We study this possibility next.

7.2.2. Managers: Bonus Payments. Figure 6 displays bonus payments for *Delegation*. As before, the data suggests outcome-dependent and adherence-dependent bonus patterns, but the descriptive data has to be interpreted with caution. The main empirical challenge in assessing adherence-dependence stems from our setting, in which managers determine bonus payments based on five task outcomes. Managers may choose to view the machine signal (to assess algorithm adherence) for any subset of the five tasks, and even if they view all machine signals for tasks with a bad outcome, they may find that the DM followed the machine on one task but overrode it on another. This complicates inference both for managers and for us: it weakens the direct link between bonus payments, task-by-task outcomes, and algorithm adherence. To address these challenges, we provide

[22] For example, they could observe two bad outcomes, one after a follow and one after an override.

Figure 6 Bonus Payments to DM



(a) $BadOutcome = 0$ (b) $BadOutcome = 1$

in Appendix C.3.1 a correlated random-effects model (Mundlak 1978), and report here only the main results. With reference to Table EC.8, we make several observations that align with earlier results.

In terms of outcome-dependence, consistent with Hypothesis 1C, each additional bad task reduces the bonus (-6.378, $p < 0.01$). More importantly, we find evidence of adherence-dependent bonus payments for bad outcomes. We stratify the data by the number of bad task outcomes in the round and find that, for each bad task outcome, managers lower the bonus payments more after they observe an unsuccessful machine override, when the round overall resulted in one bad outcome (-0.510, $p = 0.53$), two bad outcomes (-1.168, $p = 0.04$), and three bad outcomes (-0.496, $p < 0.01$). Although not immediately predicted by our theoretical model, these (non-)effects make intuitive sense in light of our previous results. The manager might be less inclined to engage in the cognitive operations required to condition bonus on machine adherence when a single bad outcome (while salient) does not reduce their average payoff much, but more inclined when two or three bad outcomes reduce the payoff substantially.[23] For good outcomes, as in Studies 1 and 2, we find no evidence for adherence-dependent bonus payments.

7.2.3. Managers: Mechanism Evidence (from $ViewM$ data). We briefly study managers' decisions to view the machine signal s_M. First, managers view s_M in 41% of all decisions, supporting the assumption that procedurally fairness-minded managers want to take the machine signal into account when evaluating the DM's decision. Second, to examine whether the decision to view

[23] More than three bad outcomes result in a negative payoff, such that no bonus can be paid.

the machine signal itself depends on outcomes, we estimate a linear probability model, $ViewM = \beta_1 BadOutcome + \beta_2 Round + \beta_3(BadOutcome \times Round)$; see Table EC.10 for the results. Managers are, on average, more likely to view the machine signal after a bad outcome ($\beta_1 = 0.082, p = 0.087$). Moreover, while the tendency to view the machine signal following good outcomes decreases over time ($\beta_2 = -0.007, p = 0.02$), it remains high following bad outcomes ($\beta_2 + \beta_3 = -0.005, p = 0.16$). This asymmetric viewing pattern aligns with our assumptions about outcome-dependent weights ω_b and ω_g in Section 3.4.

7.2.4. Does Algorithm Adherence "Pay Off"? To properly link bonus payments to decisions, we return to the DM's perspective. Our data again confirms the prediction implied by Proposition 1: despite anticipating possible blame when chance yields a bad outcome, in equilibrium, the DM still is more likely to (correctly) override the machine signal. Specifically, when signals conflict, incorrectly following the less precise machine signals reduces the DM's average bonus by 3.03 points.[24]

8. Discussion and Conclusion

While algorithmic under- and over-reliance is often explained by cognitive biases, difficulty assessing an algorithm's true quality, or strategic (agency) behavior under delegation, we show a new mechanism: algorithm mis-reliance can arise endogenously from organizational context, through the interaction between a manager and a decision maker. We identify recommendation-dependent preferences that induce DMs to over-rely on algorithmic input when they possess more precise private evidence that contradicts the algorithm. Importantly, this is the first study to endogenize these preferences in a coherent delegated decision-making framework, *and* to test them empirically, allowing us to precisely measure how strongly they arise in institutional settings where a manager delegates a decision task to an AI-assisted DM. Because we study manager-DM interactions in a setting where standard agency reasons for over- or under-reliance do not apply, the negative performance impact of delegation arise solely from behavioral reasons that we model in a framework with fairness-minded managers and noisy DMs.

8.1. Behavioral Mechanisms - How does Algorithm Adherence Arise?

Our data provides empirical evidence that is consistent with the key predictions, and the underlying behavioral mechanisms, of our model (Section 3). The manager is concerned with procedural fairness and hence values the ex ante quality of the DM's choice, as reflected in the DM's expected payoff. However, because the DM is boundedly rational and may make mistakes, an override is only a noisy signal of the DM's private information. The manager therefore infers the expected payoff faced by the

[24] This is based on a simple regression $BonusReceive_{it} = \beta_0 + \beta_1 FollowM_{it} + \varepsilon_{it}$ with standard errors clustered at the DM level. We find that $\beta_1 = -3.03$ ($p = 0.08$). See Table EC.11 for details.

DM from the only observable indicator, namely whether the machine was overridden and the outcome was bad. Since a mistaken override is very costly, while a correct override yields only a modest gain relative to the corresponding outcomes when the DM follows the machine, observing an override lowers the manager's assessment of the expected payoff, and hence the bonus. Omission–commission bias reinforces this penalty: because overriding is an act of commission, the bonus falls more after a bad override than it rises, if at all, after a good one. Overall, managers appear to blame DMs for bad outcomes, even though they know that the DM likely made the correct decision given the available information. The key implication of these behavioral mechanisms is that managers create incentives for a DM to (incorrectly) follow the machine even when the DM has more precise private information that points in the opposite direction.

8.2. Performance Effects - When does Algorithm Adherence Matter?

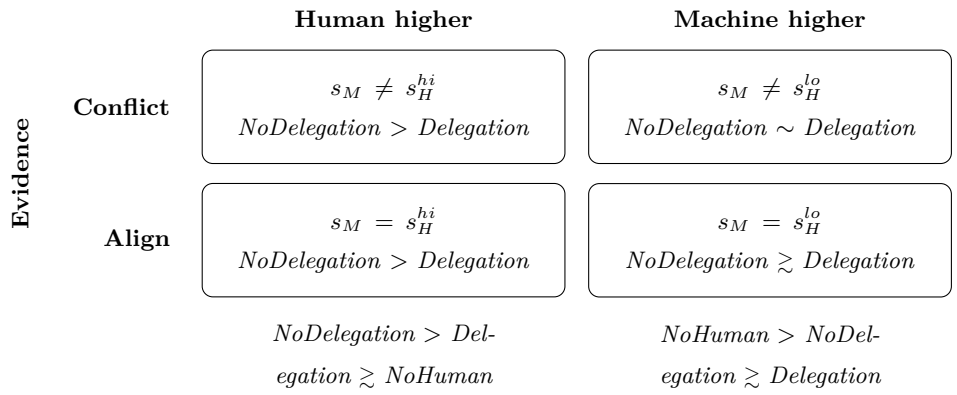
Our studies were specifically designed, first and foremost, to tease apart subtle behavioral mechanisms that might be at play when firms delegate machine-assisted decisions to human DMs. But our results also allow for cautious predictions of the conditions under which decision delegation to humans and/or machines affect performance. Specifically, in light of fast improvements of prediction technology, how might our conclusions generalize beyond the specific setting and parameters of our studies?

Figure 7 summarizes the main patterns from all three studies, decomposing performance by informational state (succinctly captured in our model by $\mathbf{s}$; see Table EC.12 for more detail). We find that *Delegation* deteriorates performance when the DM has precise private information that contradicts the machine signal ($s_M \neq s_H^{hi}$) i.e., precisely in the informational state that provides the most compelling (theoretical) reason to keep a human in the decision making loop in the first place. Because *Delegation* does not unambiguously lead to the predicted increase in machine reliance when it is desirable ($s_M \neq s_H^{lo}$), and increases the DMs decision errors when signals align ($s_M = s_H^{\theta}$), the overall effect of delegating machine-assisted tasks is negative in all studies. Notably, since we purposefully removed from our setting any standard agency reasons for why delegation might fail to achieve first-best performance, we can attribute performance loss entirely to behavioral reasons. This then leaves open an important question: how can managers leverage human expertise without triggering harmful human behavior?

8.3. Performance Effects - What can Managers Do?

Towards an answer to this question, *if* human bias outweighs the potential of human value, the firm might consider removing the human DM from the decision process entirely. Figure 7 (Table EC.12) provides a performance comparison with such a *NoHuman* policy in which the decision always follows the machine signal. By construction, *NoHuman* cannot tap into valuable human expertise, but it eliminates bias from a human as well behavioral incentive effects that arise from the interaction between

Figure 7 Performance by Informational State: *Delegation*, *NoDelegation*, *NoHuman*



a human manager and DM in *Delegation*. We find that, while *Delegation* falls short of a *NoHuman* policy in our data, *NoDelegation* outperforms it. In other words, human DMs in our *NoDelegation* treatment manage to capture some of the theoretical value that, for our studies' parameters, derives from informational states in which a better-informed human overrides a less precise machine signal. Yet, while encouraging, this result still leaves open the question of how firms could recap some of the value of keeping a human DM in the (delegated) decision loop.

One possible direction is to design a process that preserves the very reason a human DM belongs in the loop (i.e., access to relevant tacit knowledge) while mitigating the frictions that arise when that knowledge is non-codifiable and inherently unobservable. In our setting, the manager cannot observe the DM's private information, and this informational gap is precisely what gives rise to the behavioral mechanisms we predict and document in the data. This suggests a natural remedy: require the DM to explicitly communicate their private information to the manager. Doing so, however, calls for a carefully designed process that accounts for incentives and behavioral responses. For example, if managers require DMs to report their judgment, DMs can always justify their choices to align with the machine's predictions, unless those predictions are clearly incorrect. In other words, whatever the DMs communicates *after* the decision may amount to cheap talk. Instead, the manager could require DMs to record their own assessment *before* accessing the AI's prediction. Afterward, the DM may choose whether to follow or override the machine's recommendation. Since the AI prediction is not yet revealed at the time of the initial assessment, the DM cannot "hide behind" the machine to avoid blame, thereby reducing the incentive for over-reliance.[25] Our studies suggest that these remediation approaches may be effective: when the manager can observe the DM's private knowledge in the *DelegationObs* treatment of Study 2, performance loss from delegation disappears.

[25] In fact, when both the DM's judgment and the machine's prediction are informative, the AI is more likely a priori to align with the DM's assessment, which creates an incentive to report truthfully. While promising, however, this approach places additional demands on employees and may increase overall processing time. As such, delegation introduces a trade-off between accuracy and efficiency: if managers prioritize accuracy, they must be prepared for potentially slower processing.

8.4. Organizational Implications and Predictions.

Overall, our results highlight fundamental behavioral issues at the interface between DMs and those who shape the DMs' organizational context, including some implications that are outside the scope of what we theoretically model and empirically measure. For example, our study provides a conceptual basis of low employee morale that may ensue when decision makers feel pressured into following algorithmic advice that goes against their own better judgment (documented, e.g., in Bannon 2023).

Our results also yield managerial implications for the implementation of AI in organizations. Managers should be cautious not to view the successful implementation of AI as a purely technological challenge. Our results suggest that part of AI's value derives from the flexibility of allowing human decision makers to override algorithmic advice when appropriate, rather than relying on decision rules that rigidly enforce AI recommendations. Yet this same human involvement can also dissipate a substantial share of AI's potential value through behavioral frictions, stemming not only from general human biases but also, importantly, from delegation-related distortions that arise when humans interact with algorithmic recommendations. Taken together, these findings suggest that organizations seeking to capture the full value of AI must focus not only on improving and integrating algorithms, but also on designing processes and incentives that mitigate these behavioral frictions.

A related, and subtle, implication of the behavioral mechanism we document is that organizations may (unknowingly) adopt machines that do not necessarily outperform human expertise. When overrides are rare, because DMs follow the machine to avoid potential blame, the organization observes few counterfactual human decisions. This make it difficult to evaluate whether the machine truly outperforms human judgment, and inferior technologies may therefore persist.

References

Albright A (2024) The hidden effects of algorithmic recommendations. *Preprint. Last accessed March 26* URL https://apalbright.github.io/pdfs/albright-algo-recs-PAPER.pdf.

Anderson RA, Nichols S, Pizarro DA (2024) Praise is for actions that are neither expected nor required. *Personality and Social Psychology Bulletin* 01461672241289833.

Artinger FM, Artinger S, Gigerenzer G (2019) C. Y. A.: Frequency and causes of defensive decisions in public administration. *Business Research* 12(1):9–25, URL http://dx.doi.org/10.1007/s40685-018-0074-2.

Artinger FM, Marx-Fleck S, Junker NM, Gigerenzer G, Artinger S, van Dick R (2025) Coping with uncertainty: The interaction of psychological safety and authentic leadership in their effects on defensive decision making. *Journal of Business Research* 190:115240, URL http://dx.doi.org/10.1016/j.jbusres.2025.115240.

Athey SC, Bryan KA, Gans JS (2020) The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings* 110:80–84.

Balakrishnan M, Ferreira KJ, Tong J (2026) Human-algorithm collaboration with private information: Naïve advice-weighting behavior and mitigation. *Management Science* 72(1):265–284.

Banker S, Khetani S (2019) Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing* 38(4):500–515.

Bannon L (2023) When AI overrules the nurses caring for you. *The Wall Street Journal* URL https://www.wsj.com/articles/ai-medical-diagnosis-nurses-f881b0fe?st=rgc9ddwyuyio3bv&reflink=desktopwebshare_permalink.

Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Baron J, Hershey JC (1988) Outcome bias in decision evaluation. *Journal of Personality and Social Psychology* 54(4):569–579, URL http://dx.doi.org/10.1037/0022-3514.54.4.569.

Bastani H, Cachon GP (2025) The human-AI contracting paradox. *SSRN Electronic Journal* URL http://dx.doi.org/10.2139/ssrn.5962739, 31 pages. Date Written: December 24, 2025; Posted: December 29, 2025; Last revised: December 24, 2025.

Bauer K, von Zahn M, Hinz O (2023) Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research* 34(4):1582–1602.

Beer R, Qi A, Rios I (2026) Behavioral externalities of process automation. *Management Science* 72(1):575–593.

Bolton G, Brandts J, Ockenfels A (2005) Fair procedures: Evidence from games involving lotteries. *The Economic Journal* 115:1054–1076.

Bostyn DH, Roets A (2016) The morality of action: The asymmetry between judgments of praise and blame in the action–omission effect. *Journal of Experimental Social Psychology* 63:19–25.

Boyacı 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.

Brock WA, Durlauf SN (2001) Discrete choice with social interactions. *The Review of Economic Studies* 68(2):235–260.

Brockner J, Wiesenfeld BM (1996) An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures. *Psychological bulletin* 120(2):189.

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5(CSCW1):1–21.

Caro F, de Tejada Cuenca AS (2023) Believing in analytics: Managers' adherence to price recommendations from a DSS. *Manufacturing & Service Operations Management* 25(2):524–542.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825.

Chassang S, Zehnder C (2016) Rewards and punishments: Informal contracting through social preferences. *Theoretical Economics* 11(3):1145–1179.

Chen DL, Schonger M, Wickens C (2016) oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

Cox JC, Servátka M, Vadovič R (2017) Status quo effects in fairness games: reciprocal responses to acts of commission versus acts of omission. *Experimental Economics* 20(1):1–18.

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, URL `http://dx.doi.org/10.1177/00222437251332898`.

Dai T, Taylor T (2025) Designing enterprise ai systems: Hallucination, creativity, and moral hazard. *SSRN Electronic Journal* URL `http://dx.doi.org/10.2139/ssrn.5996714`, date Written: December 31, 2025; Posted: January 5, 2026.

Dana J, Weber RA, Kuang JX (2007) Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory* 33(1):67–80.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Eisenhardt KM (1989) Agency theory: An assessment and review. *Academy of Management Review* 14(1):57–74, URL `http://dx.doi.org/10.5465/amr.1989.4279003`.

Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114:817–868.

Feier T, Gogoll J, Uhl M (2022) Hiding behind machines: Artificial agents may help to evade punishment. *Science and Engineering Ethics* 28(2):19.

Feldman MS, Pentland BT (2003) Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly* 48(1):94–118, URL `http://dx.doi.org/10.2307/3556620`.

Feldman MS, Pentland BT, D'Adderio L, Dittrich K, Rerup C, Seidl D, eds. (2021) *The Cambridge Handbook of Routine Dynamics* (Cambridge: Cambridge University Press).

Froomkin AM, Kerr IR, Pineau J (2019) When AIs outperform doctors: Confronting the challenges of a tort-induced over-reliance on machine learning. *Arizona Law Review* 61(1):33–99, URL `https://journals.librarypublishing.arizona.edu/arizlrev/article/id/6843/`.

Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly* 45(3):1527–1556.

Gill MJ (2026) Rule breaking in organizations: An integrative review. *Academy of Management Annals* 20(1):285–314, URL `http://dx.doi.org/10.5465/annals.2023.0170`, published online 12 December 2025.

Grand-Clément J, Pauphilet J (2024) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science, ePub ahead of print June 10*, URL `http://dx.doi.org/10.1287/mnsc.2023.01851`.

Grossman Z (2014) Strategic ignorance and the robustness of social preferences. *Management Science* 60(11):2659–2665.

Guan X, Qi A, Wang S (2025) Why the best machine may not be the best: Incentivizing human-machine collaboration. Working Paper 5283571, SSRN, URL `http://dx.doi.org/10.2139/ssrn.5283571`, date written: June 1, 2025; posted to SSRN June 12, 2025; last revised June 17, 2025.

Guglielmo S, Malle BF (2019) Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one* 14(3):e0213544.

Gurdal MY, Miller JB, Rustichini A (2013) Why blame? *Journal of Political Economy* 121(6):1205–1247.

Hey J, Orme C (1994) Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6):1291–1326.

Hou YTY, Jung MF (2021) Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–25.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Lehmann CA, Haubitz CB, Fügener A, Thonemann UW (2022) The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management* 31(9):3419–3434.

Lerner JS, Tetlock PE (1999) Accounting for the effects of accountability. *Psychological Bulletin* 125(2):255–275, URL `http://dx.doi.org/10.1037/0033-2909.125.2.255`.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Loomes G, Moffat P, Sugden R (2002) A microeconometric test of alternative stochastic theories of risky choice. *The Journal of Risk and Uncertainty* 24(2):103–130.

Luce RD (2005) *Individual choice behavior: A theoretical analysis* (Courier Corporation).

Luong A, Kumar N, Lang KR (2020) Algorithmic decision-making: Examining the interplay of people, technology, and organizational practices through an economic experiment. *Baruch College Zicklin School of Business Research Paper* (2020-02):03.

McKelvey RD, Palfrey TR (1995) Quantal response equilibria for normal form games. *Games and Economic Behavior* 10(1):6–38.

McKinsey & Company (2023) The state of AI in 2023: Generative AI's breakout year. *McKinsey & Company* URL `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year`.

McLaughlin B, Spiess J (2025) Algorithmic assistance with recommendation-dependent preferences. URL `https://arxiv.org/abs/2208.07626`.

Morrison EW (2006) Doing the job well: An investigation of pro-social rule breaking. *Journal of Management* 32(1):5–28, URL `http://dx.doi.org/10.1177/0149206305277790`.

Mundlak Y (1978) On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society* 69–85.

Ouchi WG (1979) A conceptual framework for the design of organizational control mechanisms. *Management Science* 25(9):833–848, URL `http://dx.doi.org/10.1287/mnsc.25.9.833`.

Salvato C, Rerup C (2018) Routine regulation: Balancing conflicting goals in organizational routines. *Administrative Science Quarterly* 63(1):170–209.

Snyder C, Keppler S, Leider S (2026) Algorithm reliance: Fast and slow. *Management Science* 72(1):368–385.

Spencer B, Rerup C (2024) The dynamics of inferential interpretation in experiential learning: Deciphering hidden goals from ambiguous experience. *Administrative Science Quarterly* 69(4):962–1005.

Spranca M, Minsk E, Baron J (1991) Omission and commission in judgment and choice. *Journal of experimental social psychology* 27(1):76–105.

Tetlock PE (1985) Accountability: The neglected social context of judgment and choice. *Research in Organizational Behavior*, volume 7, 297–332 (JAI Press), URL `https://time.com/wp-content/uploads/2015/05/1985_accountability_neglected_social_context_of_judegment.pdf`.

Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R (2023) Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1):1–38.

Weitzner G (2024) Reputational algorithm aversion. *arXiv* 2402.

Wooldridge JM (2019) Correlated random effects models with unbalanced panels. *Journal of Econometrics* 211(1):137–150.

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. *Journal of Behavioral Decision Making* 32(4):403–414.

Online Appendix

Managerial Blame and Defensive Adherence: How Organizations Create AI Over-Reliance.

A. Supplementary Theory and Proofs

We first present supplementary theoretical results and subsequently collect all proofs.

A.1. Supplementary Theory

A.1.1. DM's Posterior Belief. The DM updates their belief about the true state of the world, $\mu_{\mathbf{s}} \equiv \mu^{\theta}_{s_M,s_H} \equiv \Pr(Y=R|s_M,s_H,\theta)$, using Bayes' rule. Given the prior $\mu_0 = 1/2$ we have:

$$\begin{aligned}
\mu^{hi}_{R,R} &= 1-\mu^{hi}_{B,B} = \frac{q_{s_M}\cdot q^{hi}_{s_H}}{q_{s_M}\cdot q^{hi}_{s_H}+(1-q_{s_M})(1-q^{hi}_{s_H})} > 1/2\\
\mu^{hi}_{B,R} &= 1-\mu^{hi}_{R,B} = \frac{(1-q_{s_M})\cdot q^{hi}_{s_H}}{(1-q_{s_M})\cdot q^{hi}_{s_H}+q_{s_M}\cdot(1-q^{hi}_{s_H})} > 1/2\\
\mu^{lo}_{R,R} &= 1-\mu^{lo}_{B,B} = \frac{q_{s_M}\cdot q^{lo}_{s_H}}{q_{s_M}\cdot q^{lo}_{s_H}+(1-q_{s_M})(1-q^{lo}_{s_H})} > 1/2\\
\mu^{lo}_{R,B} &= 1-\mu^{lo}_{B,R} = \frac{q_{s_M}\cdot(1-q^{lo}_{s_H})}{q_{s_M}\cdot(1-q^{lo}_{s_H})+(1-q_{s_M})\cdot q^{lo}_{s_H}} > 1/2
\end{aligned} \tag{EC.1}$$

Where $q^{hi}_{s_H}$, $q^{lo}_{s_H}$, and q_{s_M} represent the precision of the Human and Machine signals.

A.1.2. Social Preference Model Formulation. Our specification in Eq. 2 adopts a piecewise linear formulation for the manager's disutility from fairness concerns. When $\gamma = 0$, the model reduces to a standard Fehr–Schmidt inequity aversion specification based solely on ex-post payoff differences.

More generally, the manager's utility can be written as

$$\mathcal{U}^{ma} = r(a,y) - b - (\alpha+\gamma)\,\Psi\big(x,\phi(b)\big),$$

where the fairness-relevant index is given by

$$x := \lambda\, r(a,y) + (1-\lambda)\,\tilde{r}(s_M,a,y), \qquad \lambda := \frac{\alpha}{\alpha+\gamma} \in (0,1),$$

and $\phi(b)$ is a fairness benchmark that is increasing in the bonus b. In our baseline specification, we set $\phi(b) = 2b$, which corresponds to an equitable split in which the DM receives 50% of the total surplus. The term $\Psi(x,\phi(b))$ captures the manager's disutility from perceived unfairness, measured as a function of the deviation between the fairness-relevant index x and the benchmark $\phi(b)$. Our baseline specification corresponds to $\Psi(x,\phi(b)) = |x-\phi(b)|$.

For the qualitative structure of the results, it is sufficient that (i) $\phi(b)$ is increasing in b; (ii) $\Psi(x,\phi(b)) = \psi(x-\phi(b))$ for some even, strictly increasing function ψ with $\psi(0)=0$; and (iii) $\alpha > 0$ and $\gamma > 0$, so that fairness concerns depend jointly on realized outcomes and ex-ante procedural quality.

A.1.3. The Equilibria. The following two lemmas summarize the manager's and the DM's best responses, and the subsequent lemma establishes the resulting equilibria.

LEMMA EC.1 **(Manager's best response)**. *In optimality,*

1. *The manager pays a bonus b^* to the DM, such that*
$$b^* = \begin{cases} \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right), & \alpha+\gamma > \frac{1}{2} \\ 0, & \alpha+\gamma \leq \frac{1}{2} \end{cases}$$
2. *The manager prefers the action $\overline{a} \equiv \arg\max_a \mathbb{E}_Y[r(a,y)|\boldsymbol{s}]$. In the same vane, we denote by $\underline{a}$ the alternative action.*

Part 1 of the lemma stems form the first-order optimality conditions for the manager's optimization problem, which imply when the manager has sufficiently strong fairness concerns, i.e., $\alpha+\gamma > \frac{1}{2}$, they would pay a bonus $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right) > 0$ to the DM, and a bonus $b^* = 0$ otherwise. Given b^*, the manager's utility is maximized when the DM takes the action $\overline{a} \equiv \arg\max_a \mathbb{E}_Y[r(a,y)|\mathbf{s}]$, hence the part 2 of the lemma.

The stochastic choice model defined in Eq. 1 implies that the DM may not always take the best response, i.e., $a^* \equiv \arg\max_a \mathbb{E}_Y[\mathcal{U}^{dm}|\mathbf{s}]$, however, better decisions are made more often than worse ones. In the following, we characterize the DM' best response.

LEMMA EC.2 **(DM's best response)**. *In optimality, the DM's best response is:*
$$a^* = \begin{cases} \arg\max_a \mathbb{E}_Y[b^*|\boldsymbol{s}], & b^* > 0 \\ \textit{any } a \in \{\overline{a}, \underline{a}\}, & b^* = 0 \end{cases}$$

Recall that $\mathcal{U}^{dm} = b$, then the lemma stems form the first-order optimality conditions for the DM's optimization problem. When $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right) > 0$, then the DM's best response is to maximize the bonus payment. When $b^* = 0$, the DM would be indifferent between the two actions.

After the true state is realized, the manager updates their belief about the DM's posterior belief at the time of decision making and uses this belief to evaluate the ex-ante quality of the DM's decision. Let $\mu_{\overline{a}}$ and $\mu_{\underline{a}}$ denote the manager's beliefs about the ex-ante probability that the DM's chosen decision matches the true state when the DM chooses $\overline{a}$ and $\underline{a}$, respectively. Let $\omega_{\overline{a}}$ and $\omega_{\underline{a}}$ denote the corresponding omission–commission scale parameters, which depend on whether the DM follows or overrides the machine signal. Combining the best responses of the manager and the DM, the following lemma characterizes the possible equilibria of the game.

LEMMA EC.3 **(Equilibria)**. *Let $\kappa = \frac{(\omega_{\underline{a}}\mu_{\underline{a}} - \omega_{\overline{a}}\mu_{\overline{a}})(\overline{r}-\underline{r}) + (\omega_{\underline{a}} - \omega_{\overline{a}})\underline{r}}{(2\mu_{\mathbf{s}}-1)(\overline{r}-\underline{r})}$. The following equilibria may arise in the game:*

1. *If $\alpha+\gamma \leq \frac{1}{2}$, then $b^* = 0$ and $a^* \in \{\overline{a}, \underline{a}\}$.*
2. *If $\alpha+\gamma > \frac{1}{2}$ and $\frac{\alpha}{\gamma} \leq \kappa$, then $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right)$ and $a^* = \underline{a}$.*
3. *If $\alpha+\gamma > \frac{1}{2}$ and $\frac{\alpha}{\gamma} > \kappa$, then $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right)$ and $a^* = \overline{a}$.*

For the rest of the manuscript, we assume that $\alpha+\gamma > \frac{1}{2}$ and $\frac{\alpha}{\gamma} > \overline{\kappa}$, where $\overline{\kappa} := \sup_{\omega,\mu} \kappa$.

A.1.4. DM Fairness Concerns In the main model, we assume that the DM's utility is given by $\mathcal{U}^{dm} = b$. This assumption is made for simplicity and expositional clarity, as it isolates the role of the manager's distributional and procedural concerns. We now show that allowing the DM to share the same fairness concerns does not alter the qualitative equilibrium structure.

Suppose instead that the DM's utility is given by

$$\mathcal{U}^{dm} = b - \Big|\alpha\big(2b - r(a,y)\big) + \gamma\big(2b - \tilde{r}(s_M, a, y)\big)\Big|,$$

so that the DM is additionally sensitive to unfair treatment by the manager in exactly the same way as the manager evaluates fairness.

COROLLARY EC.1 **(Robustness to DM fairness concerns)**. *Suppose the DM's utility is given by*

$$\mathcal{U}^{dm} = b - \Big|\alpha\big(2b - r(a,y)\big) + \gamma\big(2b - \tilde{r}(s_M, a, y)\big)\Big|.$$

Then the equilibrium characterization in Lemma EC.3 applies whenever $\alpha+\gamma > \frac{1}{2}$. If $\alpha+\gamma \le \frac{1}{2}$, the equilibrium is unique and satisfies $b^* = 0$ and $a^* = \underline{a}$.

The intuition is straightforward. When $b^* = 0$, the DM's material payoff is fixed at zero and they cannot affect it. The DM's utility therefore depends only on the fairness term. In this case, they minimize their disutility from unfairness by choosing the action that minimizes the manager's expected payoff, which yields $a^* = \underline{a}$. In contrast, when $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M, a, y)\right)$, the fairness disutility is exactly neutralized, so that the absolute-value term equals zero. The DM's objective then reduces to maximizing the bonus payment b^*, as in the baseline specification. Accordingly, the equilibrium characterization for $\alpha+\gamma > \frac{1}{2}$ in Lemma EC.3 remains unchanged.

Since throughout the paper we assume $\alpha+\gamma > \frac{1}{2}$ and $\frac{\alpha}{\gamma} > \overline{\kappa}$, the equilibrium of interest is unaffected by introducing fairness concerns for the DM.

A.1.5. Stochastic decision making. Next, we impose a mild regularity condition on the DM's stochastic choice formulation that guarantees well behaved comparative statics in the resulting QRE.

LEMMA EC.4 **(Monotonicity of Choice Probability in β)**. *Suppose that $\beta > \underline{\beta}$, the DM's choice probability, $\mathcal{P}(A = \overline{a}|\boldsymbol{s})$, is strictly decreasing in β.*

COROLLARY EC.2. *The derivative of the optimal choice probability, $\mathcal{P}(A = \overline{a}|\boldsymbol{s})$, with respect to any variable x, is uniformly bounded:*

$$\left|\frac{d\mathcal{P}}{dx}\right| \le L(\beta)$$

where $L(\beta)$ is strictly decreasing in β.

A.1.6. Ex-ante decision quality. We next analyze how the manager's assessment of the DM's decision systematically depends on whether the DM had chosen to follow or override the machine signal s_M.

Note first that assessing the expected payoff $\mathbb{E}_Y[r(a,y) \mid s_M]$ of a decision[26] is challenging, because the manager cannot observe the DM's full information $\mathbf{s}$ *and* because of the DM's noisy decision process $\mathcal{P}(\mathbf{s}) \equiv$

[26] Formally, the expectation should be written as $\mathbb{E}_Y[r(a,Y) \mid y, s_M]$, since the realization of y is observed ex post. We use the abbreviated notation for expositional convenience, and maintain this convention throughout.

$\Pr(A=\overline{a}|\mathbf{s})$ from Eq. 1.[27] To evaluate a decision in this environment, the manager must rely only on the observables (a, y, s_M) to infer the DM's information $\mathbf{s}$. Formally, this requires forming a belief over the human signal and its precision, $\Pr(s_H, \theta \mid s_M, a, y)$, using Bayes' rule:

$$\Pr(s_H, \theta \mid s_M, a, y) = \frac{\Pr(A=a|\mathbf{s}) \Pr(s_M, s_H|y, \theta) \Pr(\Theta=\theta) \Pr(Y=y)}{\Pr(s_M, a, y)} \tag{EC.2}$$

We use the following notation to express the manager's beliefs explicitly for all possible cases:

$$p_{ij}^k \equiv \Pr(A=s_H, \theta=k|i,j), \quad q_{ij}^k \equiv \Pr(A\neq s_H, \theta=k|i,j)$$

where $i=1$ indicates $a=s_M$ and $i=0$ indicates $a\neq s_M$, while $j=1$ indicates $y=a$ and $j=0$ indicates $y\neq a$. Furthermore, define

$$\mathcal{P}^{hi}_{\text{align}} := \mathcal{P}(s_M, s_H=s_M, \theta=hi), \quad \mathcal{P}^{hi}_{\text{conf}} := \mathcal{P}(s_M, s_H\neq s_M, \theta=hi)$$
$$\mathcal{P}^{lo}_{\text{align}} := \mathcal{P}(s_M, s_H=s_M, \theta=lo), \quad \mathcal{P}^{lo}_{\text{conf}} := \mathcal{P}(s_M, s_H\neq s_M, \theta=lo)$$

Finally, the manager's belief over the human signal and its precision, $\Pr(s_H, \theta \mid s_M, a, y)$, for all possible combinations of (s_M, a, y) is given by:

$$p^{\text{hi}}_{11} = \frac{\mathcal{P}^{hi}_{\text{align}} q^{hi}_{s_H} \mu_\theta}{\mathcal{P}^{hi}_{\text{align}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}})(1-q^{hi}_{s_H})\mu_\theta + \mathcal{P}^{lo}_{\text{align}} q^{lo}_{s_H}(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}}(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3a}$$

$$q^{\text{hi}}_{11} = \frac{(1-\mathcal{P}^{hi}_{\text{conf}})(1-q^{hi}_{s_H})\mu_\theta}{\mathcal{P}^{hi}_{\text{align}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}})(1-q^{hi}_{s_H})\mu_\theta + \mathcal{P}^{lo}_{\text{align}} q^{lo}_{s_H}(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}}(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3b}$$

$$p^{\text{lo}}_{11} = \frac{\mathcal{P}^{lo}_{\text{align}} q^{lo}_{s_H}(1-\mu_\theta)}{\mathcal{P}^{hi}_{\text{align}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}})(1-q^{hi}_{s_H})\mu_\theta + \mathcal{P}^{lo}_{\text{align}} q^{lo}_{s_H}(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}}(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3c}$$

$$q^{\text{lo}}_{11} = \frac{\mathcal{P}^{lo}_{\text{conf}}(1-q^{lo}_{s_H})(1-\mu_\theta)}{\mathcal{P}^{hi}_{\text{align}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}})(1-q^{hi}_{s_H})\mu_\theta + \mathcal{P}^{lo}_{\text{align}} q^{lo}_{s_H}(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}}(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3d}$$

$$p^{\text{hi}}_{01} = \frac{\mathcal{P}^{hi}_{\text{conf}} q^{hi}_{s_H} \mu_\theta}{\mathcal{P}^{hi}_{\text{conf}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{align}})(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{lo}_{\text{conf}}) q^{lo}_{s_H}(1-\mu_\theta) + (1-\mathcal{P}^{lo}_{\text{align}})(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3e}$$

$$q^{\text{hi}}_{01} = \frac{(1-\mathcal{P}^{hi}_{\text{align}})(1-q^{hi}_{s_H})\mu_\theta}{\mathcal{P}^{hi}_{\text{conf}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{align}})(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{lo}_{\text{conf}}) q^{lo}_{s_H}(1-\mu_\theta) + (1-\mathcal{P}^{lo}_{\text{align}})(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3f}$$

$$p^{\text{lo}}_{01} = \frac{(1-\mathcal{P}^{lo}_{\text{conf}}) q^{lo}_{s_H}(1-\mu_\theta)}{\mathcal{P}^{hi}_{\text{conf}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{align}})(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{lo}_{\text{conf}}) q^{lo}_{s_H}(1-\mu_\theta) + (1-\mathcal{P}^{lo}_{\text{align}})(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3g}$$

$$q^{\text{lo}}_{01} = \frac{(1-\mathcal{P}^{lo}_{\text{align}})(1-q^{lo}_{s_H})(1-\mu_\theta)}{\mathcal{P}^{hi}_{\text{conf}} q^{hi}_{s_H} \mu_\theta + (1-\mathcal{P}^{hi}_{\text{align}})(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{lo}_{\text{conf}}) q^{lo}_{s_H}(1-\mu_\theta) + (1-\mathcal{P}^{lo}_{\text{align}})(1-q^{lo}_{s_H})(1-\mu_\theta)} \tag{EC.3h}$$

$$p^{\text{hi}}_{10} = \frac{\mathcal{P}^{hi}_{\text{align}}(1-q^{hi}_{s_H})\mu_\theta}{\mathcal{P}^{hi}_{\text{align}}(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}}) q^{hi}_{s_H} \mu_\theta + \mathcal{P}^{lo}_{\text{align}}(1-q^{lo}_{s_H})(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}} q^{lo}_{s_H}(1-\mu_\theta)} \tag{EC.3i}$$

$$q^{\text{hi}}_{10} = \frac{(1-\mathcal{P}^{hi}_{\text{conf}}) q^{hi}_{s_H} \mu_\theta}{\mathcal{P}^{hi}_{\text{align}}(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}}) q^{hi}_{s_H} \mu_\theta + \mathcal{P}^{lo}_{\text{align}}(1-q^{lo}_{s_H})(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}} q^{lo}_{s_H}(1-\mu_\theta)} \tag{EC.3j}$$

$$p^{\text{lo}}_{10} = \frac{\mathcal{P}^{lo}_{\text{align}}(1-q^{lo}_{s_H})(1-\mu_\theta)}{\mathcal{P}^{hi}_{\text{align}}(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}}) q^{hi}_{s_H} \mu_\theta + \mathcal{P}^{lo}_{\text{align}}(1-q^{lo}_{s_H})(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}} q^{lo}_{s_H}(1-\mu_\theta)} \tag{EC.3k}$$

$$q^{\text{lo}}_{10} = \frac{\mathcal{P}^{lo}_{\text{conf}} q^{lo}_{s_H}(1-\mu_\theta)}{\mathcal{P}^{hi}_{\text{align}}(1-q^{hi}_{s_H})\mu_\theta + (1-\mathcal{P}^{hi}_{\text{conf}}) q^{hi}_{s_H} \mu_\theta + \mathcal{P}^{lo}_{\text{align}}(1-q^{lo}_{s_H})(1-\mu_\theta) + \mathcal{P}^{lo}_{\text{conf}} q^{lo}_{s_H}(1-\mu_\theta)} \tag{EC.3l}$$

[27] The manager under full information $\mathbf{s}$ would be able to perfectly distinguish between ex-ante good and bad decisions, formally: $\mathbb{E}_Y[r(\overline{a}, y) \mid \mathbf{s}] > \mathbb{E}_Y[r(\underline{a}, y) \mid \mathbf{s}]$ for $y=a$ and $y\neq a$.

$$p_{00}^{\text{hi}} = \frac{\mathcal{P}_{\text{conf}}^{hi}(1-q_{s_H}^{hi})\mu_\theta}{\mathcal{P}_{\text{conf}}^{hi}(1-q_{s_H}^{hi})\mu_\theta + (1-\mathcal{P}_{\text{align}}^{hi})q_{s_H}^{hi}\mu_\theta + (1-\mathcal{P}_{\text{conf}}^{lo})(1-q_{s_H}^{lo})(1-\mu_\theta) + (1-\mathcal{P}_{\text{align}}^{lo})q_{s_H}^{lo}(1-\mu_\theta)} \tag{EC.3m}$$

$$q_{00}^{\text{hi}} = \frac{(1-\mathcal{P}_{\text{align}}^{hi})q_{s_H}^{hi}\mu_\theta}{\mathcal{P}_{\text{conf}}^{hi}(1-q_{s_H}^{hi})\mu_\theta + (1-\mathcal{P}_{\text{align}}^{hi})q_{s_H}^{hi}\mu_\theta + (1-\mathcal{P}_{\text{conf}}^{lo})(1-q_{s_H}^{lo})(1-\mu_\theta) + (1-\mathcal{P}_{\text{align}}^{lo})q_{s_H}^{lo}(1-\mu_\theta)} \tag{EC.3n}$$

$$p_{00}^{\text{lo}} = \frac{(1-\mathcal{P}_{\text{conf}}^{lo})(1-q_{s_H}^{lo})(1-\mu_\theta)}{\mathcal{P}_{\text{conf}}^{hi}(1-q_{s_H}^{hi})\mu_\theta + (1-\mathcal{P}_{\text{align}}^{hi})q_{s_H}^{hi}\mu_\theta + (1-\mathcal{P}_{\text{conf}}^{lo})(1-q_{s_H}^{lo})(1-\mu_\theta) + (1-\mathcal{P}_{\text{align}}^{lo})q_{s_H}^{lo}(1-\mu_\theta)} \tag{EC.3o}$$

$$q_{00}^{\text{lo}} = \frac{(1-\mathcal{P}_{\text{align}}^{lo})q_{s_H}^{lo}(1-\mu_\theta)}{\mathcal{P}_{\text{conf}}^{hi}(1-q_{s_H}^{hi})\mu_\theta + (1-\mathcal{P}_{\text{align}}^{hi})q_{s_H}^{hi}\mu_\theta + (1-\mathcal{P}_{\text{conf}}^{lo})(1-q_{s_H}^{lo})(1-\mu_\theta) + (1-\mathcal{P}_{\text{align}}^{lo})q_{s_H}^{lo}(1-\mu_\theta)} \tag{EC.3p}$$

This updating shapes the manager's assessment of the DM's ex-ante decision quality. Under full rationality, $\mathcal{P}(\mathbf{s}) = 1$, the manager would infer from a machine override (i.e., $a \neq s_M$) with certainty that the DM observed conflicting signals and followed a strong human signal. Under bounded rationality, $\mathcal{P}(\mathbf{s}) < 1$, this inference is noisy but the implication remains the same: the decision to override a machine signal is more diagnostic of following correctly a more precise human signal than is the decision to follow a machine signal, i.e.,

$$\Pr\left(s_H = a, \theta = hi | s_M, \underbrace{a \neq s_M}_{\text{override}}, y\right) > \Pr\left(s_H = a, \theta = hi | s_M, \underbrace{a = s_M}_{\text{follow}}, y\right). \tag{EC.4}$$

This mirrors the logic of models in career-concerns settings (e.g., Weitzner 2024), where high-ability DMs receive more precise signals than the machine and low-ability DMs less precise ones. In such settings, overriding the machine always leads to a higher bonus, as pay depends on the posterior probability of being high-ability.

Overrides of machine signals thus increase the manager's belief that the DM made a correct decision in a favorable decision environment $\mathbf{s}$. But the manager's assessment of a decision also takes into account its expected payoff $\mathbb{E}_Y[r(a,y) \mid \mathbf{s}]$, based on their belief about the human signal precision θ and its realization s_H. Formally,

$$\mathbb{E}_Y[r(a,y)|s_M = a, \theta = hi, s_H = s_M] = \mu_{RR}^{hi} \cdot \overline{r} + (1-\mu_{RR}^{hi}) \cdot \underline{r} \tag{EC.5a}$$

$$\mathbb{E}_Y[r(a,y)|s_M = a, \theta = hi, s_H \neq s_M] = (1-\mu_{BR}^{hi}) \cdot \overline{r} + \mu_{BR}^{hi} \cdot \underline{r} \tag{EC.5b}$$

$$\mathbb{E}_Y[r(a,y)|s_M = a, \theta = lo, s_H = s_M] = \mu_{RR}^{lo} \cdot \overline{r} + (1-\mu_{RR}^{lo}) \cdot \underline{r} \tag{EC.5c}$$

$$\mathbb{E}_Y[r(a,y)|s_M = a, \theta = lo, s_H \neq s_M] = \mu_{RB}^{lo} \cdot \overline{r} + (1-\mu_{RB}^{lo}) \cdot \underline{r} \tag{EC.5d}$$

$$\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = hi, s_H = s_M] = (1-\mu_{RR}^{hi}) \cdot \overline{r} + \mu_{RR}^{hi} \cdot \underline{r} \tag{EC.5e}$$

$$\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = hi, s_H \neq s_M] = \mu_{BR}^{hi} \cdot \overline{r} + (1-\mu_{BR}^{hi}) \cdot \underline{r} \tag{EC.5f}$$

$$\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = lo, s_H = s_M] = (1-\mu_{RR}^{lo}) \cdot \overline{r} + \mu_{RR}^{lo} \cdot \underline{r} \tag{EC.5g}$$

$$\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = lo, s_H \neq s_M] = (1-\mu_{RB}^{lo}) \cdot \overline{r} + \mu_{RB}^{lo} \cdot \underline{r} \tag{EC.5h}$$

If $\theta = lo$, regardless of the realization of the less precise human signal s_H, following the machine signal s_M has higher expected payoff than overriding it. Similarly, if $\theta = hi$, following s_M signal has higher expected payoff regardless of the realization of the more precise s_H: an ex-ante good decision ($a = s_H$) has higher expected payoff if it coincides with following an aligning machine signal ($s_M = s_H$), and an ex-ante bad

decision ($a \neq s_H$) has higher expected payoff if the decision follows the conflicting machine signal ($s_M \neq s_H$) rather than the DM overriding both signals. Formally,

$$\begin{aligned}
&&\text{OVERRIDE:} && \text{FOLLOW:} \\
s_H = a, \theta = hi: \quad & \underbrace{\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = hi, s_H \neq s_M]}_{\textit{ex-ante correct}} < \underbrace{\mathbb{E}_Y[r(a,y)|s_M = a, \theta = hi, s_H = s_M]}_{\textit{ex-ante correct}}, && \text{(EC.6a)} \\
s_H \neq a, \theta = hi: \quad & \underbrace{\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = hi, s_H = s_M]}_{\textbf{ex-ante MISTAKE}} < \underbrace{\mathbb{E}_Y[r(a,y)|s_M = a, \theta = hi, s_H \neq s_M]}_{\textbf{ex-ante MISTAKE}}, && \text{(EC.6b)} \\
s_H = a, \theta = lo: \quad & \underbrace{\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = lo, s_H \neq s_M]}_{\textbf{ex-ante MISTAKE}} < \underbrace{\mathbb{E}_Y[r(a,y)|s_M = a, \theta = lo, s_H = s_M]}_{\textit{ex-ante correct}}, && \text{(EC.6c)} \\
s_H \neq a, \theta = lo: \quad & \underbrace{\mathbb{E}_Y[r(a,y)|s_M \neq a, \theta = lo, s_H = s_M]}_{\textbf{ex-ante MISTAKE}} < \underbrace{\mathbb{E}_Y[r(a,y)|s_M = a, \theta = lo, s_H \neq s_M]}_{\textit{ex-ante correct}}. && \text{(EC.6d)}
\end{aligned}$$

Note that the expectation enables the manager to rank decisions not only by whether they are ex-ante good or bad, i.e., (EC.6c) and (EC.6d), but also by their relative expected impact on payoffs—for example, as in (EC.6b), even a bad decision such as over-relying on the machine signal ($a = s_M$) when $s_H \neq s_M$ and $\theta = hi$ yields a higher expected payoff than an alternative that is worse, such as going against both signals ($a \neq s_M$) when $s_H = s_M$ and $\theta = hi$, and the same for good decisions as in (EC.6a).

Combining these, the manager evaluates $\mathbb{E}_Y[r(a,y) \mid s_M]$ for all possible combinations of (s_M, a, y) as follows. These scenarios are defined by whether the DM's action aligns with the actual state, and whether it aligns with the machine signal.

Action aligns with the Machine signal and the true state:

$$\begin{aligned}
&\mathbb{E}_Y[r(a,y) \mid s_M = a, y = a] = \\
&\Pr(A = s_H, \theta = hi | s_M, a = s_M, y = a) \cdot \left(\mu^{hi}_{RR} \cdot \overline{r} + (1 - \mu^{hi}_{RR}) \cdot \underline{r}\right) \\
&+ \Pr(A \neq s_H, \theta = hi | s_M, a = s_M, y = a) \cdot \left(\mu^{hi}_{BR} \cdot \underline{r} + (1 - \mu^{hi}_{BR}) \cdot \overline{r}\right) \\
&+ \Pr(A = s_H, \theta = lo | s_M, a = s_M, y = a) \cdot \left(\mu^{lo}_{RR} \cdot \overline{r} + (1 - \mu^{lo}_{RR}) \cdot \underline{r}\right) \\
&+ \Pr(A \neq s_H, \theta = lo | s_M, a = s_M, y = a) \cdot \left(\mu^{lo}_{RB} \cdot \overline{r} + (1 - \mu^{lo}_{RB}) \cdot \underline{r}\right)
\end{aligned} \tag{EC.7}$$

Action does NOT align with the Machine signal but aligns the true state:

$$\begin{aligned}
&\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y = a] = \\
&\Pr(A = s_H, \theta = hi | s_M, a \neq s_M, y = a) \cdot \left(\mu^{hi}_{BR} \cdot \overline{r} + (1 - \mu^{hi}_{BR}) \cdot \underline{r}\right) \\
&+ \Pr(A \neq s_H, \theta = hi | s_M, a \neq s_M, y = a) \cdot \left(\mu^{hi}_{RR} \cdot \underline{r} + (1 - \mu^{hi}_{RR}) \cdot \overline{r}\right) \\
&+ \Pr(A = s_H, \theta = lo | s_M, a \neq s_M, y = a) \cdot \left(\mu^{lo}_{RB} \cdot \underline{r} + (1 - \mu^{lo}_{RB}) \cdot \overline{r}\right) \\
&+ \Pr(A \neq s_H, \theta = lo | s_M, a \neq s_M, y = a) \cdot \left(\mu^{lo}_{RR} \cdot \underline{r} + (1 - \mu^{lo}_{RR}) \cdot \overline{r}\right)
\end{aligned} \tag{EC.8}$$

Action aligns with the Machine signal but does NOT align with the true state:

$$\begin{aligned}
&\mathbb{E}_Y[r(a,y) \mid s_M = a, y \neq a] = \\
&\Pr(A = s_H, \theta = hi | s_M, a = s_M, y \neq a) \cdot \left(\mu^{hi}_{RR} \cdot \overline{r} + (1 - \mu^{hi}_{RR}) \cdot \underline{r}\right) \\
&+ \Pr(A \neq s_H, \theta = hi | s_M, a = s_M, y \neq a) \cdot \left(\mu^{hi}_{BR} \cdot \underline{r} + (1 - \mu^{hi}_{BR}) \cdot \overline{r}\right) \\
&+ \Pr(A = s_H, \theta = lo | s_M, a = s_M, y \neq a) \cdot \left(\mu^{lo}_{RR} \cdot \overline{r} + (1 - \mu^{lo}_{RR}) \cdot \underline{r}\right) \\
&+ \Pr(A \neq s_H, \theta = lo | s_M, a = s_M, y \neq a) \cdot \left(\mu^{lo}_{RB} \cdot \overline{r} + (1 - \mu^{lo}_{RB}) \cdot \underline{r}\right)
\end{aligned} \tag{EC.9}$$

Action does NOT align with the Machine signal and the true state:

$$\begin{aligned}
&\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y \neq a] = \\
&\Pr(A = s_H, \theta = hi | s_M, a \neq s_M, y \neq a) \cdot \left(\mu_{BR}^{hi} \cdot \overline{r} + (1 - \mu_{BR}^{hi}) \cdot \underline{r}\right) \\
+&\Pr(A \neq s_H, \theta = hi | s_M, a \neq s_M, y \neq a) \cdot \left(\mu_{RR}^{hi} \cdot \underline{r} + (1 - \mu_{RR}^{hi}) \cdot \overline{r}\right) \\
+&\Pr(A = s_H, \theta = lo | s_M, a \neq s_M, y \neq a) \cdot \left(\mu_{RB}^{lo} \cdot \underline{r} + (1 - \mu_{RB}^{lo}) \cdot \overline{r}\right) \\
+&\Pr(A \neq s_H, \theta = lo | s_M, a \neq s_M, y \neq a) \cdot \left(\mu_{RR}^{lo} \cdot \underline{r} + (1 - \mu_{RR}^{lo}) \cdot \overline{r}\right)
\end{aligned} \tag{EC.10}$$

Overall, the DM's choice to follow or override the machine signal thus has two opposing effects on the manager's assessment of the DM's decision, $\mathbb{E}_Y[r(a,y) \mid s_M]$. While overriding the machine signal increases the manager's belief that the DM correctly followed a high precision human signal (Eq. EC.4), following the machine signal increases the expected payoff $\mathbb{E}_Y[r(a,y) \mid \mathbf{s}]$. In what follows, we establish three properties of the manager's assessment of ex-ante decision quality and then show, in the subsequent proposition, that the latter effect dominates.

LEMMA EC.5 **(Effect of DM Noise on Manager's Assessment)**. *Suppose $\beta > \underline{\beta}$, the manager's assessment of the ex-ante decision quality, $\mathbb{E}_Y[r(a,y) \mid s_M]$, is strictly decreasing in β.*

This lemma shows that as the DM becomes more noisy or less rational (higher β), the manager systematically evaluates the DM's decision as worse in expectation. In other words, greater decision noise reduces perceived ex-ante decision quality.

LEMMA EC.6 **(Asymmetric Sensitivity to Noise)**. *The manager's assessment of the ex-ante decision quality, when the DM overrides the machine signal, $\mathbb{E}_Y[r(a,y) \mid s_M \neq a]$, decreases more rapidly with β than when the DM follows the machine signal, $\mathbb{E}_Y[r(a,y) \mid s_M = a]$.*

This lemma formalizes that the ex-ante assessment deteriorates faster when the DM overrides the machine signal than when the DM follows it. Thus, the manager penalizes noisy deviations from the machine more strongly than noisy adherence to it.

LEMMA EC.7 **(Effect of Machine Precision on Manager's Assessment)**. *Suppose $\beta > \underline{\beta}$, the manager's assessment of the ex-ante decision quality is strictly increasing in q_{s_M} when the DM follows the machine's signal, $\mathbb{E}_Y[r(a,y) \mid s_M = a]$, and strictly decreasing in q_{s_M} when the DM overrides the machine's signal, $\mathbb{E}_Y[r(a,y) \mid s_M \neq a]$.*

This lemma captures that the manager values the DM's decision more when the DM follows a more precise machine signal, but values it less when the DM overrides a more precise machine signal. In this sense, machine precision amplifies the evaluation gap between following and overriding.

The three lemmas establish how the manager evaluates the DM's decision as a function of (i) the DM's noise level, (ii) whether the DM follows or overrides the machine, and (iii) the precision of the machine signal. Together, they imply that—even when the DM occasionally receives a superior human signal—the manager's belief updating and expected-payoff assessment jointly favor decisions that follow the machine. The next proposition formalizes this result: at equilibrium, the manager always assigns higher ex-ante expected decision quality to actions aligned with the machine signal than to actions that override it.

PROPOSITION EC.1 **(Rational Assessment of Ex-ante Quality)**. *Suppose* $q^{hi}_{s_H} - q_{s_M} \le \overline{q}$*, at equilibrium, the ex-ante expected payoff from the DM's decision satisfies:* $\mathbb{E}_Y[r(a,y) \mid \underbrace{s_M = a}_{follow}] > \mathbb{E}_Y[r(a,y) \mid \underbrace{s_M \neq a}_{override}]$.

The condition $q^{hi}_{s_H} - q_{s_M} \le \overline{q}$ is merely a sufficient condition that rules out the corner case in which the machine signal is effectively uninformative.

A.1.7. DM's Decision Precision. We now turn to our central question: how does the delegation of machine-assisted decisions affect the quality of these decisions? To contrast delegated decisions ($i = d$) with non-delegated ones ($i = n$), the following lemma characterizes the DM's response to bonus payments when the manager is fairness-minded but does not exhibit omission–commission bias (i.e., $\omega(a, y, s_M) = 1$). The subsequent proposition then generalizes this result by incorporating the manager's omission–commission distortions.

LEMMA EC.8 **(Decision Precision Benchmark)**. *Suppose* $\beta > \underline{\beta}$ *and* $\omega(a, y, s_M) = 1$*. At equilibrium:*

- *When* $\theta = hi$ *(the human signal is more precise):*
  - $0.5 < \mathcal{P}_d(\theta = hi, s_M, s_H) < \mathcal{P}_n(\theta = hi, s_M, s_H) < 1$.
- *When* $\theta = lo$ *(the machine signal is more precise):*
  - $0.5 < \mathcal{P}_d(\theta = lo, s_M = s_H) < \mathcal{P}_n(\theta = lo, s_M = s_H) < 1$.
  - $0.5 < \mathcal{P}_n(\theta = lo, s_M \neq s_H) < \mathcal{P}_d(\theta = lo, s_M \neq s_H) < 1$.

Lemma EC.9 shows that, even when the manager is free of omission–commission concerns, delegation systematically increases the DM's reliance on the machine signal in states where signals conflict. When the human signal is more precise ($\theta = hi$, $\overline{a} = s_H$), delegation lowers the probability that the DM follows the human signal relative to the no delegation benchmark, thereby increasing algorithmic over reliance when signals conflict. When the machine signal is more precise ($\theta = lo$, $\overline{a} = s_M$), delegation increases reliance on the machine in conflicting signal states, mitigating algorithmic under reliance. By contrast, when signals align, delegation reduces decision quality relative to the no delegation benchmark, as it dampens the DM's responsiveness to the signals.

The next proposition generalizes this benchmark by allowing the manager to exhibit omission–commission bias (i.e., $\omega_g > 1$ and $\omega_b < 1$), thereby characterizing how such distortions further affect the DM's reliance on the machine.

PROPOSITION EC.2 **(Decision Precision under Omission–Commission Bias)**. *Suppose* $\beta > \underline{\beta}$ *and* $\omega(s_M, a = s_M, y) = 1$*,* $\omega(s_M, a \neq s_M, y = a) = \omega_g > 1$*,* $\omega(s_M, a \neq s_M, y \neq a) = \omega_b < 1$*. At equilibrium, for each informational state* $\mathbf{s} = (s_M, s_H, \theta)$ *there exist constants* $\xi_{\mathbf{s}} > 0$*,* $\zeta_{\mathbf{s}} > 0$*, and* $\varphi_{\mathbf{s}} \in \mathbb{R}$ *such that the following decision-precision ordering holds.*

(i) Suppose that $\theta = hi$ *and* $s_H = s_M$*.*

- *If* $\xi_{(s_M, s_H = s_M, \theta = hi)}(\omega_g - 1) - \zeta_{(s_M, s_H = s_M, \theta = hi)}(1 - \omega_b) \ge \varphi_{(s_M, s_H = s_M, \theta = hi)}$*, then*

$$0.5 < \mathcal{P}_d(A = s_H \mid \theta = hi, s_M = s_H) \le \mathcal{P}_n(A = s_H \mid \theta = hi, s_M = s_H) < 1.$$

- *Otherwise,*

$$0.5 < \mathcal{P}_n(A = s_H \mid \theta = hi, s_M = s_H) < \mathcal{P}_d(A = s_H \mid \theta = hi, s_M = s_H) < 1.$$

(ii) Suppose that $\theta = hi$ *and* $s_H \neq s_M$*.*

- *If* $\xi_{(s_M, s_H \neq s_M, \theta = hi)}(\omega_g - 1) - \zeta_{(s_M, s_H \neq s_M, \theta = hi)}(1 - \omega_b) \geq \varphi_{(s_M, s_H \neq s_M, \theta = hi)}$*, then*

$$0.5 < \mathcal{P}_n(A = s_H \mid \theta = hi, s_M \neq s_H) \leq \mathcal{P}_d(A = s_H \mid \theta = hi, s_M \neq s_H) < 1.$$

- *Otherwise,*

$$0.5 < \mathcal{P}_d(A = s_H \mid \theta = hi, s_M \neq s_H) < \mathcal{P}_n(A = s_H \mid \theta = hi, s_M \neq s_H) < 1.$$

(iii) Suppose that $\theta = lo$ *and* $s_H = s_M$*.*

- *If* $\xi_{(s_M, s_H = s_M, \theta = lo)}(\omega_g - 1) - \zeta_{(s_M, s_H = s_M, \theta = lo)}(1 - \omega_b) \geq \varphi_{(s_M, s_H = s_M, \theta = lo)}$*, then*

$$0.5 < \mathcal{P}_d(A = s_M \mid \theta = lo, s_M = s_H) \leq \mathcal{P}_n(A = s_M \mid \theta = lo, s_M = s_H) < 1.$$

- *Otherwise,*

$$0.5 < \mathcal{P}_n(A = s_M \mid \theta = lo, s_M = s_H) < \mathcal{P}_d(A = s_M \mid \theta = lo, s_M = s_H) < 1.$$

(iv) Suppose that $\theta = lo$ *and* $s_H \neq s_M$*.*

- *If* $\xi_{(s_M, s_H \neq s_M, \theta = lo)}(\omega_g - 1) - \zeta_{(s_M, s_H \neq s_M, \theta = lo)}(1 - \omega_b) \geq \varphi_{(s_M, s_H \neq s_M, \theta = lo)}$*, then*

$$0.5 < \mathcal{P}_d(A = s_M \mid \theta = lo, s_M \neq s_H) \leq \mathcal{P}_n(A = s_M \mid \theta = lo, s_M \neq s_H) < 1.$$

- *Otherwise,*

$$0.5 < \mathcal{P}_n(A = s_M \mid \theta = lo, s_M \neq s_H) < \mathcal{P}_d(A = s_M \mid \theta = lo, s_M \neq s_H) < 1.$$

The proposition characterizes how delegation affects decision precision when managerial evaluation is distorted by omission–commission bias. For each informational state $\mathbf{s}$, the result hinges on a linear threshold condition. The constants $\xi_{\mathbf{s}}$ and $\zeta_{\mathbf{s}}$ summarize how strongly omission–commission bias tilts evaluation following overrides that lead to good versus bad outcomes, while $\varphi_{\mathbf{s}}$ captures the baseline wedge that determines the precision ordering in the absence of such bias.

The left-hand side of the inequality aggregates the net evaluative pressure induced by omission–commission bias—rewards for successful overrides scaled by $\xi_{\mathbf{s}}$ and penalties for unsuccessful overrides scaled by $\zeta_{\mathbf{s}}$. The right-hand side, $\varphi_{\mathbf{s}}$, represents the benchmark ordering implied by payoff fundamentals alone. When the bias-induced term dominates this baseline wedge, omission–commission bias overturns the baseline effect of delegation on decision precision; when it does not, it reinforces the baseline ordering.

Note that by Lemma EC.9, $\varphi_{\mathbf{s}} < 0$ when the human and machine signals align, i.e., $\mathbf{s} = (s_M, s_H = s_M, \theta)$, while $\varphi_{\mathbf{s}} > 0$ when the signals conflict, i.e., $\mathbf{s} = (s_M, s_H \neq s_M, \theta)$. Define

$$\hat{\omega}_{hi, s_M \neq s_H} := \frac{\zeta_{(s_M, s_H \neq s_M, \theta = hi)}}{\xi_{(s_M, s_H \neq s_M, \theta = hi)}}, \qquad \hat{\omega}_{lo, s_M \neq s_H} := \frac{\zeta_{(s_M, s_H \neq s_M, \theta = lo)}}{\xi_{(s_M, s_H \neq s_M, \theta = lo)}}.$$

In line with the empirical and experimental literature, we focus on environments in which machine overrides that lead to bad outcomes are punished much more severely than overrides that lead to good outcomes are rewarded, that is,

$$\frac{\omega_g - 1}{1 - \omega_b} < \overline{\omega} \leq \min\{\hat{\omega}_{hi, s_M \neq s_H}, \hat{\omega}_{lo, s_M \neq s_H}\}.$$

Under this (sufficient) condition, our baseline prediction is preserved: delegation systematically increases the decision maker's reliance on the machine when the machine and human signals conflict.

A.1.8. Observable Human Signal. We now consider the special case in which the human signal (s_H) and its precision ($q^{\theta}_{s_H}$) are publicly observed by both the DM and the manager. In this setting, the manager directly observes the information available to the DM at the time of decision-making, eliminating any need to infer the DM's decision process from realized decisions and outcomes.

Our goal is to isolate the role of informational asymmetry between the manager and the DM in the delegated setting. In particular, we compare delegated decisions in which the human signal is private information of the DM ($i = d$) with those in which the human signal is publicly observed by the manager ($i = o$). When s_H is public, deviations from the decision suggested by the more precise signal can be evaluated unambiguously, and the manager's bonus decision no longer reflects uncertainty about the DM's informational basis.

The next proposition characterizes equilibrium bonus payments when the human signal is observable ($i = o$).

PROPOSITION EC.3 **(Bonus Observable)**. *At equilibrium, bonus payments satisfy the following.*

A. When $\theta = hi$ (i.e., $\bar{a} = s_H$):

- *Suppose a bad outcome ($y \neq a$). Then $b_o(\boldsymbol{s}, a = s_H, y \neq a) > b_o(\boldsymbol{s}, a \neq s_H, y \neq a)$ if either $s_H = s_M$ or ($s_H \neq s_M$ and $\omega_b > \hat{\omega}_b$); the inequality reverses when $s_H \neq s_M$ and $\omega_b \leq \hat{\omega}_b$.*
- *Suppose a good outcome ($y = a$). Then $b_o(\boldsymbol{s}, a = s_H, y = a) > b_o(\boldsymbol{s}, a \neq s_H, y = a)$ if either $s_H \neq s_M$ or ($s_H = s_M$ and $\omega_g < \hat{\omega}^{hi}_g$); the inequality reverses when $s_H = s_M$ and $\omega_g \geq \hat{\omega}^{hi}_g$.*

B. When $\theta = lo$ (i.e., $\bar{a} = s_M$):

- *Suppose a bad outcome ($y \neq a$). Then $b_o(\boldsymbol{s}, a = s_M, y \neq a) > b_o(\boldsymbol{s}, a \neq s_M, y \neq a)$.*
- *Suppose a good outcome ($y = a$). Then $b_o(\boldsymbol{s}, a = s_M, y = a) > b_o(\boldsymbol{s}, a \neq s_M, y = a)$ if $\omega_g < \hat{\omega}^{lo}_g$, and the inequality reverses if $\omega_g \geq \hat{\omega}^{lo}_g$.*

In other words, when the manager's omission–commission considerations are not too extreme, that is, when $\omega_b > \hat{\omega}_b$ and $\omega_g < \min\{\hat{\omega}^{lo}_g, \hat{\omega}^{hi}_g\}$, the manager continues to rank ex-ante optimal decisions above ex-ante non-optimal ones. In this case, observability of the human signal implies that the manager rewards ex-ante optimal decisions more generously than ex-ante non-optimal ones.

The structure of bonus payments has immediate and intuitive implications for the DM's decision-making and, in particular, for the DM's propensity to override the machine signal. The following proposition characterizes equilibrium machine-reliance behavior across delegation with a private human signal ($i = d$), delegation with a public human signal ($i = o$), and no delegation ($i = n$).

PROPOSITION EC.4 **(Machine Reliance with Observable.)**. *Suppose* $\beta > \underline{\beta}$ *and* $(\omega_g - 1)/(1 - \omega_b) < \overline{\omega}$. *When signals conflict, i.e.,* $s_H \neq s_M$*:*

A. ***Overreliance****: When* $\theta = hi$ *(i.e.,* $\bar{a} = s_H$*):* $0.5 < \mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_o(\mathbf{s}'_{hi}) < \mathcal{P}_n(\mathbf{s}'_{hi}) < 1$.

B. ***Underreliance****: When* $\theta = lo$ *(i.e.,* $\bar{a} = s_M$*):* $0.5 < \mathcal{P}_n(\mathbf{s}'_{lo}) < \mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_d(\mathbf{s}'_{lo}) < 1$.

The proposition shows that eliminating information asymmetry about the human signal does not eliminate the effect of delegation on machine reliance completely: delegation still increases reliance on the machine signal due to the manager's omission–commission bias. However, the magnitude of this effect is attenuated relative to environments in which the human signal remains private information.

To generate sharp predictions for our experimental setting in Study 2, we specialize the model to an environment in which the human signal is always more precise than the machine signal, i.e., $\mu_\theta = 1$, which corresponds to the design of the experiment. We then compare the delegated setting with information asymmetry regarding the human's signal ($i = d$) to the delegated setting without such asymmetry ($i = o$). To rule out cases in which extreme omission–commission considerations overturn basic performance rankings, we restrict attention to moderate levels of omission–commission bias, such that $\omega_b > \hat{\omega}_b$ and $\omega_g < \min\{\hat{\omega}_g^{lo}, \hat{\omega}_g^{hi}\}$. Under this restriction, the manager continues to rank ex-ante optimal decisions above ex-ante non-optimal ones in the observable setting ($i = o$), allowing us to isolate how information asymmetry shapes bonus payments under delegation.

PROPOSITION EC.5 **(Observable vs Unobservable)**. *Assume* $\mu_\theta = 1$, $\omega_b > \hat{\omega}_b$ *and* $\omega_g < \min\{\hat{\omega}_g^{lo}, \hat{\omega}_g^{hi}\}$. *Let* $b_o(\boldsymbol{s}, a, y)$ *and* $b_d(s_M, a, y)$ *denote the equilibrium bonus when the human signal* s_H *is observable and unobservable, respectively. In equilibrium, the following hold for both* $y = a$ *and* $y \neq a$*:*

1. $b_o(\underbrace{\boldsymbol{s}, a = \underline{a}}_{\substack{ex\text{-}ante \\ bad}}, y) < b_d(s_M, a, y) < b_o(\underbrace{\boldsymbol{s}, a = \overline{a}}_{\substack{ex\text{-}ante \\ good}}, y)$.
2. $b_d(s_M, a, y) < \mathbb{E}_{s_H}[b_o(\boldsymbol{s}, a, y)]$.

The bonus in $i = d$ can be interpreted as a convex combination of the bonuses paid for ex-ante optimal and ex-ante non-optimal decisions in $i = o$, with weights given by the manager's equilibrium belief that the decision was ex-ante optimal. Therefore, the unobservable bonus must lie between these two observable benchmarks. Moreover, as shown in Proposition EC.4, decision precision is lower under $i = d$ than under $i = o$, and the manager anticipates this in equilibrium. Consequently, the expected bonus paid under $i = d$ is strictly lower than the expected bonus under $i = o$.

A.1.9. View Feature. In Studies 1 and 3, managers can choose whether to view the machine signal before assigning a bonus. We label this decision by $ViewM \in \{0, 1\}$, where $ViewM = 1$ indicates that the manager observes the machine signal and can infer whether the decision maker followed or overrode it. To derive a formal prediction comparing bonus payments conditional on this choice, we analyze a setting in which managers who do not view the machine signal are subject to the same procedural fairness considerations as those who do. Under this assumption, differences in bonus payments across $ViewM = 0$ and $ViewM = 1$ arise solely from the information available to the manager at the time of deciding the bonus.

PROPOSITION EC.6 **(View vs No View)**. *Let $b_{v=1}(s_M, a, y)$ and $b_{v=0}(a, y)$ denote the equilibrium bonus when the manager views and not views the machine signal, respectively. In equilibrium the following holds for both $y = a$ and $y \neq a$:*

$$\min\{b_{v=1}(s_M, a \neq s_M, y), b_{v=1}(s_M, a = s_M, y)\} < b_{v=0}(a, y) < \max\{b_{v=1}(s_M, a \neq s_M, y), b_{v=1}(s_M, a = s_M, y)\}.$$

When the manager does not view the machine signal, the bonus cannot be conditioned on whether the DM followed or overrode the machine and therefore pools across these two cases. As a result, the equilibrium bonus under $ViewM = 0$ must lie between the bonuses paid when the manager observes the machine signal and conditions on following versus overriding ($ViewM = 1$). This pooling logic implies that bonus payments without viewing reflect an average of the bonuses paid after following and overriding, whereas viewing the machine signal allows the manager to differentiate between these decisions in assessment.

A.2. Mathematical Proofs

Proof of Lemma EC.1: Let $F := \alpha \cdot r(a, y) + \gamma \cdot \tilde{r}(a, s_M, y)$ and $K := 2(\alpha + \gamma)$. We reformulate the manager's utility function as:

$$\mathcal{U}^{ma} = \begin{cases} r(a, y) - F + (K - 1)b, & b \leq F/K \\ r(a, y) + F - (K + 1)b, & b > F/K \end{cases} \tag{EC.11}$$

Suppose $b \leq F/K$, if $\alpha + \gamma \leq \frac{1}{2}$ then $K - 1 \leq 0$ and $b^* = 0$, and if if $\alpha + \gamma > \frac{1}{2}$ then $K - 1 > 0$ and $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a, y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M, a, y)\right)$. Now, suppose $b > F/K$, then $-(K + 1) < 0$ and $b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a, y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M, a, y)\right)$. This concludes the proof of part 1.

For part 2, if the manager were to choose the action herself, she would select the payoff–maximizing action $\overline{a}$, which maximizes $\mathbb{E}_Y[r(a, y) \mid \mathbf{s}]$. To establish this formally, we analyze the manager's preferences under the two possible bonus regimes. When $b^* = 0$, it is immediate that $\overline{a} \equiv \arg\max_a \mathbb{E}_Y[r(a, y) \mid \mathbf{s}]$ maximizes the manager's expected utility. When $b^* > 0$, the manager's utility is $\mathcal{U}^{ma*} = C_1\, r(a, y) - C_2\, \tilde{r}(s_M, a, y)$, where

$$C_1 = \frac{\alpha + 2\gamma}{2(\alpha + \gamma)} > 0, \qquad C_2 = \frac{\gamma}{2(\alpha + \gamma)} > 0, \qquad \frac{C_1}{C_2} = \frac{\alpha}{\gamma} + 2.$$

Define $\Delta_r := \mathbb{E}_Y[r(\overline{a}, y) - r(\underline{a}, y) \mid \mathbf{s}] = (2\mu_{\mathbf{s}} - 1)(\overline{r} - \underline{r})$, and $\Delta_{\tilde{r}} := \mathbb{E}_Y[\tilde{r}(\overline{a}, s_M, y) - \tilde{r}(\underline{a}, s_M, y) \mid \mathbf{s}] = (\omega_{\overline{a}}\mu_{\overline{a}} - \omega_{\underline{a}}\mu_{\underline{a}})(\overline{r} - \underline{r}) + (\omega_{\overline{a}} - \omega_{\underline{a}})\underline{r}$, where $\mu_{\overline{a}}$ and $\mu_{\underline{a}}$ denote the manager's expected beliefs regarding the posterior induced by decisions $\overline{a}$ and $\underline{a}$, and $\omega_{\overline{a}}$ and $\omega_{\underline{a}}$ are the corresponding omission–commission scale parameters. Let

$$\kappa_0 \equiv \frac{(\omega_{\overline{a}}\mu_{\overline{a}} - \omega_{\underline{a}}\mu_{\underline{a}})(\overline{r} - \underline{r}) + (\omega_{\overline{a}} - \omega_{\underline{a}})\underline{r}}{(2\mu_{\mathbf{s}} - 1)(\overline{r} - \underline{r})} - 2.$$

The decision $\overline{a}$ maximizes the manager's expected utility if and only if

$$\frac{\alpha}{\gamma} \geq \kappa_0. \tag{EC.12}$$

Note that when $\omega_{\overline{a}} = \omega_{\underline{a}} = 1$, the inequality always holds. Define $\overline{\kappa_0} := \sup_{\omega,\mu} \kappa_0$; we assume $\frac{\alpha}{\gamma} > \overline{\kappa} \geq \overline{\kappa_0}$ so the inequality EC.12 holds. This concludes the proof of part 2.

Proof of Lemma EC.3: Part 1 is directly resulted from lemmas EC.1 and EC.2. We know from lemma EC.2 that when $b > 0$ the DM chooses a decision such that $a^* = \arg\max_a \mathbb{E}_Y[\frac{\alpha}{\alpha+\gamma} r(a, y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M, a, y) | \mathbf{s}]$. Define

$\Delta_r := \mathbb{E}_Y[r(\overline{a},y) - r(\underline{a},y) \mid \mathbf{s}] = (2\mu_{\mathbf{s}} - 1)(\overline{r} - \underline{r})$, and $\Delta_{\tilde{r}} := \mathbb{E}_Y[\tilde{r}(\overline{a}, s_M, y) - \tilde{r}(\underline{a}, s_M, y) \mid \mathbf{s}] = (\omega_{\overline{a}}\mu_{\overline{a}} - \omega_{\underline{a}}\mu_{\underline{a}})(\overline{r} - \underline{r}) + (\omega_{\overline{a}} - \omega_{\underline{a}})\underline{r}$. It is now evident that, given $b > 0$, if $\frac{\alpha}{\gamma} \leq \kappa$ that $a^* = \underline{a}$, and if $\frac{\alpha}{\gamma} > \kappa$ that $a^* = \overline{a}$. This conlcudes the proof. ■

Proof of Proposition 1: From Lemma EC.3, recall $\kappa = \frac{(\omega_{\underline{a}}\mu_{\underline{a}} - \omega_{\overline{a}}\mu_{\overline{a}})(\overline{r} - \underline{r}) + (\omega_{\underline{a}} - \omega_{\overline{a}})\underline{r}}{(2\mu_{\mathbf{s}} - 1)(\overline{r} - \underline{r})}$. Now let $\overline{\kappa} := \sup_{\omega,\mu} \kappa$. The proposition follows directly from Lemma EC.3 together with the assumptions $\alpha + \gamma > \frac{1}{2}$ and $\frac{\alpha}{\gamma} > \overline{\kappa}$.

Proof of Lemma EC.4: We first compute the derivative of the optimal choice probability with respect to β and then identify sufficient conditions under which $\mathcal{P}$ decreases in β.

Let $\mathcal{P} \equiv \mathcal{P}(A = \overline{a} \mid \mathbf{s})$, and define

$$u_1(\mathcal{P}) = \mathbb{E}_Y[\mathcal{U}^{dm}(\overline{a}, y) \mid \mathbf{s}], \qquad u_0(\mathcal{P}) = \mathbb{E}_Y[\mathcal{U}^{dm}(\underline{a}, y) \mid \mathbf{s}].$$

By definition,

$$\mathcal{P} = \frac{u_1(\mathcal{P})^{1/\beta}}{u_1(\mathcal{P})^{1/\beta} + u_0(\mathcal{P})^{1/\beta}}.$$

Rearranging gives

$$\ln(1 - \mathcal{P}) - \ln(\mathcal{P}) = \frac{1}{\beta}\big(\ln u_0(\mathcal{P}) - \ln u_1(\mathcal{P})\big). \tag{EC.13}$$

Differentiating (EC.13) with respect to β yields

$$\frac{d\mathcal{P}}{d\beta} = \frac{(1/\beta^2)\ln\big(\frac{u_0(\mathcal{P})}{u_1(\mathcal{P})}\big)}{\frac{1}{\mathcal{P}(1-\mathcal{P})} + \frac{1}{\beta}\Big(\frac{u_0'(\mathcal{P})}{u_0(\mathcal{P})} - \frac{u_1'(\mathcal{P})}{u_1(\mathcal{P})}\Big)}. \tag{EC.14}$$

When incentives are aligned, we have $u_0(\mathcal{P}) < u_1(\mathcal{P})$, so the numerator in (EC.14) is negative. For equilibrium stability, the denominator must be positive, i.e.,

$$\frac{u_1'(\mathcal{P})}{u_1(\mathcal{P})} - \frac{u_0'(\mathcal{P})}{u_0(\mathcal{P})} < \frac{\beta}{\mathcal{P}(1-\mathcal{P})}. \tag{EC.15}$$

Recall that in equilibrium $\mathcal{U}^{dm} = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M, a, y)\right)$. Let

$$u_1(\mathcal{P}) = \tfrac{1}{2}\Big(\lambda\big(\hat{\mu}\,\overline{r} + (1-\hat{\mu})\,\underline{r}\big) + (1-\lambda) f(\mathcal{P})\Big), \quad u_0(\mathcal{P}) = \tfrac{1}{2}\Big(\lambda\big(\hat{\mu}\,\underline{r} + (1-\hat{\mu})\,\overline{r}\big) + (1-\lambda) g(\mathcal{P})\Big),$$

where $\hat{\mu} > 1/2$ is the posterior belief after observing the signal realizations, and $\lambda = \frac{\alpha}{\alpha+\gamma}$.

Define

$$H = \frac{\sup_{\mathcal{P}} |(1-\lambda) f'(\mathcal{P})|}{\inf_{\mathcal{P}} u_1(\mathcal{P})} + \frac{\sup_{\mathcal{P}} |(1-\lambda) g'(\mathcal{P})|}{\inf_{\mathcal{P}} u_0(\mathcal{P})} = \frac{\sup_{\mathcal{P}} |f'(\mathcal{P})|}{\hat{\mu}\,\overline{r} + (1-\hat{\mu})\,\underline{r}} + \frac{\sup_{\mathcal{P}} |g'(\mathcal{P})|}{\hat{\mu}\,\underline{r} + (1-\hat{\mu})\,\overline{r}}.$$

A sufficient condition ensuring that the denominator of (EC.14) is positive is

$$\beta > \underline{\beta}_0, \quad \text{where} \quad \underline{\beta}_0 = \frac{H}{4} \leq \underline{\beta}.$$

This bound is restrictive but not necessary; the regularity result continues to hold for a broad range of parameter values. Under these conditions, the derivative in (EC.14) is negative, implying that $\frac{d\mathcal{P}}{d\beta} < 0$, which completes the proof. ■

Proof of Corollary EC.2: Start from the log-odds identity (Eq. (EC.13)):

$$\Phi(\mathcal{P}, \beta) := \ln(1 - \mathcal{P}) - \ln \mathcal{P} - \frac{1}{\beta}\big(\ln u_0(\mathcal{P}) - \ln u_1(\mathcal{P})\big) = 0.$$

Differentiate implicitly with respect to x (treating β as fixed):

$$\frac{\partial \Phi}{\partial \mathcal{P}} \cdot \frac{d\mathcal{P}}{dx} + \frac{\partial \Phi}{\partial x} = 0,$$

hence

$$\frac{d\mathcal{P}}{dx} = -\frac{\partial_x \Phi}{\partial_{\mathcal{P}} \Phi} = \frac{\frac{1}{\beta}\big(\partial_x \ln u_0 - \partial_x \ln u_1\big)}{\frac{1}{\mathcal{P}(1-\mathcal{P})} + \frac{1}{\beta}\big(\frac{u_0'(\mathcal{P})}{u_0(\mathcal{P})} - \frac{u_1'(\mathcal{P})}{u_1(\mathcal{P})}\big)}.$$

Define

$$M_x := \sup_{\mathcal{P}} \Big|\partial_x \ln u_0(\mathcal{P}) - \partial_x \ln u_1(\mathcal{P})\Big|,$$

Therefore, taking absolute values,

$$\left|\frac{d\mathcal{P}}{dx}\right| \le \frac{(1/\beta)\, M_x}{\frac{1}{\mathcal{P}(1-\mathcal{P})} - \frac{1}{\beta}\left|\frac{u_0'}{u_0} - \frac{u_1'}{u_1}\right|}.$$

Recall that in equilibrium $\mathcal{U}^{dm} = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a,y) + \frac{\gamma}{\alpha+\gamma}\tilde{r}(s_M,a,y)\right)$. Let

$$u_1(\mathcal{P}) = \tfrac{1}{2}\Big(\lambda\big(\hat{\mu}\,\overline{r} + (1-\hat{\mu})\,\underline{r}\big) + (1-\lambda) f(\mathcal{P})\Big), \quad u_0(\mathcal{P}) = \tfrac{1}{2}\Big(\lambda\big(\hat{\mu}\,\underline{r} + (1-\hat{\mu})\,\overline{r}\big) + (1-\lambda) g(\mathcal{P})\Big),$$

where $\hat{\mu} > 1/2$ is the posterior belief after observing the signal realizations, and $\lambda = \frac{\alpha}{\alpha+\gamma}$. Define

$$H = \frac{\sup_{\mathcal{P}} |(1-\lambda) f'(\mathcal{P})|}{\inf_{\mathcal{P}} u_1(\mathcal{P})} + \frac{\sup_{\mathcal{P}} |(1-\lambda) g'(\mathcal{P})|}{\inf_{\mathcal{P}} u_0(\mathcal{P})} = \frac{\sup_{\mathcal{P}} |f'(\mathcal{P})|}{\hat{\mu}\,\overline{r} + (1-\hat{\mu})\,\underline{r}} + \frac{\sup_{\mathcal{P}} |g'(\mathcal{P})|}{\hat{\mu}\,\underline{r} + (1-\hat{\mu})\,\overline{r}}.$$

By the definition of H we have $\left|\frac{u_0'}{u_0} - \frac{u_1'}{u_1}\right| \le H$. Substituting into the previous display yields

$$\left|\frac{d\mathcal{P}}{dx}\right| \le \frac{(1/\beta) M_x}{(4\beta - H)/\beta} = \frac{M_x}{4\beta - H}.$$

Define $L(\beta) := M_x/(4\beta - H)$. Since $M_x \ge 0$ is constant (given primitives) and the denominator $4\beta - H$ is strictly increasing in β, $L(\beta)$ is strictly decreasing there. This proves the corollary. ■

Proof of Lemma EC.5: The expected payoff, $\mathbb{E}_Y[r(a,y) \mid s_M]$, under different scenarios from equations EC.7-EC.10 could be reformulated (expanded) as the following:

Let

$$\begin{aligned}
\mathbb{E}_Y[r(a,y) \mid s_M = a, y = a] &= \Big(P^{fl}_{a=y} \cdot \nu^{fl}_{a=y} + (1 - P^{fl}_{a=y}) \cdot \omega^{fl}_{a=y}\Big)(\overline{r} - \underline{r}) + \underline{r} \\
\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y = a] &= \Big(P^{ov}_{a=y} \cdot \nu^{ov}_{a=y} + (1 - P^{ov}_{a=y}) \cdot \omega^{ov}_{a=y}\Big)(\overline{r} - \underline{r}) + \underline{r} \\
\mathbb{E}_Y[r(a,y) \mid s_M = a, y \neq a] &= \Big(P^{fl}_{a\neq y} \cdot \nu^{fl}_{a\neq y} + (1 - P^{fl}_{a\neq y}) \cdot \omega^{fl}_{a\neq y}\Big)(\overline{r} - \underline{r}) + \underline{r} \\
\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y \neq a] &= \Big(P^{ov}_{a\neq y} \cdot \nu^{ov}_{a\neq y} + (1 - P^{ov}_{a\neq y}) \cdot \omega^{ov}_{a\neq y}\Big)(\overline{r} - \underline{r}) + \underline{r}
\end{aligned} \tag{EC.16}$$

where P^i_j is the manager's overall belief that the DM took the ex-ante correct decision, ν^i_j is the expected value of the ex-ante correct decision, while ω^i_j is the expected value of the ex-ante incorrect decision for $i \in \{fl, ov\}$ standing for *follow* the machine signal ($a = s_M$), and *override* the machine signal ($a \neq s_M$) and for $j \in \{a = y, a \neq y\}$ standing for good, and bad outcome.

In particular, the following mapping applies:

$$P^{fl}_{a=y} = p^{\text{hi}}_{11} + p^{\text{lo}}_{11} + q^{\text{lo}}_{11}, \quad P^{ov}_{a=y} = p^{\text{hi}}_{01}, \quad P^{fl}_{a\neq y} = p^{\text{hi}}_{10} + p^{\text{lo}}_{10} + q^{\text{lo}}_{10}, \quad P^{ov}_{a\neq y} = p^{\text{hi}}_{00} \tag{EC.17}$$

$$
\begin{aligned}
\nu^{fl}_{a=y} &= \frac{p^{\text{hi}}_{11}}{p^{\text{hi}}_{11}+p^{\text{lo}}_{11}+q^{\text{lo}}_{11}}\mu^{hi}_{RR} + \frac{p^{\text{lo}}_{11}}{p^{\text{hi}}_{11}+p^{\text{lo}}_{11}+q^{\text{lo}}_{11}}\mu^{lo}_{RR} + \frac{q^{\text{lo}}_{11}}{p^{\text{hi}}_{11}+p^{\text{lo}}_{11}+q^{\text{lo}}_{11}}\mu^{lo}_{RB}, \quad \omega^{fl}_{a=y} = 1-\mu^{hi}_{BR}\\
\nu^{ov}_{a=y} &= \mu^{hi}_{BR}, \quad \omega^{ov}_{a=y} = \frac{q^{\text{hi}}_{01}}{q^{\text{hi}}_{01}+p^{\text{lo}}_{01}+q^{\text{lo}}_{01}}(1-\mu^{hi}_{RR}) + \frac{p^{\text{lo}}_{01}}{q^{\text{hi}}_{01}+p^{\text{lo}}_{01}+q^{\text{lo}}_{01}}(1-\mu^{lo}_{RB}) + \frac{q^{\text{lo}}_{01}}{q^{\text{hi}}_{01}+p^{\text{lo}}_{01}+q^{\text{lo}}_{01}}(1-\mu^{lo}_{RR})\\
\nu^{fl}_{a\neq y} &= \frac{p^{\text{hi}}_{10}}{p^{\text{hi}}_{10}+p^{\text{lo}}_{10}+q^{\text{lo}}_{10}}\mu^{hi}_{RR} + \frac{p^{\text{lo}}_{10}}{p^{\text{hi}}_{10}+p^{\text{lo}}_{10}+q^{\text{lo}}_{10}}\mu^{lo}_{RR} + \frac{q^{\text{lo}}_{10}}{p^{\text{hi}}_{10}+p^{\text{lo}}_{10}+q^{\text{lo}}_{10}}\mu^{lo}_{RB}, \quad \omega^{fl}_{a\neq y} = 1-\mu^{hi}_{BR}\\
\nu^{ov}_{a\neq y} &= \mu^{hi}_{BR}, \quad \omega^{ov}_{a\neq y} = \frac{q^{\text{hi}}_{00}}{q^{\text{hi}}_{00}+p^{\text{lo}}_{00}+q^{\text{lo}}_{00}}(1-\mu^{hi}_{RR}) + \frac{p^{\text{lo}}_{00}}{q^{\text{hi}}_{00}+p^{\text{lo}}_{00}+q^{\text{lo}}_{00}}(1-\mu^{lo}_{RB}) + \frac{q^{\text{lo}}_{00}}{q^{\text{hi}}_{00}+p^{\text{lo}}_{00}+q^{\text{lo}}_{00}}(1-\mu^{lo}_{RR})
\end{aligned}
\tag{EC.18}
$$

Now we check the derivatives starting with the scenario where the DM follows the machine signal and outcome is good.

$$
\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_Y[r(a,y)\mid s_M=a, y=a] = \Big(\frac{\mathrm{d}}{\mathrm{d}\beta}P^{fl}_{a=y}\cdot(\nu^{fl}_{a=y}-\omega^{fl}_{a=y}) + P^{fl}_{a=y}\cdot\frac{\mathrm{d}}{\mathrm{d}\beta}\nu^{fl}_{a=y}\Big)(\overline{r}-\underline{r}) \tag{EC.19}
$$

Note that Lemma EC.4 implies $\frac{\mathrm{d}}{\mathrm{d}\beta}P^{fl}_{a=y}<0$; that is as the DM becomes more noisy in decision making, the manager's belief that the DM took the correct ex-ante decision weakens. Therefore, $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_Y[r(a,y)\mid s_M=a, y=a]<0$ requires that:

$$
\nu^{fl}_{a=y}-\omega^{fl}_{a=y} > \frac{P^{fl}_{a=y}}{-\frac{\mathrm{d}}{\mathrm{d}\beta}P^{fl}_{a=y}}\frac{\mathrm{d}}{\mathrm{d}\beta}\nu^{fl}_{a=y} \tag{EC.20}
$$

Conceptually, any change in the rationality parameter β redistributes probability mass both *between* ex-ante correct and incorrect decisions and *within* each of those groups (among the different correct or incorrect cases). The inequality above guarantees that the former, between-group effect dominates, thereby determining the sign of the derivative. It yields:

$$
\mu^{hi}_{BR} > 1-\Big(\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}\mu^{hi}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}\mu^{lo}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}\mu^{lo}_{RB}\Big). \tag{EC.21}
$$

Similarly, $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_Y[r(a,y)\mid s_M\neq a, y=a]<0$, $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_Y[r(a,y)\mid s_M=a, y\neq a]<0$, and $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_Y[r(a,y)\mid s_M\neq a, y\neq a]<0$ respectively require:

$$
\begin{aligned}
\mu^{hi}_{BR} &> 1-\Big(\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{01}}{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{01}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{01}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{01}}\mu^{hi}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{01}}{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{01}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{01}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{01}}\mu^{lo}_{RB} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{01}}{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{01}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{01}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{01}}\mu^{lo}_{RR}\Big),\\
\mu^{hi}_{BR} &> 1-\Big(\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{10}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{10}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{10}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{10}}\mu^{hi}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{10}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{10}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{10}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{10}}\mu^{lo}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{10}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{10}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{10}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{10}}\mu^{lo}_{RB}\Big),\\
\mu^{hi}_{BR} &> 1-\Big(\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{00}}{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{00}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{00}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{00}}\mu^{hi}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{00}}{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{00}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{00}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{00}}\mu^{lo}_{RB} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{00}}{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{hi}}_{00}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{00}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{00}}\mu^{lo}_{RR}\Big).
\end{aligned}
\tag{EC.22}
$$

If all derivative terms in the denominators on the right-hand sides have the same sign, then each fraction appearing in a parenthesis is a nonnegative weight and the three weights sum to one. Then each parenthesis equals the convex combination of $\mu^{\cdot}_{\cdot}$ terms. We have each $\mu^{\cdot}_{\cdot}>\frac{1}{2}$; hence the convex combination exceeds 1/2. Consequently, because the left-hand side of every displayed inequality is $\mu^{hi}_{BR}>\frac{1}{2}$, the displayed inequalities hold.

To handle the mixed-sign case we impose a mild regularity condition. First, for the inequality EC.21 to hold, it is sufficient that:

$$
\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}\mu^{hi}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}\mu^{lo}_{RR} + \frac{\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}{\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{hi}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}p^{\text{lo}}_{11}+\frac{\mathrm{d}}{\mathrm{d}\beta}q^{\text{lo}}_{11}}\mu^{lo}_{RB} \le \tfrac{1}{2} \tag{EC.23}
$$

This can be reformulated as:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{hi}}\cdot(\mu_{RR}^{hi}-\tfrac{1}{2})+\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{lo}}\cdot(\mu_{RR}^{lo}-\tfrac{1}{2})+\frac{\mathrm{d}}{\mathrm{d}\beta}q_{11}^{\mathrm{lo}}\cdot(\mu_{RB}^{lo}-\tfrac{1}{2})\leq 0 \tag{EC.24}$$

Note that $\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{hi}}+\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{lo}}+\frac{\mathrm{d}}{\mathrm{d}\beta}q_{11}^{\mathrm{lo}}<0$, i.e., as the DM's becomes noisier in decision making, the manager's overall belief that the DM correctly followed the machine signal weakens. Therefore, since $\mu_{RR}^{hi}>\mu_{RR}^{lo}>\mu_{RB}^{lo}$, for the above inequality to hold, it suffices to show that $\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{hi}}<0$.

Let $P_1=\mathcal{P}_{\mathrm{align}}^{hi}, P_2=\mathcal{P}_{\mathrm{conf}}^{hi}, P_3=\mathcal{P}_{\mathrm{align}}^{lo}, P_4=\mathcal{P}_{\mathrm{conf}}^{lo}, A=q_{s_H}^{hi}\mu_\theta, B=(1-q_{s_H}^{hi})\mu_\theta, C=q_{s_H}^{lo}(1-\mu_\theta), D=(1-q_{s_H}^{lo})(1-\mu_\theta)$, and $L=AP_1+B(1-P_2)+CP_3+DP_4$. We have:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{hi}}=\frac{AB(P_1'(1-P_2)+P_1P_2')+AC(P_1'P_3-P_1P_3')+AD(P_1'P_4-P_1P_4')}{L^2} \tag{EC.25}$$

Since $\frac{\mathrm{d}}{\mathrm{d}\beta}P_i<0$, we have $P_1'(1-P_2)+P_1P_2'<0$, so the first term in the numerator is always negative. The second and third terms, however, may be positive or negative. For sufficiently large β, we have $P_1\approx P_2\approx P_3$ and $P_1'\approx P_2'\approx P_3'$, and therefore $P_1'P_3-P_1P_3'\approx 0$ and $P_1'P_4-P_1P_4'\approx 0$. Hence, there exists $\underline{\beta}_0$ such that for all $\beta>\underline{\beta}_0$, we have $\frac{\mathrm{d}}{\mathrm{d}\beta}p_{11}^{\mathrm{hi}}<0$. We assume $\beta>\underline{\beta}\geq\underline{\beta}_0$, and therefore the inequality holds. The analysis for the other inequalities follows a similar path and they hold for any $\beta>\underline{\beta}$, hence the proof is concluded. ■

Proof of Lemma EC.6: Variation in the rationality parameter β directly affects the manager's assessment by altering her belief about whether the DM's decision corresponds to an ex-ante optimal choice or an ex-ante mistake (the *between-group effect*). It also indirectly changes the manager's belief about the relative likelihood of different cases within the set of ex-ante optimal (or suboptimal) decisions (the *within-group effect*). To see this, suppose the DM follows the machine signal. As β increases and the DM becomes noisier in decision-making, the manager's belief that the DM correctly followed the machine declines. At the same time, the probability mass within the three possible ex-ante optimal decision cases—$(\theta=hi, s_M, s_H=s_M)$, $(\theta=lo, s_M, s_H=s_M)$, and $(\theta=lo, s_M, s_H\neq s_M)$—also shifts.

We show below that the *within-group effect* following an *override* of the machine signal reduces the manager's perceived decision quality more sharply than in the case of a *follow*, and that the *direct (between-group) effect* is also stronger in the override case.

Let

$$P_1=\mathcal{P}_{\mathrm{align}}^{hi},\quad P_2=\mathcal{P}_{\mathrm{conf}}^{hi},\quad P_3=\mathcal{P}_{\mathrm{align}}^{lo},\quad P_4=\mathcal{P}_{\mathrm{conf}}^{lo},$$

$$A=q_{s_H}^{hi}\mu_\theta,\quad B=(1-q_{s_H}^{hi})\mu_\theta,\quad C=q_{s_H}^{lo}(1-\mu_\theta),\quad D=(1-q_{s_H}^{lo})(1-\mu_\theta).$$

$$D_{fl}=AP_1+CP_3+DP_4,\quad D_{ov}=B(1-P_1)+D(1-P_3)+C(1-P_4)$$

Then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\beta}\frac{p_{11}^{\mathrm{hi}}}{p_{11}^{\mathrm{hi}}+p_{11}^{\mathrm{lo}}+q_{11}^{\mathrm{lo}}}&=\frac{AC(P_1'P_3-P_1P_3')+AD(P_1'P_4-P_1P_4')}{D_{fl}^2},\\ \frac{\mathrm{d}}{\mathrm{d}\beta}\frac{q_{01}^{\mathrm{hi}}}{q_{01}^{\mathrm{hi}}+p_{01}^{\mathrm{lo}}+q_{01}^{\mathrm{lo}}}&=\frac{BD[-P_1'(1-P_3)+(1-P_1)P_3']+BC[-P_1'(1-P_4)+(1-P_1)P_4']}{D_{ov}^2}.\end{aligned} \tag{EC.26}$$

Lemma EC.4 implies,

$$\frac{1-P_1}{1-P_3}<\frac{|P_1'|}{|P_3'|}\quad\text{and}\quad\frac{1-P_1}{1-P_4}<\frac{|P_1'|}{|P_4'|}.$$

Hence, the second derivative in the override case is positive. Intuitively, as β increases and the DM becomes noisier, the *within-group effect* for an *override* amplifies the direct effect: not only does the manager's belief that the DM made an ex-ante mistake rise, but within the set of ex-ante mistakes, probability mass shifts toward the most severe case—where signals are aligned and the DM has gone against both. Consequently, the manager's valuation of the DM's decision declines even more sharply.

For a *follow*, two cases arise:

1. If $\frac{|P_1'|}{|P_3'|} < \frac{P_1}{P_3}$ ($\frac{|P_1'|}{|P_4'|} < \frac{P_1}{P_4}$), the within-group effect mitigates the direct effect—while the manager's belief that the DM made an ex-ante optimal decision decreases, probability mass among optimal cases shifts toward the best one.
2. If $\frac{|P_1'|}{|P_3'|} \geq \frac{P_1}{P_3}$ ($\frac{|P_1'|}{|P_4'|} \geq \frac{P_1}{P_4}$), the within-group effect reinforces the direct effect, as in an override. In this case, since $D_{ov} < D_{fl}$, we obtain

$$\left| \frac{\mathrm{d}}{\mathrm{d}\beta} \frac{p_{11}^{\mathrm{hi}}}{p_{11}^{\mathrm{hi}} + p_{11}^{\mathrm{lo}} + q_{11}^{\mathrm{lo}}} \right| < \frac{\mathrm{d}}{\mathrm{d}\beta} \frac{q_{01}^{\mathrm{hi}}}{q_{01}^{\mathrm{hi}} + p_{01}^{\mathrm{lo}} + q_{01}^{\mathrm{lo}}},$$

showing that the *within-group effect* after an *override* is stronger than that after a *follow*.

Therefore, to show that $\mathbb{E}_Y[r(a,y) \mid s_M \neq a]$ decreases more rapidly in β than $\mathbb{E}_Y[r(a,y) \mid s_M = a]$ for both $y = a$ and $y \neq a$, it suffices to establish that $|\frac{\mathrm{d}}{\mathrm{d}\beta} p_{01}^{\mathrm{hi}}| > |\frac{\mathrm{d}}{\mathrm{d}\beta} q_{11}^{\mathrm{hi}}|$ and $|\frac{\mathrm{d}}{\mathrm{d}\beta} p_{00}^{\mathrm{hi}}| > |\frac{\mathrm{d}}{\mathrm{d}\beta} q_{10}^{\mathrm{hi}}|$, respectively. In other words, as the DM becomes noisier in decision-making, the manager's belief that overriding was ex-ante the correct decision should decline faster than the increase in the belief that following was ex-ante a mistake. We start with the case of $y = a$. First, we simplify the notation. Let $p = p_{01}^{\mathrm{hi}}$ and $q = q_{11}^{\mathrm{hi}}$. All constants A, B, C, D are positive. By Lemma EC.4, p is decreasing in β ($p' < 0$) and q is increasing ($q' > 0$). The inequality to be proven is $|p'| > |q'|$, which is equivalent to $-p' > q'$. The functions can be written as $p = N_p/D_p$ and $q = N_q/D_q$. Their derivatives are $p' = (N_p' D_p - N_p D_p')/D_p^2$ and $q' = (N_q' D_q - N_q D_q')/D_q^2$. Let $N_{p'}$ and $N_{q'}$ denote the numerators of these derivatives. The inequality is equivalent to:

$$\frac{-N_{p'}}{D_p^2} > \frac{N_{q'}}{D_q^2}$$

Note that $0 < D_p < D_q$. Thus, it would suffice to show that $-N_{p'} \geq N_{q'}$ for the inequality to hold. We can evaluate them as follows:

$$\begin{aligned}
-N_{p'} =& AB(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2 \cdot (1-P_1) - P_2 \cdot \frac{\mathrm{d}}{\mathrm{d}\beta}P_1) + AD(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2 \cdot (1-P_3) - P_2 \cdot \frac{\mathrm{d}}{\mathrm{d}\beta}P_3) \\
&+ AC(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2 \cdot (1-P_4) - P_2 \cdot \frac{\mathrm{d}}{\mathrm{d}\beta}P_4) \\
N_{q'} =& BA(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2 \cdot P_1 - (1-P_2) \cdot \frac{\mathrm{d}}{\mathrm{d}\beta}P_1) + BC(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2 \cdot P_3 - (1-P_2) \cdot \frac{\mathrm{d}}{\mathrm{d}\beta}P_3) \\
&+ BD(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2 \cdot P_4 - (1-P_2) \cdot \frac{\mathrm{d}}{\mathrm{d}\beta}P_4)
\end{aligned} \tag{EC.27}$$

Since $P_i' < 0$ both $-N_{p'}$ and $N_{q'}$ are positive. We want to show $\Delta := -N_{p'} - N_{q'} \geq 0$,

$$\begin{aligned}\Delta = &(-\frac{\mathrm{d}}{\mathrm{d}\beta}P_2)\Big[BAP_1 + BCP_3 + BDP_4 - \big(AB(1-P_1) + AD(1-P_3) + AC(1-P_4)\big)\Big] \\ &+ (-\frac{\mathrm{d}}{\mathrm{d}\beta}P_1)\big[\,P_2AB - (1-P_2)BA\,\big] \\ &+ (-\frac{\mathrm{d}}{\mathrm{d}\beta}P_3)\big[\,P_2AD - (1-P_2)BC\,\big] \\ &+ (-\frac{\mathrm{d}}{\mathrm{d}\beta}P_4)\big[\,P_2AC - (1-P_2)BD\,\big].\end{aligned} \tag{EC.28}$$

Since each $-\frac{\mathrm{d}}{\mathrm{d}\beta}P_i > 0$ and decreases in β ($\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}P_i \geq 0$), $\Delta(\beta)$ is a decreasing function of β. Moreover, as $\beta \to \infty$, each $\frac{\mathrm{d}}{\mathrm{d}\beta}P_i \to 0$ and hence $\lim_{\beta\to\infty}\Delta(\beta) = 0$. Because $\Delta(\beta)$ is nonnegative and weakly decreasing, it follows that $\Delta(\beta) \geq 0$ for all $\beta > 0$. This concludes the proof. ■

Proof of Lemma EC.7: We want to show that $\mathbb{E}_Y[r(a,y) \mid s_M = a, y = a]$ and $\mathbb{E}_Y[r(a,y) \mid s_M = a, y \neq a]$ are strictly increasing and $\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y = a]$ and $\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y \neq a]$ are strictly decreasing in q_{s_M}. Let:

$$\begin{aligned} p_1^{fl,g} = p_{11}^{hi}, \quad p_2^{fl,g} = q_{11}^{hi}, \quad p_3^{fl,g} = p_{11}^{lo}, \quad p_4^{fl,g} = q_{11}^{lo}, \\ p_1^{fl,b} = p_{10}^{hi}, \quad p_2^{fl,b} = q_{10}^{hi}, \quad p_3^{fl,b} = p_{10}^{lo}, \quad p_4^{fl,b} = q_{10}^{lo}, \\ p_1^{ov,g} = p_{01}^{hi}, \quad p_2^{ov,g} = q_{01}^{hi}, \quad p_3^{ov,g} = p_{01}^{lo}, \quad p_4^{ov,g} = q_{01}^{lo}, \\ p_1^{ov,b} = p_{00}^{hi}, \quad p_2^{ov,b} = q_{00}^{hi}, \quad p_3^{ov,b} = p_{00}^{lo}, \quad p_4^{ov,b} = q_{00}^{lo}, \end{aligned} \tag{EC.29}$$

and

$$\begin{aligned} v_1^{fl} = \mu_{R,R}^{hi}, \quad v_2^{fl} = 1 - \mu_{B,R}^{hi}, \quad v_3^{fl} = \mu_{R,R}^{lo}, \quad v_4^{fl} = \mu_{R,B}^{lo}, \\ v_1^{ov} = \mu_{B,R}^{hi}, \quad v_2^{ov} = 1 - \mu_{R,R}^{hi}, \quad v_3^{ov} = 1 - \mu_{R,B}^{lo}, \quad v_4^{ov} = 1 - \mu_{R,R}^{lo}. \end{aligned} \tag{EC.30}$$

Using the above definitions, we can reformulate the manager's adjusted payoff as:

$$\begin{aligned} &\mathbb{E}_Y[r(a,y) \mid s_M = a, y = a] = \sum_{i=1}^{4} p_i^{fl,g} \cdot v_i^{fl} \cdot (\overline{r} - \underline{r}) + \underline{r}, \quad \mathbb{E}_Y[r(a,y) \mid s_M = a, y \neq a] = \sum_{i=1}^{4} p_i^{fl,b} \cdot v_i^{fl} \cdot (\overline{r} - \underline{r}) + \underline{r}, \\ &\mathbb{E}_Y[r(a,y) \mid s_M \neq a, y = a] = \sum_{i=1}^{4} p_i^{ov,g} \cdot v_i^{ov} \cdot (\overline{r} - \underline{r}) + \underline{r}, \quad \mathbb{E}_Y[r(a,y) \mid s_M \neq a, y \neq a] = \sum_{i=1}^{4} p_i^{ov,b} \cdot v_i^{ov} \cdot (\overline{r} - \underline{r}) + \underline{r}. \end{aligned} \tag{EC.31}$$

Applying the first order derivative for follow-good yields:

$$\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}\mathbb{E}_Y[r(a,y) \mid s_M = a, y = a] = \Big(\sum_{i=1}^{4} p_i^{fl,g} \cdot \frac{\mathrm{d}}{\mathrm{d}q_{s_M}} v_i^{fl} + \sum_{i=1}^{4} \frac{\mathrm{d}}{\mathrm{d}q_{s_M}} p_i^{fl,g} \cdot v_i^{fl}\Big) \cdot (\overline{r} - \underline{r}), \tag{EC.32}$$

where $\sum_{i=1}^{4} p_i^{fl,g} \cdot \frac{\mathrm{d}}{\mathrm{d}q_{s_M}} v_i^{fl}$ captures the direct effect of an increase in q_{s_M} on the expected payoff, holding the belief weights fixed. The second term, $\sum_{i=1}^{4} \frac{\mathrm{d}}{\mathrm{d}q_{s_M}} p_i^{fl,g} \cdot v_i^{fl}$ reflects the indirect effect through the change in the manager's belief. Note that $\frac{\mathrm{d}}{\mathrm{d}q_{s_M}} v_i^{fl} > 0$ for all i, as the payoffs in this case are from either correctly following the machine or wrongly overriding it. Therefore, we get for the first part $\sum_{i=1}^{4} p_i^{fl,g} \cdot \frac{\mathrm{d}}{\mathrm{d}q_{s_M}} v_i^{fl} > 0$. The second part, however, can be both positive or negative. Hence, it is not straightforward to determine the sign of $\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}\mathbb{E}_Y[r(a,y) \mid s_M = a, y = a]$. Nonetheless, note that $\sum_{i=1}^{4} p_i^{fl,g} = 1$ so we have $\sum_{i=1}^{4} \frac{\mathrm{d}}{\mathrm{d}q_{s_M}} p_i^{fl,g} = 0$. As such, in most cases the indirect effect of an increase in q_{s_M} on the adjusted payoff through the change

in the manager's belief is either negligible or points to the same direction as the first part. However, in boundary situations there could be cases where the second part becomes arbitrarily too negative. As such, $\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}\mathbb{E}_Y[r(a,y)\mid s_M=a,y=a]>0$ requires:

$$-\sum_{i=1}^{4} p_i^{fl,g}\cdot\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}v_i^{fl} < \sum_{i=1}^{4}\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}p_i^{fl,g}\cdot v_i^{fl} \tag{EC.33}$$

We know from Corollary EC.2 that the derivative of $\mathcal{P}$ with respect to q_{s_M} is bounded such that $|\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}\mathcal{P}|\le L_{q_{s_M}}(\beta)$. Where $L_{q_{s_M}}(\beta)\ge 0$ is a decreasing function of the rationality parameter β. Using this bound we can calculate the bounds of $\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}p_i^{fl,g}$ for $i=1,2,3,4$ as:

$$|\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}p_i^{fl,g}|\le\Psi_i^{fl,g}(\beta)\equiv\frac{c_i\cdot L_{q_{s_M}(\beta)}}{D}\big(1+\frac{\eta_i\cdot N}{D}\big) \tag{EC.34}$$

where

$$\begin{aligned}
&c_1=q_{s_H}^{hi}\mu_\theta,\quad c_2=(1-q_{s_H}^{hi})\mu_\theta,\quad c_3=q_{s_H}^{lo}(1-\mu_\theta),\quad c_4=(1-q_{s_H}^{lo})(1-\mu_\theta)\\
&\eta_1=\mathcal{P}_{\text{align}}^{hi},\quad \eta_2=(1-\mathcal{P}_{\text{conf}}^{hi}),\quad \eta_3=\mathcal{P}_{\text{align}}^{lo},\quad \eta_4=\mathcal{P}_{\text{conf}}^{lo}\\
&N=\sum_{i=1}^{4}c_i,\quad D=\sum_{i=1}^{4}c_i\cdot\eta_i
\end{aligned} \tag{EC.35}$$

Applying these bounds yields:

$$-\Psi^{fl,g}(\beta)\cdot(\mu_{RR}^{hi}+\mu_{BR}^{hi}-1)\le\sum_{i=1}^{4}\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}p_i^{fl,g}\cdot v_i^{fl} \tag{EC.36}$$

where $\Psi^{fl,g}(\beta)=\max_i\Psi_i^{fl,g}(\beta)$. Note that $\lim_{\beta\to\infty}L_{q_{s_M}}(\beta)=\lim_{\beta\to\infty}\Psi^{fl,g}(\beta)=0$, therefore there exists a $\underline{\beta}_{\Psi^{fl,g}}\ge 0$ such that for $\beta>\underline{\beta}_{\Psi^{fl,g}}$:

$$-\sum_{i=1}^{4}p_i^{fl,g}\cdot\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}v_i^{fl}<-\Psi^{fl,g}(\beta)\cdot(\mu_{RR}^{hi}+\mu_{BR}^{hi}-1)\le\sum_{i=1}^{4}\frac{\mathrm{d}}{\mathrm{d}q_{s_M}}p_i^{fl,g}\cdot v_i^{fl} \tag{EC.37}$$

Similarly, we can calculate $\underline{\beta}_{\Psi^{fl,b}}$, $\underline{\beta}_{\Psi^{ov,g}}$, and $\underline{\beta}_{\Psi^{ov,b}}$. We assume $\beta>\underline{\beta}\ge\max\{\underline{\beta}_{\Psi^{fl,g}},\underline{\beta}_{\Psi^{fl,b}},\underline{\beta}_{\Psi^{ov,g}},\underline{\beta}_{\Psi^{ov,b}}\}$ so the proof is concluded. ■

Proof of Proposition EC.1: Let $\Delta\tilde{r}_{good}\equiv\mathbb{E}_Y[r(a,y)\mid s_M=a,y=a]-\mathbb{E}_Y[r(a,y)\mid s_M\ne a,y=a]$ and $\Delta\tilde{r}_{bad}\equiv\mathbb{E}_Y[r(a,y)\mid s_M=a,y\ne a]-\mathbb{E}_Y[r(a,y)\mid s_M\ne a,y\ne a]$. We want to show $\Delta\hat{r}_k>0$ for $k\in\{good,bad\}$. Lemma EC.7 implies that the differences $\Delta\tilde{r}_{good}$ and $\Delta\tilde{r}_{bad}$ increase strictly with q_{s_M}. We will first show how $\Delta\tilde{r}_{good}$ and $\Delta\tilde{r}_{bad}$ behave as q_{s_M} approaches its bounds, then we will derive the range on q_{s_M} where the proposition holds.

Note that $q_{s_M}\in(q_{s_H}^{lo},q_{s_H}^{hi})$, and the expected payoff of DM's decision when they correctly override the machine signal is $v_1^{ov}=\mu_{BR}^{hi}$, while the expected payoff when they wrongly follow the machine signal is $v_2^{fl}=1-\mu_{BR}^{hi}$; where $\mu_{BR}^{hi}=1-\mu_{RB}^{hi}$ is the DM's posterior belief after observing two conflicting signals that the state matches the human signal.

$$\lim_{q_{s_M}\to q_{s_H}^{hi}}\mu_{BR}^{hi}=1/2 \tag{EC.38}$$

This implies then that $\lim_{q_{s_M}\to q_{s_H}^{hi}}v_1^{ov}=\lim_{q_{s_M}\to q_{s_H}^{hi}}v_2^{fl}$. Consequently, we get:

$$\lim_{q_{s_M}\to q_{s_H}^{hi}}\sum_{i=1}^{4}p_i^{fl,k}\cdot v_i^{fl}>\lim_{q_{s_M}\to q_{s_H}^{hi}}\sum_{i=1}^{4}p_i^{ov,k}\cdot v_i^{ov},\quad\text{for } k\in\{g,b\} \tag{EC.39}$$

that is $\lim_{q_{s_M}\to q^{hi}_{s_H}} \Delta\tilde{r}_k > 0$ for $k \in \{good, bad\}$. In other words, when the high precision human signal does not provide any additional precision in decision making over the machine signal, the manager always assigns a higher ex-ante value to following the machine signal than overriding it, all else kept constant. This implies that there exists a threshold $\underline{q}_h \geq q^{lo}_{s_H}$ that for $q_{s_M} > \underline{q}_h$ we have $\Delta\tilde{r}_k > 0$ for $k \in \{good, bad\}$. Now let $q_{s_M} \to q^{lo}_{s_H}$ and evaluate $\Delta\tilde{r}_k$ for $k \in \{good, bad\}$ at $q_{s_M} = q^{lo}_{s_H}$.

$$\lim_{q_{s_M}\to q^{lo}_{s_H}} \Delta\tilde{r}_k \equiv \Delta\tilde{r}_k(q_{s_M} = q^{lo}_{s_H}, q^{lo}_{s_H}) \tag{EC.40}$$

Here, $\Delta\tilde{r}_k(q_{s_M} = q^{lo}_{s_H}, q^{lo}_{s_H})$ is an increasing function in $q^{lo}_{s_H}$ and its sign depends on $q^{hi}_{s_H} - q^{lo}_{s_H}$ as shown below:

$$\begin{aligned} &\lim_{q^{lo}_{s_H}\to q^{hi}_{s_H}} \mu^{hi}_{BR} = 1/2 \\ &\lim_{q^{lo}_{s_H}\to 1/2} \mu^{lo}_{BR} = \lim_{q^{lo}_{s_H}\to 1/2} \mu^{lo}_{RR} = 1/2 \end{aligned} \tag{EC.41}$$

This implies for $k \in \{good, bad\}$:

$$\begin{aligned} &\lim_{q^{lo}_{s_H}\to q^{hi}_{s_H}} \Delta\hat{r}_k(q_{s_M} = q^{lo}_{s_H}, q^{lo}_{s_H}) > 0 \\ &\lim_{q^{lo}_{s_H}\to 1/2} \Delta\hat{r}_k(q_{s_M} = q^{lo}_{s_H}, q^{lo}_{s_H}) < 0 \end{aligned} \tag{EC.42}$$

Therefore, there exists a threshold $\underline{q}_l > 1/2$ such that, for $q^{lo}_{s_H} > \underline{q}_l$, we have

$$\Delta\hat{r}_k(q_{s_M} = q^{lo}_{s_H}, q^{lo}_{s_H}) > 0 \quad \text{for } k \in \{good, bad\}.$$

In this case, $q^{lo}_{s_H} > \underline{q}_l$ implies $\underline{q}_h = q^{lo}_{s_H}$, and thus $\Delta\hat{r}_k > 0$ for $k \in \{good, bad\}$ for any q_{s_M}. By contrast, when $q^{lo}_{s_H} \leq \underline{q}_l$, we obtain $\underline{q}_h > q^{lo}_{s_H}$, and $\Delta\hat{r}_k > 0$ for $k \in \{good, bad\}$ only when $q_{s_M} > \underline{q}_h$. Overall, we can equivalently state that there exists a threshold $\underline{q}$ such that, whenever

$$q^{hi}_{s_H} - q_{s_M} \leq \overline{q},$$

we have $\Delta\hat{r}_k > 0$ for $k \in \{good, bad\}$. This concludes the proof. ■

Proof of Theorem 1: The bonus payment is determined directly by the manager's assessment of both the ex-ante and ex-post decision quality:

$$b^* = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma}\, r(a, y) \,+\, \frac{\gamma}{\alpha+\gamma}\, \tilde{r}(s_M, a, y)\right).$$

Therefore, to establish the inequalities stated in the proposition, it suffices to verify the following:

(A) $\frac{\alpha}{\alpha+\gamma}\, r(a, y = a) \,+\, \frac{\gamma}{\alpha+\gamma}\, \tilde{r}(s_M, a, y = a) \,>\, \frac{\alpha}{\alpha+\gamma}\, r(a, y \neq a) \,+\, \frac{\gamma}{\alpha+\gamma}\, \tilde{r}(s_M, a, y \neq a)$.

(B) $\tilde{r}(s_M, a = s_M, y \neq a) \,>\, \tilde{r}(s_M, a \neq s_M, y \neq a)$.

(C) Verify the inequalities both when

$\tilde{r}(s_M, a = s_M, y = a) \,>\, \tilde{r}(s_M, a \neq s_M, y = a)$ and when $\tilde{r}(s_M, a = s_M, y = a) \,\leq\, \tilde{r}(s_M, a \neq s_M, y = a)$.

Part (A). The inequality in part (A) holds whenever

$$\frac{\alpha}{\gamma} \,>\, \frac{\tilde{r}(s_M, a, y \neq a) - \tilde{r}(s_M, a, y = a)}{\overline{r} - \underline{r}}.$$

Note that

$$\tilde{r}(s_M, a, y \neq a) - \tilde{r}(s_M, a, y = a) = (\omega_{y\neq a}\mu_{y\neq a} - \omega_{y=a}\mu_{y=a})(\overline{r} - \underline{r}) + (\omega_{y\neq a} - \omega_{y=a})\,\underline{r},$$

where $\mu_{y=a}$ and $\mu_{y\neq a}$ denote the posterior beliefs about the state that the manager attributes to the DM after observing a good and bad outcome, respectively, and $\omega_{y=a}$ and $\omega_{y\neq a}$ are the corresponding omission–commission scaling parameters. By definition, $\omega_{y=a} \geq 1$ and $\omega_{y\neq a} \leq 1$. Hence

$$\tilde{r}(s_M, a, y \neq a) - \tilde{r}(s_M, a, y = a) \;\leq\; (\mu_{y\neq a} - \mu_{y=a})(\overline{r} - \underline{r}).$$

We assume $\frac{\alpha}{\gamma} > \overline{\kappa} \geq \sup_{\mu}(\mu_{y\neq a} - \mu_{y=a})$, so the inequality holds.

For parts (B) and (C), recall that $\tilde{r}(s_M, a, y) = \omega(a, y, s_M) \cdot \mathbb{E}_Y[r(a, y) \mid s_M]$, and that Proposition EC.1 establishes

$$\mathbb{E}_Y[r(a,y) \mid \underbrace{s_M = a}_{\text{follow}}] > \mathbb{E}_Y[r(a,y) \mid \underbrace{s_M \neq a}_{\text{override}}].$$

Part (B). When the outcome is bad ($y \neq a$), the omission–commission parameters satisfy $\omega(s_M, a = s_M, y \neq a) = 1$ and $\omega(s_M, a \neq s_M, y \neq a) = \omega_b < 1$. Since follows and overrides differ only in ω, while their expected monetary payoff satisfies the inequality above, managers assign a higher ex-ante value to a *follow* than to an *override*:

$$\tilde{r}(s_M,\ a = s_M,\ y \neq a) > \tilde{r}(s_M,\ a \neq s_M,\ y \neq a).$$

Part (C). When the outcome is good ($y = a$), the parameters satisfy $\omega(s_M, a = s_M, y = a) = 1$ and $\omega(s_M, a \neq s_M, y = a) = \omega_g > 1$. In this case, the larger omission–commission weight applied to overrides can overturn the advantage of following. There exists a threshold $\omega^* = \frac{\mathbb{E}_Y[r(a,y)|s_M=a, y=a]}{\mathbb{E}_Y[r(a,y)|s_M\neq a, y=a]} > 1$ such that:

- If $\omega_g > \omega^*$, overrides are rewarded more than follows:

$$\tilde{r}(s_M,\ a = s_M,\ y = a) < \tilde{r}(s_M,\ a \neq s_M,\ y = a).$$

- If $\omega_g = \omega^*$, overrides and follows receive the same ex-ante evaluation:

$$\tilde{r}(s_M,\ a = s_M,\ y = a) = \tilde{r}(s_M,\ a \neq s_M,\ y = a).$$

- If $\omega_g < \omega^*$, the expected-payoff advantage dominates the commission premium, and follows are rewarded more than overrides:

$$\tilde{r}(s_M,\ a = s_M,\ y = a) > \tilde{r}(s_M,\ a \neq s_M,\ y = a).$$

This concludes the proof. ■

Proof of Lemma EC.9: As is evident from Eq. 1, the key comparative statistic for evaluating differences in the DM's decision precision across delegated ($i = d$) and non-delegated ($i = n$) settings is the ratio of the DM's expected utility from taking the manager-preferred action $\overline{a}$ to that from taking the alternative action $\underline{a}$. In the stochastic choice framework, this ratio uniquely determines the DM's choice probabilities:

$$\frac{\mathbb{E}_Y[\mathcal{U}_i^{dm}(\underline{a}, y) \mid \mathbf{s}]}{\mathbb{E}_Y[\mathcal{U}_i^{dm}(\overline{a}, y) \mid \mathbf{s}]}.$$

We evaluate this statistic for the four relevant informational states:

$$(\theta = hi,\, s_M = s_H), \qquad (\theta = hi,\, s_M \neq s_H), \qquad (\theta = lo,\, s_M = s_H), \qquad (\theta = lo,\, s_M \neq s_H).$$

For each case, under $i = d$, we analyze the setting with assuming $\omega(a, y, s_M) = 1$, corresponding to bonus payments chosen by a rational but fairness-minded manager under delegation. Recall that, $b^* = \frac{1}{2}\left(\lambda \cdot r(a,y) + (1-\lambda)\cdot \tilde{r}(s_M, a, y)\right)$ where $\lambda = \frac{\alpha}{\alpha+\gamma}$. This implies for the case of $(\theta = hi,\, s_M = s_H)$:

$$\begin{aligned}\mathbb{E}_Y[\mathcal{U}_d^{dm}(\overline{a}, y)|\theta = hi, s_M = s_H] &= \tfrac{1}{2}\Big(\lambda \cdot \big(\mu_{RR}^{hi}\cdot \overline{r} + (1-\mu_{RR}^{hi})\cdot \underline{r}\big) + (1-\lambda)\cdot\big(\hat{\mu}^1_{hi,s_M=s_H}\cdot\overline{r} + (1-\hat{\mu}^1_{hi,s_M=s_H})\cdot\underline{r}\big)\Big)\\ &= \tfrac{1}{2}\big(\mu^1_{hi,s_M=s_H}\cdot\overline{r} + (1-\mu^1_{hi,s_M=s_H})\cdot\underline{r}\big)\end{aligned} \tag{EC.43}$$

where $\mu^1_{hi,s_M=s_H} = \lambda\,\mu^{hi}_{RR} + (1-\lambda)\,\hat{\mu}^1_{hi,s_M=s_H}$, and where $\hat{\mu}^1_{hi,s_M=s_H}$ represents the DM's expectation of how the manager perceives the DM's posterior belief, given that the manager does not observe the human signal (the manager's assessment of the ex-ante decision quality). Under $\omega(a, y, s_M) = 1$, this perceived posterior reduces to a weighted average of $\{\mu^{hi}_{RR},\, 1-\mu^{hi}_{BR},\, \mu^{lo}_{RR},\, \mu^{lo}_{RB}\}$ (see Eqs. EC.7–EC.9).

By the same reasoning, we obtain:

$$\begin{aligned}
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\underline{a}, y)|\theta = hi, s_M = s_H] &= \tfrac{1}{2}\big(\mu^0_{hi,s_M=s_H}\cdot\overline{r} + (1-\mu^0_{hi,s_M=s_H})\cdot\underline{r}\big)\\
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\overline{a}, y)|\theta = hi, s_M \neq s_H] &= \tfrac{1}{2}\big(\mu^1_{hi,s_M\neq s_H}\cdot\overline{r} + (1-\mu^1_{hi,s_M\neq s_H})\cdot\underline{r}\big)\\
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\underline{a}, y)|\theta = hi, s_M \neq s_H] &= \tfrac{1}{2}\big(\mu^0_{hi,s_M\neq s_H}\cdot\overline{r} + (1-\mu^0_{hi,s_M\neq s_H})\cdot\underline{r}\big)\\
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\overline{a}, y)|\theta = lo, s_M = s_H] &= \tfrac{1}{2}\big(\mu^1_{lo,s_M=s_H}\cdot\overline{r} + (1-\mu^1_{lo,s_M=s_H})\cdot\underline{r}\big)\\
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\underline{a}, y)|\theta = lo, s_M = s_H] &= \tfrac{1}{2}\big(\mu^0_{lo,s_M=s_H}\cdot\overline{r} + (1-\mu^0_{lo,s_M=s_H})\cdot\underline{r}\big)\\
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\overline{a}, y)|\theta = lo, s_M \neq s_H] &= \tfrac{1}{2}\big(\mu^1_{lo,s_M\neq s_H}\cdot\overline{r} + (1-\mu^1_{lo,s_M\neq s_H})\cdot\underline{r}\big)\\
\mathbb{E}_Y[\mathcal{U}_d^{dm}(\underline{a}, y)|\theta = lo, s_M \neq s_H] &= \tfrac{1}{2}\big(\mu^0_{lo,s_M\neq s_H}\cdot\overline{r} + (1-\mu^0_{lo,s_M\neq s_H})\cdot\underline{r}\big)
\end{aligned} \tag{EC.44}$$

In the *no delegation* setting ($i = n$), expected utilities reduce to the standard posterior-weighted payoffs:

$$\mathbb{E}_Y[\mathcal{U}_n^{dm}(a, y)\,|\,\mathbf{s}] = \Pr(Y = a\,|\,\mathbf{s})\,\overline{r} \;+\; \Pr(Y \neq a\,|\,\mathbf{s})\,\underline{r},$$

giving the expressions listed below:

$$\begin{aligned}
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\overline{a}, y)|\theta = hi, s_M = s_H] &= \mu^{hi}_{RR}\cdot\overline{r} + (1-\mu^{hi}_{RR})\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\underline{a}, y)|\theta = hi, s_M = s_H] &= (1-\mu^{hi}_{RR})\cdot\overline{r} + \mu^{hi}_{RR}\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\overline{a}, y)|\theta = hi, s_M \neq s_H] &= \mu^{hi}_{BR}\cdot\overline{r} + (1-\mu^{hi}_{BR})\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\underline{a}, y)|\theta = hi, s_M \neq s_H] &= (1-\mu^{hi}_{BR})\cdot\overline{r} + \mu^{hi}_{BR}\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\overline{a}, y)|\theta = lo, s_M = s_H] &= \mu^{lo}_{RR}\cdot\overline{r} + (1-\mu^{lo}_{RR})\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\underline{a}, y)|\theta = lo, s_M = s_H] &= (1-\mu^{lo}_{RR})\cdot\overline{r} + \mu^{lo}_{RR}\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\overline{a}, y)|\theta = lo, s_M \neq s_H] &= \mu^{lo}_{RB}\cdot\overline{r} + (1-\mu^{lo}_{RB})\cdot\underline{r}\\
\mathbb{E}_Y[\mathcal{U}_n^{dm}(\underline{a}, y)|\theta = lo, s_M \neq s_H] &= (1-\mu^{lo}_{RB})\cdot\overline{r} + \mu^{lo}_{RB}\cdot\underline{r}
\end{aligned} \tag{EC.45}$$

Note that $\omega(s_M, a, y) = 1$ implies that $\mu^1_{hi,s_M=s_H}$ and $\mu^0_{hi,s_M\neq s_H}$ reduce to weighted averages of $\{\mu^{hi}_{RR},\, 1-\mu^{hi}_{BR},\, \mu^{lo}_{RR},\, \mu^{lo}_{RB}\}$, while $\mu^1_{hi,s_M\neq s_H}$ and $\mu^0_{hi,s_M=s_H}$ reduce to weighted averages of $\{1-\mu^{hi}_{RR},\, \mu^{hi}_{BR},\, 1-\mu^{lo}_{RR},\, 1-$

$\mu^{lo}_{RB}\}$, where the weights reflect the manager's belief about the DM's (unobserved) human signal (see Eqs. EC.7-EC.10). Also recall that $\mu^{hi}_{RR} > \mu^{hi}_{BR} > 1/2$, $\mu^{lo}_{RR} > \mu^{lo}_{RB} > 1/2$, and $\mu^{hi}_{RR} > \mu^{lo}_{RR}$. Therefore, when $\theta = hi$, it is clear for the both cases of aligning signals ($s_H = s_M$) and conflicting signals ($s_H \neq s_M$) that:

$$\begin{aligned}
\mu^1_{hi,s_M=s_H} < \mu^{hi}_{RR} \quad &\rightarrow \quad \mathbb{E}_Y[\mathcal{U}^{dm}_d(\overline{a},y)|\theta=hi,s_M=s_H] < \mathbb{E}_Y[\mathcal{U}^{dm}_n(\overline{a},y)|\theta=hi,s_M=s_H] \\
\mu^0_{hi,s_M=s_H} > 1-\mu^{hi}_{RR} \quad &\rightarrow \quad \mathbb{E}_Y[\mathcal{U}^{dm}_d(\underline{a},y)|\theta=hi,s_M=s_H] > \mathbb{E}_Y[\mathcal{U}^{dm}_n(\underline{a},y)|\theta=hi,s_M=s_H] \\
\mu^1_{hi,s_M\neq s_H} < \mu^{hi}_{BR} \quad &\rightarrow \quad \mathbb{E}_Y[\mathcal{U}^{dm}_d(\overline{a},y)|\theta=hi,s_M\neq s_H] < \mathbb{E}_Y[\mathcal{U}^{dm}_n(\overline{a},y)|\theta=hi,s_M\neq s_H] \\
\mu^0_{hi,s_M\neq s_H} > 1-\mu^{hi}_{BR} \quad &\rightarrow \quad \mathbb{E}_Y[\mathcal{U}^{dm}_d(\underline{a},y)|\theta=hi,s_M\neq s_H] > \mathbb{E}_Y[\mathcal{U}^{dm}_n(\underline{a},y)|\theta=hi,s_M\neq s_H]
\end{aligned} \tag{EC.46}$$

Now, applying Eq. 1 implies

$$\mathcal{P}_d(A=s_H|\theta=hi,s_M,s_H) < \mathcal{P}_n(A=s_H|\theta=hi,s_M,s_H). \tag{EC.47}$$

When $\theta = lo$, let

$$\begin{aligned}
u^d_{align} &= \frac{\mathbb{E}_Y[\mathcal{U}^{dm}_d(\underline{a},y)|\theta=lo,s_M=s_H]}{\mathbb{E}_Y[\mathcal{U}^{dm}_d(\overline{a},y)|\theta=lo,s_M=s_H]} = \frac{\mu^0_{lo,s_M=s_H}\cdot\overline{r}+(1-\mu^0_{lo,s_M=s_H})\cdot\underline{r}}{\mu^1_{lo,s_M=s_H}\cdot\overline{r}+(1-\mu^1_{lo,s_M=s_H})\cdot\underline{r}}, \\
u^n_{align} &= \frac{\mathbb{E}_Y[\mathcal{U}^{dm}_n(\underline{a},y)|\theta=lo,s_M=s_H]}{\mathbb{E}_Y[\mathcal{U}^{dm}_n(\overline{a},y)|\theta=lo,s_M=s_H]} = \frac{(1-\mu^{lo}_{RR})\cdot\overline{r}+\mu^{lo}_{RR}\cdot\underline{r}}{\mu^{lo}_{RR}\cdot\overline{r}+(1-\mu^{lo}_{RR})\cdot\underline{r}}, \\
u^d_{conflict} &= \frac{\mathbb{E}_Y[\mathcal{U}^{dm}_d(\underline{a},y)|\theta=lo,s_M\neq s_H]}{\mathbb{E}_Y[\mathcal{U}^{dm}_d(\overline{a},y)|\theta=lo,s_M\neq s_H]} = \frac{\mu^0_{lo,s_M\neq s_H}\cdot\overline{r}+(1-\mu^0_{lo,s_M\neq s_H})\cdot\underline{r}}{\mu^1_{lo,s_M=s_H}\cdot\overline{r}+(1-\mu^1_{lo,s_M=s_H})\cdot\underline{r}}, \\
u^n_{conflict} &= \frac{\mathbb{E}_Y[\mathcal{U}^{dm}_n(\underline{a},y)|\theta=lo,s_M\neq s_H]}{\mathbb{E}_Y[\mathcal{U}^{dm}_n(\overline{a},y)|\theta=lo,s_M\neq s_H]} = \frac{(1-\mu^{lo}_{RB})\cdot\overline{r}+\mu^{lo}_{RB}\cdot\underline{r}}{\mu^{lo}_{RB}\cdot\overline{r}+(1-\mu^{lo}_{RB})\cdot\underline{r}}.
\end{aligned} \tag{EC.48}$$

Note that here, the action $\overline{a}$ corresponds with *following* the machine signal, and $\underline{a}$ with *overriding* it. Therefore, Lemma EC.6 implies that both $u^d_{align}(\beta)$ and $u^d_{conflict}(\beta)$ are decreasing functions in β. By definition we have

$$\begin{aligned}
\mu^1_{lo,s_M=s_H} &= \lambda\cdot\mu^{lo}_{RR}+(1-\lambda)\cdot\hat{\mu}^1_{lo,s_M=s_H}, \\
\mu^0_{lo,s_M=s_H} &= \lambda\cdot(1-\mu^{lo}_{RR})+(1-\lambda)\cdot\hat{\mu}^0_{lo,s_M=s_H} \\
\mu^1_{lo,s_M\neq s_H} &= \lambda\cdot\mu^{lo}_{RB}+(1-\lambda)\cdot\hat{\mu}^1_{lo,s_M\neq s_H}, \\
\mu^0_{lo,s_M\neq s_H} &= \lambda\cdot(1-\mu^{lo}_{RB})+(1-\lambda)\cdot\hat{\mu}^0_{lo,s_M\neq s_H},
\end{aligned} \tag{EC.49}$$

where the $\hat{\mu}$ terms represent the DM's expectation of how the manager perceives the DM's posterior belief about the state, given that the manager does not observe the human signal. These expectations are derived from Eqs. EC.7–EC.10. Under $\omega(a,y,s_M)=1$, the terms $\hat{\mu}^1_{lo,s_M,s_H}$ reduce to weighted averages of $\{\mu^{hi}_{RR}, 1-\mu^{hi}_{BR}, \mu^{lo}_{RR}, \mu^{lo}_{RB}\}$, whereas the terms $\hat{\mu}^0_{lo,s_M=s_H}$ reduce to weighted averages of $\{1-\mu^{hi}_{RR}, \mu^{hi}_{BR}, 1-\mu^{lo}_{RR}, 1-\mu^{lo}_{RB}\}$. The weights reflect the manager's belief about the DM's unobserved human signal and depend on the parameter β. For the case $\theta = lo$ and $s_H = s_M$, in the limit ($\beta \to \infty$) we obtain:

$$\begin{aligned}
\lim_{\beta\to\infty}\mu^1_{lo,s_M=s_H} =&\Big(\mu^{lo}_{RR}q^{hi}_{s_H}\mu_\theta+(1-\mu^{lo}_{RR})(1-q^{hi}_{s_H})\mu_\theta\Big)\mu^{hi}_{RR}\\
&+\Big(\mu^{lo}_{RR}(1-q^{hi}_{s_H})\mu_\theta+(1-\mu^{lo}_{RR})q^{hi}_{s_H}\mu_\theta\Big)(1-\mu^{hi}_{BR})\\
&+\Big(\mu^{lo}_{RR}q^{lo}_{s_H}(1-\mu_\theta)+(1-\mu^{lo}_{RR})(1-q^{lo}_{s_H})(1-\mu_\theta)\Big)\mu^{lo}_{RR}\\
&+\Big(\mu^{lo}_{RR}(1-q^{lo}_{s_H})(1-\mu_\theta)+(1-\mu^{lo}_{RR})q^{lo}_{s_H}(1-\mu_\theta)\Big)\mu^{lo}_{RB}\\
\lim_{\beta\to\infty}\mu^0_{lo,s_M=s_H} =&\Big(\mu^{lo}_{RR}(1-q^{hi}_{s_H})\mu_\theta+(1-\mu^{lo}_{RR})q^{hi}_{s_H}\mu_\theta\Big)\mu^{hi}_{BR}\\
&+\Big(\mu^{lo}_{RR}q^{hi}_{s_H}\mu_\theta+(1-\mu^{lo}_{RR})(1-q^{hi}_{s_H})\mu_\theta\Big)(1-\mu^{hi}_{RR})\\
&+\Big(\mu^{lo}_{RR}(1-q^{lo}_{s_H})(1-\mu_\theta)+(1-\mu^{lo}_{RR})q^{lo}_{s_H}(1-\mu_\theta)\Big)(1-\mu^{lo}_{RB})\\
&+\Big(\mu^{lo}_{RR}q^{lo}_{s_H}(1-\mu_\theta)+(1-\mu^{lo}_{RR})(1-q^{lo}_{s_H})(1-\mu_\theta)\Big)(1-\mu^{lo}_{RR})
\end{aligned}\tag{EC.50}$$

Consequently, we get:

$$\lim_{\beta\to\infty}\frac{\hat\mu^0_{lo,s_M=s_H}}{\hat\mu^1_{lo,s_M=s_H}}>\frac{1-\mu^{lo}_{RR}}{\mu^{lo}_{RR}}\tag{EC.51}$$

Since $\frac{\hat\mu^0_{lo,s_M=s_H}}{\hat\mu^1_{lo,s_M=s_H}}$ is decreasing in β, we get $\frac{\hat\mu^0_{lo,s_M=s_H}}{\hat\mu^1_{lo,s_M=s_H}}>\frac{1-\mu^{lo}_{RR}}{\mu^{lo}_{RR}}$ for all $\beta>0$, which is equivalent of $u^d_{align}>u^n_{align}$. Applying Eq. 1 implies

$$\mathcal{P}_d(A=s_H|\theta=lo,s_M,s_H=s_M)<\mathcal{P}_n(A=s_H|\theta=lo,s_M,s_H=s_M).\tag{EC.52}$$

Similarly, for the case of $\theta=lo$ and $s_H\neq s_M$, in the limits $(\beta\to0)$ and $(\beta\to\infty)$ we have:

$$\begin{aligned}
\lim_{\beta\to0}\mu^1_{lo,s_M\neq s_H} =&\Big(\mu^{lo}_{RB}\frac{q^{hi}_{s_H}\mu}{q^{hi}_{s_H}\mu+1-\mu}+(1-\mu^{lo}_{RB})\frac{(1-q^{hi}_{s_H})\mu}{(1-q^{hi}_{s_H})\mu+1-\mu}\Big)\mu^{hi}_{RR}\\
&+\Big(\mu^{lo}_{RB}\frac{q^{lo}_{s_H}(1-\mu)}{q^{hi}_{s_H}\mu+1-\mu}+(1-\mu^{lo}_{RB})\frac{(1-q^{lo}_{s_H})(1-\mu)}{(1-q^{hi}_{s_H})\mu+1-\mu}\Big)\mu^{lo}_{RR}\\
&+\Big(\mu^{lo}_{RB}\frac{(1-q^{lo}_{s_H})(1-\mu)}{q^{hi}_{s_H}\mu+1-\mu}+(1-\mu^{lo}_{RB})\frac{q^{lo}_{s_H}(1-\mu)}{(1-q^{hi}_{s_H})\mu+1-\mu}\Big)\mu^{lo}_{RB},\\
\lim_{\beta\to0}\mu^0_{lo,s_M\neq s_H} =&\mu^{hi}_{BR},\\
\lim_{\beta\to\infty}\mu^1_{lo,s_M\neq s_H} =&\Big(\mu^{lo}_{RB}q^{hi}_{s_H}\mu_\theta+(1-\mu^{lo}_{RB})(1-q^{hi}_{s_H})\mu_\theta\Big)\mu^{hi}_{RR}\\
&+\Big(\mu^{lo}_{RB}(1-q^{hi}_{s_H})\mu_\theta+(1-\mu^{lo}_{RB})q^{hi}_{s_H}\mu_\theta\Big)(1-\mu^{hi}_{BR})\\
&+\Big(\mu^{lo}_{RB}q^{lo}_{s_H}(1-\mu_\theta)+(1-\mu^{lo}_{RB})(1-q^{lo}_{s_H})(1-\mu_\theta)\Big)\mu^{lo}_{RR}\\
&+\Big(\mu^{lo}_{RB}(1-q^{lo}_{s_H})(1-\mu_\theta)+(1-\mu^{lo}_{RB})q^{lo}_{s_H}(1-\mu_\theta)\Big)\mu^{lo}_{RB},\\
\lim_{\beta\to\infty}\mu^0_{lo,s_M\neq s_H} =&\Big(\mu^{lo}_{RB}(1-q^{hi}_{s_H})\mu_\theta+(1-\mu^{lo}_{RB})q^{hi}_{s_H}\mu_\theta\Big)\mu^{hi}_{BR}\\
&+\Big(\mu^{lo}_{RB}q^{hi}_{s_H}\mu_\theta+(1-\mu^{lo}_{RB})(1-q^{hi}_{s_H})\mu_\theta\Big)(1-\mu^{hi}_{RR})\\
&+\Big(\mu^{lo}_{RB}(1-q^{lo}_{s_H})(1-\mu_\theta)+(1-\mu^{lo}_{RB})q^{lo}_{s_H}(1-\mu_\theta)\Big)(1-\mu^{lo}_{RB})\\
&+\Big(\mu^{lo}_{RB}q^{lo}_{s_H}(1-\mu_\theta)+(1-\mu^{lo}_{RB})(1-q^{lo}_{s_H})(1-\mu_\theta)\Big)(1-\mu^{lo}_{RR}).
\end{aligned}\tag{EC.53}$$

From this we get:

$$\lim_{\beta\to\infty}\frac{\hat{\mu}^0_{lo,s_M\neq s_H}}{\hat{\mu}^1_{lo,s_M\neq s_H}}<\frac{1-\mu^{lo}_{RB}}{\mu^{lo}_{RB}} \tag{EC.54}$$

However, when β approaches 0, depending on the parameter values, $\lim_{\beta\to 0}\frac{\hat{\mu}^0_{lo,s_M\neq s_H}}{\hat{\mu}^1_{lo,s_M\neq s_H}}$ can be larger, equal, or smaller than $\frac{1-\mu^{lo}_{RB}}{\mu^{lo}_{RB}}$. In particular, for sufficiently high values of $q^{hi}_{s_H}$ it is always larger, while for sufficiently low values of $q^{hi}_{s_H}$ it is always smaller. Therefore, in general, there must exist a $\underline{\beta}_0\geq 0$ where for $\beta>\underline{\beta}_0$ we have $\frac{\hat{\mu}^0_{lo,s_M\neq s_H}}{\hat{\mu}^1_{lo,s_M\neq s_H}}<\frac{1-\mu^{lo}_{RB}}{\mu^{lo}_{RB}}$ or equivalently $u^d_{conflict}>u^n_{conflict}$. Applying Eq. 1 implies that for $\beta>\underline{\beta}_0$

$$\mathcal{P}_d(A=s_H|\theta=lo,s_M,s_H\neq s_M)>\mathcal{P}_n(A=s_H|\theta=lo,s_M,s_H\neq s_M). \tag{EC.55}$$

With $\underline{\beta}_0<\underline{\beta}<\beta$ the proof is concluded. ■

Proof of Proposition EC.2. Fix an informational state $\mathbf{s}=(s_M,s_H,\theta)$. The constants $\xi_{\mathbf{s}}$ and $\zeta_{\mathbf{s}}$ collect the expected-payoff terms that scale the evaluation tilt from omission–commission bias after a good versus bad override, while $\varphi_{\mathbf{s}}$ captures the residual baseline wedge that determines the precision ordering when omission–commission bias is absent. In particular, when $\omega_g=\omega_b=1$ the left-hand side of the threshold inequality is zero, so the sign of $\varphi_{\mathbf{s}}$ alone selects the relevant ordering. Define $\xi_{\mathbf{s}}$, $\zeta_{\mathbf{s}}$, and $\varphi_{\mathbf{s}}$ as follows.
If $\mathbf{s}=(s_M,s_H\neq s_M,\theta=hi)$, then

$$\begin{aligned}
\xi_{\mathbf{s}}&=\mu_{\mathbf{s}}\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y=a],\\
\zeta_{\mathbf{s}}&=(1-\mu_{\mathbf{s}})\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y\neq a],\\
\varphi_{\mathbf{s}}&=\frac{\mu_{\mathbf{s}}\overline{r}+(1-\mu_{\mathbf{s}})\underline{r}}{\mu_{\mathbf{s}}\underline{r}+(1-\mu_{\mathbf{s}})\overline{r}}\Big(\mu_{\mathbf{s}}\cdot\mathbb{E}_Y[r(a,y)\mid s_M=a,\ y\neq a]+(1-\mu_{\mathbf{s}})\cdot\mathbb{E}_Y[r(a,y)\mid s_M=a,\ y=a]\Big)\\
&\quad-\Big(\mu_{\mathbf{s}}\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y=a]+(1-\mu_{\mathbf{s}})\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y\neq a]\Big).
\end{aligned}$$

Otherwise,

$$\begin{aligned}
\xi_{\mathbf{s}}&=(1-\mu_{\mathbf{s}})\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y=a],\\
\zeta_{\mathbf{s}}&=\mu_{\mathbf{s}}\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y\neq a],\\
\varphi_{\mathbf{s}}&=\frac{\mu_{\mathbf{s}}\underline{r}+(1-\mu_{\mathbf{s}})\overline{r}}{\mu_{\mathbf{s}}\overline{r}+(1-\mu_{\mathbf{s}})\underline{r}}\Big(\mu_{\mathbf{s}}\cdot\mathbb{E}_Y[r(a,y)\mid s_M=a,\ y=a]+(1-\mu_{\mathbf{s}})\cdot\mathbb{E}_Y[r(a,y)\mid s_M=a,\ y\neq a]\Big)\\
&\quad-\Big(\mu_{\mathbf{s}}\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y\neq a]+(1-\mu_{\mathbf{s}})\cdot\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y=a]\Big).
\end{aligned}$$

From the proof of Lemma EC.9 recall that, for any $\beta>\underline{\beta}$,

$$\mathcal{P}_d(A=s_H\mid\mathbf{s})<\mathcal{P}_n(A=s_H\mid\mathbf{s})\quad\text{if and only if}\quad\frac{\mathbb{E}_Y[\mathcal{U}^{dm}_d(\underline{a},y)\mid\mathbf{s}]}{\mathbb{E}_Y[\mathcal{U}^{dm}_d(\overline{a},y)\mid\mathbf{s}]}>\frac{\mathbb{E}_Y[\mathcal{U}^{dm}_n(\underline{a},y)\mid\mathbf{s}]}{\mathbb{E}_Y[\mathcal{U}^{dm}_n(\overline{a},y)\mid\mathbf{s}]}.$$

We begin with the informational state $\mathbf{s}=(s_M,s_H=s_M,\theta=hi)$. In this case the expected utilities are

$$\begin{aligned}
&\mathbb{E}_Y[\mathcal{U}^{dm}_d(\underline{a},y)\mid s_M,s_H=s_M,\theta=hi]=\omega_b\mu^{hi}_{RR}\,\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y\neq a]+\omega_g(1-\mu^{hi}_{RR})\,\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\ y=a],\\
&\mathbb{E}_Y[\mathcal{U}^{dm}_d(\overline{a},y)\mid s_M,s_H=s_M,\theta=hi]=\mu^{hi}_{RR}\,\mathbb{E}_Y[r(a,y)\mid s_M=a,\ y=a]+(1-\mu^{hi}_{RR})\,\mathbb{E}_Y[r(a,y)\mid s_M=a,\ y\neq a],\\
&\mathbb{E}_Y[\mathcal{U}^{dm}_n(\underline{a},y)\mid s_M,s_H=s_M,\theta=hi]=\mu^{hi}_{RR}\underline{r}+(1-\mu^{hi}_{RR})\overline{r},\\
&\mathbb{E}_Y[\mathcal{U}^{dm}_n(\overline{a},y)\mid s_M,s_H=s_M,\theta=hi]=\mu^{hi}_{RR}\overline{r}+(1-\mu^{hi}_{RR})\underline{r}.
\end{aligned} \tag{EC.56}$$

For the constants we have:

$$\xi_{(s_M,s_H=s_M,\theta=hi)} = (1-\mu_{RR}^{hi})\,\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\, y=a]\,,$$
$$\zeta_{(s_M,s_H=s_M,\theta=hi)} = \mu_{RR}^{hi}\,\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\, y\neq a]\,,$$
$$\varphi_{(s_M,s_H=s_M,\theta=hi)} = \frac{\mu_{RR}^{hi}\underline{r}+(1-\mu_{RR}^{hi})\overline{r}}{\mu_{RR}^{hi}\overline{r}+(1-\mu_{RR}^{hi})\underline{r}}\Big(\mu_{RR}^{hi}\,\mathbb{E}_Y[r(a,y)\mid s_M=a,\, y=a]+(1-\mu_{RR}^{hi})\,\mathbb{E}_Y[r(a,y)\mid s_M=a,\, y\neq a]\Big)$$
$$-\Big(\mu_{RR}^{hi}\,\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\, y\neq a]+(1-\mu_{RR}^{hi})\,\mathbb{E}_Y[r(a,y)\mid s_M\neq a,\, y=a]\Big).$$

Substituting (EC.56) into the inequality above and simplifying yields that in this informational state

$$\mathcal{P}_d(A=s_H\mid\theta=hi,s_M=s_H)\le\mathcal{P}_n(A=s_H\mid\theta=hi,s_M=s_H)\qquad\text{iff}$$

$$\xi_{(s_M,s_H=s_M,\theta=hi)}(\omega_g-1)-\zeta_{(s_M,s_H=s_M,\theta=hi)}(1-\omega_b)\ge\varphi_{(s_M,s_H=s_M,\theta=hi)}.$$

Since both probabilities lie in $(0.5,1)$, this implies

$$0.5<\mathcal{P}_d(A=s_H\mid\theta=hi,s_M=s_H)\le\mathcal{P}_n(A=s_H\mid\theta=hi,s_M=s_H)<1,$$

and the reverse inequality holds otherwise.

The remaining informational states are handled analogously. Carrying out the corresponding algebra leads to the conditions stated in the proposition. ■

Proof of Theorem 2. The result follows directly from Proposition EC.2 together with the assumption

$$\frac{\omega_g-1}{1-\omega_b}<\overline{\omega}\;\le\;\min\{\hat{\omega}_{hi,s_M\neq s_H},\,\hat{\omega}_{lo,s_M\neq s_H}\},$$

where

$$\hat{\omega}_{hi,s_M\neq s_H}:=\frac{\zeta_{(s_M,s_H\neq s_M,\theta=hi)}}{\xi_{(s_M,s_H\neq s_M,\theta=hi)}},\qquad\hat{\omega}_{lo,s_M\neq s_H}:=\frac{\zeta_{(s_M,s_H\neq s_M,\theta=lo)}}{\xi_{(s_M,s_H\neq s_M,\theta=lo)}}.$$

■

Proof of Proposition EC.3. As in Proposition 1, in the Observable setting ($i=o$) the manager's equilibrium bonus is

$$b_o^*(\mathbf{s},a,y)=\frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma}r(a,y)+\frac{\gamma}{\alpha+\gamma}\,\tilde{r}(\mathbf{s},a,y)\right),$$

where

$$\tilde{r}(\mathbf{s},a,y)\equiv\omega(s_M,a,y)\cdot\mathbb{E}_Y[r(a,y)\mid\mathbf{s}]\,.$$

Fix $(\mathbf{s},y)$ and compare bonuses under the two actions $a=\bar{a}$ (the ex-ante optimal action given $\mathbf{s}$) and $a\neq\bar{a}$. Since $r(a,y)$ is the same across actions whenever the realized outcome type (good/bad) is fixed, the ranking of b_o^* across actions is determined by $\tilde{r}(\mathbf{s},a,y)$, i.e., by the product of the omission–commission term $\omega(\cdot)$ and the ex-ante payoff $\mathbb{E}_Y[r(a,y)\mid\mathbf{s}]$.

Step 1 (baseline ranking). By definition of $\bar{a}$,

$$\mathbb{E}_Y[r(\bar{a},y)\mid\mathbf{s}]>\mathbb{E}_Y[r(a\neq\bar{a},y)\mid\mathbf{s}]\,.$$

Hence, in the absence of omission–commission considerations ($\omega(s_M,a,y)=1$), the bonus is higher for the ex-ante optimal decision.

Step 2 (cases where ω can overturn the baseline). Since we assume $\omega_g > 1$ and $\omega_b < 1$, potential reversals arise only in cases where outcome is good ($y = a$) and the optimal decision is considered an omission ($\overline{a} = s_M$), or when outcome is bad ($y \neq a$) and the optimal decision is considered a commission ($\overline{a} \neq s_M$). To express the resulting thresholds compactly, define

$$\hat{\omega}_g^{lo} \equiv \frac{\mathbb{E}_Y[r(a = s_M, y = a) \mid \theta = lo, s_M, s_H]}{\mathbb{E}_Y[r(a \neq s_M, y = a) \mid \theta = lo, s_M, s_H]},$$

$$\hat{\omega}_g^{hi} \equiv \frac{\mathbb{E}_Y[r(a = s_H, y = a) \mid \theta = hi, s_M, s_H = s_M]}{\mathbb{E}_Y[r(a \neq s_H, y = a) \mid \theta = hi, s_M, s_H = s_M]},$$

and

$$\hat{\omega}_b \equiv \frac{\mathbb{E}_Y[r(a \neq s_H, y \neq a) \mid \theta = hi, s_M, s_H \neq s_M]}{\mathbb{E}_Y[r(a = s_H, y \neq a) \mid \theta = hi, s_M, s_H \neq s_M]}.$$

With these definitions, comparing $b_o^*(\mathbf{s}, a, y)$ across a reduces to comparing $\omega(s_M, a, y) \cdot \mathbb{E}_Y[r(a, y) \mid \mathbf{s}]$ across a, which yields the stated rankings: (i) when $\theta = hi$, bad outcomes: alignment ($s_H = s_M$) implies the ex-ante optimal decision strictly dominates; under conflict ($s_H \neq s_M$) the ranking depends on whether ω_b is above or below $\hat{\omega}_b$; (ii) when $\theta = hi$, good outcomes: under conflict ($s_H \neq s_M$) the ex-ante optimal decision strictly dominates, whereas under alignment ($s_H = s_M$) the ranking depends on whether ω_g is below or above $\hat{\omega}_g^{hi}$; (iii) when $\theta = lo$, bad outcomes: the ex-ante optimal decision strictly dominates; and (iv) when $\theta = lo$, good outcomes: the ranking depends on whether ω_g is below or above $\hat{\omega}_g^{lo}$. This concludes the proof. ■

Proof of Proposition EC.4. As in Proposition 1, in the Observable setting ($i = o$) the manager's equilibrium bonus is

$$b_o^*(\mathbf{s}, a, y) = \frac{1}{2}\left(\frac{\alpha}{\alpha + \gamma} r(a, y) + \frac{\gamma}{\alpha + \gamma}\, \tilde{r}(\mathbf{s}, a, y)\right),$$

where

$$\tilde{r}(\mathbf{s}, a, y) \equiv \omega(s_M, a, y) \cdot \mathbb{E}_Y[r(a, y) \mid \mathbf{s}]\,.$$

This implies that the DM's expected utility from taking a decision (a) reduces to

$$\mathbb{E}_Y[\mathcal{U}_o^{dm}(a, y) | \mathbf{s}] = \frac{\alpha + \gamma \cdot \omega}{2(\alpha + \gamma)} \mathbb{E}_Y[r(a, y) | \mathbf{s}].$$

We first show how the decision precision in $i = o$ compares to that in $i = n$, and then will turn to comparison with $i = d$. In the stochastic choice framework, the ratio uniquely determines the DM's choice probabilities:

$$\frac{\mathbb{E}_Y[\mathcal{U}_i^{dm}(\underline{a}, y) \mid \mathbf{s}]}{\mathbb{E}_Y[\mathcal{U}_i^{dm}(\overline{a}, y) \mid \mathbf{s}]}.$$

Note the in the absence of omission–commission considerations ($\omega = 1$), the DM's expected utility becomes $\mathbb{E}_Y[\mathcal{U}_o^{dm}(a, y) | \mathbf{s}] = \frac{1}{2}\mathbb{E}_Y[r(a, y) | \mathbf{s}]$. Applying the stochastic choice model in 1 yields $\mathcal{P}_o = \mathcal{P}_n$. Now the assumption $(\omega_g - 1)/(1 - \omega_b) < \overline{\omega}$ implies that machine overrides are punished overall and that the DM's expected utility from overriding the machine is scaled down compared to the case without omission–commission considerations. That is:

$$\frac{\mathbb{E}_Y[\mathcal{U}_o^{dm}(\underline{a}, y) \mid \mathbf{s}'_{hi}]}{\mathbb{E}_Y[\mathcal{U}_o^{dm}(\overline{a}, y) \mid \mathbf{s}'_{hi}]} > \frac{\mathbb{E}_Y[\mathcal{U}_n^{dm}(\underline{a}, y) \mid \mathbf{s}'_{hi}]}{\mathbb{E}_Y[\mathcal{U}_n^{dm}(\overline{a}, y) \mid \mathbf{s}'_{hi}]}$$

$$\frac{\mathbb{E}_Y[\mathcal{U}_o^{dm}(\underline{a}, y) \mid \mathbf{s}'_{lo}]}{\mathbb{E}_Y[\mathcal{U}_o^{dm}(\overline{a}, y) \mid \mathbf{s}'_{lo}]} < \frac{\mathbb{E}_Y[\mathcal{U}_n^{dm}(\underline{a}, y) \mid \mathbf{s}'_{lo}]}{\mathbb{E}_Y[\mathcal{U}_n^{dm}(\overline{a}, y) \mid \mathbf{s}'_{lo}]}$$

Applying the stochastic choice model in 1 yields

$$\mathcal{P}_o(\mathbf{s}'_{hi}) < \mathcal{P}_n(\mathbf{s}'_{hi}) \qquad \mathcal{P}_n(\mathbf{s}'_{lo}) < \mathcal{P}_o(\mathbf{s}'_{lo}).$$

Now recall from Lemma EC.9 that in the absence of omission–commission considerations, $\mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_n(\mathbf{s}'_{hi})$ and $\mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_n(\mathbf{s}'_{lo})$. We established above that in the absence of omission–commission considerations $\mathcal{P}_o = \mathcal{P}_n$. Therefore, assuming $\omega = 1$, we get $\mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_o(\mathbf{s}'_{hi})$ and $\mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_o(\mathbf{s}'_{lo})$. Now relaxing the assumption on ω and letting $\omega_g > 1$ and $\omega_b < 1$ would scale the DM's expected utility from overriding in both $i = o$ and $i = d$ in the same way so the orderings $\mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_o(\mathbf{s}'_{hi})$ and $\mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_o(\mathbf{s}'_{lo})$ are preserved. This concludes the proof.

Suppose $(\omega_g - 1)/(1 - \omega_b) < \overline{\omega}$. When signals conflict, i.e., $s_H \neq s_M$:

A. **Overreliance**: When $\theta = hi$ (i.e., $\bar{a} = s_H$): $0.5 < \mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_o(\mathbf{s}'_{hi}) < \mathcal{P}_n(\mathbf{s}'_{hi}) < 1$.

B. **Underreliance**: When $\theta = lo$ (i.e., $\bar{a} = s_M$): $0.5 < \mathcal{P}_n(\mathbf{s}'_{lo}) < \mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_d(\mathbf{s}'_{lo}) < 1$.

LEMMA EC.9 **(Decision Precision Benchmark)**. *Suppose $\beta > \underline{\beta}$ and $\omega(a, y, s_M) = 1$. At equilibrium:*

- *When $\theta = hi$ (the human signal is more precise):*
  - $0.5 < \mathcal{P}_d(\theta = hi, s_M, s_H) < \mathcal{P}_n(\theta = hi, s_M, s_H) < 1$.
- *When $\theta = lo$ (the machine signal is more precise):*
  - $0.5 < \mathcal{P}_d(\theta = lo, s_M = s_H) < \mathcal{P}_n(\theta = lo, s_M = s_H) < 1$.
  - $0.5 < \mathcal{P}_n(\theta = lo, s_M \neq s_H) < \mathcal{P}_d(\theta = lo, s_M \neq s_H) < 1$.

■

Proof of Proposition EC.4. Fix a conflict history $\mathbf{s}'_\theta$ with $s_H \neq s_M$. As in Proposition 1, in the Observable regime ($i = o$) the manager's equilibrium bonus is

$$b_o^*(\mathbf{s}, a, y) = \frac{1}{2}\left(\frac{\alpha}{\alpha+\gamma} r(a, y) + \frac{\gamma}{\alpha+\gamma}\,\tilde{r}(\mathbf{s}, a, y)\right), \qquad \tilde{r}(\mathbf{s}, a, y) \equiv \omega(s_M, a, y) \cdot \mathbb{E}_Y[r(a, y) \mid \mathbf{s}].$$

Therefore the DM's expected utility from action a in $i = o$ is proportional to the ex-ante expected payoff:

$$\mathbb{E}\big[\mathcal{U}_o^{dm}(a, y) \mid \mathbf{s}\big] = \frac{\alpha + \gamma \cdot \omega}{2(\alpha+\gamma)}\,\mathbb{E}[r(a, y) \mid \mathbf{s}]\,, \tag{EC.57}$$

where $\omega = \omega(s_M, a, y)$ denotes the omission–commission weight associated with (s_M, a, y). In the No-Delegation benchmark ($i = n$),

$$\mathbb{E}\big[\mathcal{U}_n^{dm}(a, y) \mid \mathbf{s}\big] = \mathbb{E}[r(a, y) \mid \mathbf{s}]\,. \tag{EC.58}$$

Under the stochastic choice function in (1), the probability of choosing $\overline{a}$ is strictly decreasing in the ratio

$$R_i(\mathbf{s}) \equiv \frac{\mathbb{E}\big[\mathcal{U}_i^{dm}(\underline{a}, y) \mid \mathbf{s}\big]}{\mathbb{E}\big[\mathcal{U}_i^{dm}(\overline{a}, y) \mid \mathbf{s}\big]}, \qquad i \in \{n, o, d\}.$$

Hence, to compare decision precision across regimes it suffices to compare these ratios.

Step 1 (o vs. n under conflict). Using (EC.57)–(EC.58), we can write, for any $\mathbf{s}$,

$$R_o(\mathbf{s}) = \frac{\alpha + \gamma \cdot \omega(\underline{a})}{\alpha + \gamma \cdot \omega(\overline{a})} \cdot \frac{1}{2} \cdot R_n(\mathbf{s}), \tag{EC.59}$$

where $\omega(\underline{a}) \equiv \omega(s_M, \underline{a}, y)$ and $\omega(\bar{a}) \equiv \omega(s_M, \bar{a}, y)$, and

$$R_n(\mathbf{s}) = \frac{\mathbb{E}[r(\underline{a}, y) \mid \mathbf{s}]}{\mathbb{E}[r(\bar{a}, y) \mid \mathbf{s}]}.$$

The factor 1/2 in (EC.59) is common across actions and thus does not by itself change the ranking of $\bar{a}$ vs. $\underline{a}$ within $i = o$; what matters for cross-regime comparisons is the relative weight

$$\frac{\alpha + \gamma \cdot \omega(\underline{a})}{\alpha + \gamma \cdot \omega(\bar{a})}.$$

Case $\theta = hi$ (so $\bar{a} = s_H$ under conflict). When $s_H \neq s_M$, choosing $\bar{a}$ corresponds to *overriding* the machine, while choosing $\underline{a}$ corresponds to *following* the machine. The maintained condition $(\omega_g - 1)/(1 - \omega_b) < \overline{\omega}$ implies that overrides are, on net, penalized relative to follows. Equivalently, in expectation the omission–commission term lowers the relative attractiveness of the override action, which yields

$$R_o(\mathbf{s}'_{hi}) > R_n(\mathbf{s}'_{hi}).$$

By the monotonicity of the Luce rule, this implies

$$\mathcal{P}_o(\mathbf{s}'_{hi}) < \mathcal{P}_n(\mathbf{s}'_{hi}). \tag{EC.60}$$

Case $\theta = lo$ (so $\bar{a} = s_M$ under conflict). Now $\bar{a}$ corresponds to *following* the machine and $\underline{a}$ corresponds to *overriding*. The same "overall penalty for overrides" condition implies that the omission–commission term increases the relative attractiveness of $\bar{a}$, hence

$$R_o(\mathbf{s}'_{lo}) < R_n(\mathbf{s}'_{lo}),$$

and thus

$$\mathcal{P}_n(\mathbf{s}'_{lo}) < \mathcal{P}_o(\mathbf{s}'_{lo}). \tag{EC.61}$$

Step 2 (Bring in d). First note that, as established above, when $\omega(a, y, s_M) = 1$, the Observable regime $i = o$ is equivalent to the No-Delegation benchmark $i = n$, i.e., $\mathcal{P}_o(\mathbf{s}) = \mathcal{P}_n(\mathbf{s})$. From Lemma EC.9 (stated under $\omega \equiv 1$) we know, under conflict,

$$\mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_n(\mathbf{s}'_{hi}) \quad \text{and} \quad \mathcal{P}_n(\mathbf{s}'_{lo}) < \mathcal{P}_d(\mathbf{s}'_{lo}).$$

Using $\mathcal{P}_o = \mathcal{P}_n$ when $\omega \equiv 1$, this gives the corresponding ordering between d and o:

$$\mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_o(\mathbf{s}'_{hi}) \quad \text{and} \quad \mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_d(\mathbf{s}'_{lo}) \qquad (\omega \equiv 1).$$

Finally, relaxing $\omega \equiv 1$ and allowing $\omega_g > 1$ and $\omega_b < 1$ affects the DM's expected utility in both $i = o$ and $i = d$ only through the same multiplicative term $\omega(s_M, a, y)$ attached to overrides (relative to follows). Therefore the sign of the relevant utility ratio comparison (and hence the choice probability ordering) is preserved, so the inequalities continue to hold for general (ω_g, ω_b):

$$\mathcal{P}_d(\mathbf{s}'_{hi}) < \mathcal{P}_o(\mathbf{s}'_{hi}) \quad \text{and} \quad \mathcal{P}_o(\mathbf{s}'_{lo}) < \mathcal{P}_d(\mathbf{s}'_{lo}).$$

Combining steps 1 and 2 yields the inequalities specified in the proposition. This concludes the proof. ■

Proof of Proposition EC.5. We prove the two claims in turn.

Part 1. By Eq. 3 and Proposition 1, when $\mu_\theta = 1$, the equilibrium bonus in the $i = d$ can be written as the manager's posterior-weighted average of bonuses in $i = o$ corresponding to an ex-ante good versus ex-ante bad decision:

$$b_d(s_M, a, y) = \Pr(a = \overline{a} \mid s_M, a, y)\, b_o(\mathbf{s}, a = \overline{a}, y) + \Pr(a = \underline{a} \mid s_M, a, y)\, b_o(\mathbf{s}, a = \underline{a}, y), \tag{EC.62}$$

where $\Pr(a = \underline{a} \mid s_M, a, y) = 1 - \Pr(a = \overline{a} \mid s_M, a, y)$. Both posterior weights are in $(0, 1)$, so (EC.62) expresses $b_d(s_M, a, y)$ as a *strict convex combination* of the two observable bonuses.

By Proposition EC.3, the observable bonus is strictly higher following an ex-ante good decision than following an ex-ante bad decision:

$$b_o(\mathbf{s}, a = \underline{a}, y) < b_o(\mathbf{s}, a = \overline{a}, y).$$

Therefore, since $b_d(s_M, a, y)$ is a strict convex combination of these two values, it must lie strictly between them:

$$b_o(\mathbf{s}, a = \underline{a}, y) < b_d(s_M, a, y) < b_o(\mathbf{s}, a = \overline{a}, y).$$

Part 2. Next consider the expected observable-regime bonus conditional on (s_M, a, y):

$$\mathbb{E}_{s_H}[b_o(\mathbf{s}, a, y)].$$

In equilibrium, the manager knows that the DM's decision precision is lower in $i = d$ than in $i = o$ (Proposition EC.4). Hence, conditional on observing (s_M, a, y), the posterior probability that the action taken is ex-ante optimal is strictly lower in $i = d$ than in $i = o$. Since the equilibrium bonus is increasing in the posterior weight placed on the action being ex-ante optimal (cf. (EC.62)), it follows that the bonus in $i = d$ is strictly lower than the bonus in $i = o$ averaged over different realizations of s_H:

$$b_d(s_M, a, y) < \mathbb{E}_{s_H}[b_o(\mathbf{s}, a, y)].$$

This proves both parts. ■

Proof of Proposition EC.6. When the manager does not view the machine signal ($v = 0$), but has the same procedural fairness concerns as under $v = 1$, the equilibrium bonus is the expected (view) bonus taken with respect to the manager's posterior over s_M conditional on (a, y). Hence

$$b_{v=0}(a, y) = \Pr(s_M = a \mid a, y)\, b_{v=1}(s_M, a = s_M, y) + \Pr(s_M \neq a \mid a, y)\, b_{v=1}(s_M, a \neq s_M, y),$$

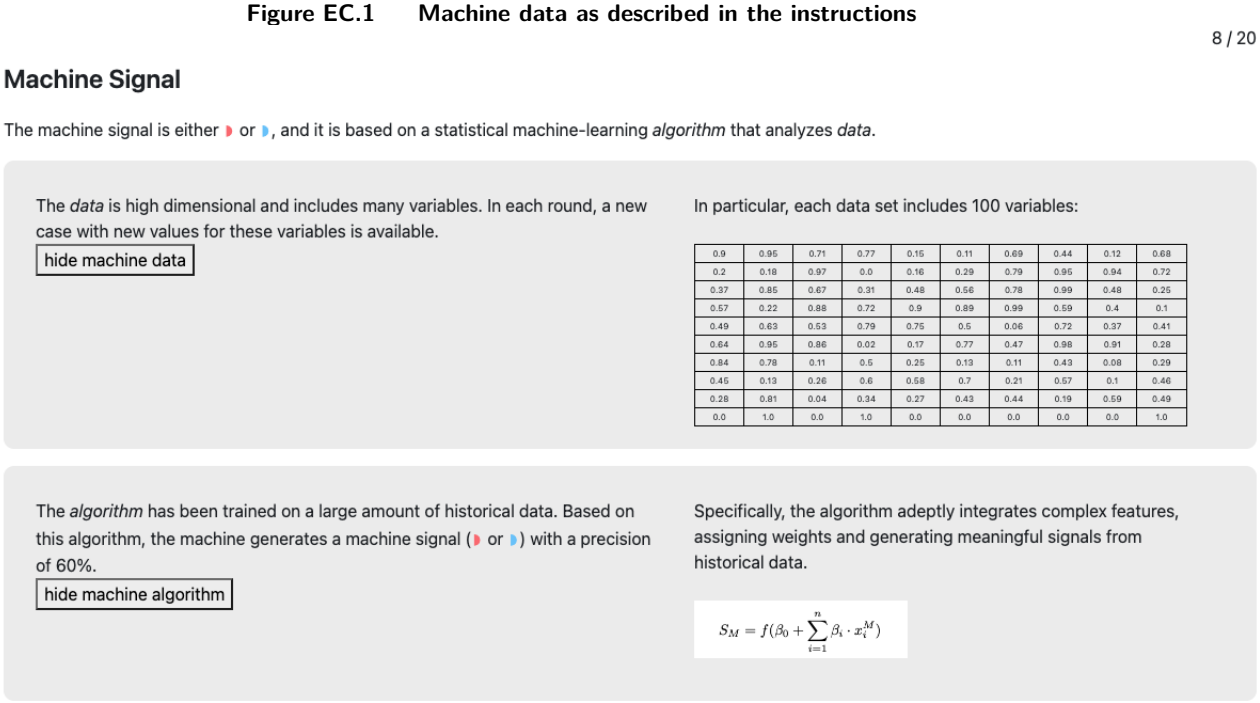
where $\Pr(s_M \neq a \mid a, y) = 1 - \Pr(s_M = a \mid a, y)$.

In equilibrium, the manager's posterior assigns positive probability to both events $s_M = a$ and $s_M \neq a$, so both weights lie in $(0, 1)$. Therefore $b_{v=0}(a, y)$ is a *strict convex combination* of the two view bonuses, implying

$$\min\{b_{v=1}(s_M, a \neq s_M, y),\, b_{v=1}(s_M, a = s_M, y)\} < b_{v=0}(a, y) < \max\{b_{v=1}(s_M, a \neq s_M, y),\, b_{v=1}(s_M, a = s_M, y)\}.$$

This holds for both $y = a$ and $y \neq a$. ■

Figure EC.1 Machine data as described in the instructions



B. Implementation

B.1. Session Planning

We provide an overview of all experimental sessions in Table EC.1. The table reports the number of sessions, participants, and treatment allocations, ensuring full transparency about the data collection process.

B.2. Data Generation

For each decision task, the true state of the world, y, and the human signal precision, θ, are first drawn independently. The state y is drawn from a Bernoulli distribution with parameter μ_0, while the precision type θ is drawn according to $\mu_\theta \equiv \Pr(\Theta = hi)$. Then, the signals s_H and s_M are generated independently from two Bernoulli distributions, each defined by precision parameters $q^\theta_{s_H}$ and q_{s_M} respectively.

Participants learn general details about the machine's prediction process from the instructions. We do *not* frame s_M explicitly as a decision *recommendation*, to avoid demand effects that could possibly arise simply from the term "recommendation". But participants are told that machine signals (either *red* or *blue*) are based on a statistical *algorithm* that analyzes *data*. They are further told that the *data* is high dimensional and included many variables, and that a new case with new values for these variables would be available in each round of the experiment.

Specifically, let X^M consist of n independent features, $X^M = (x_1^M, x_2^M, \ldots, x_n^M)$, with $n = 100$, so that the data are sufficiently complex for an algorithm to add substantive predictive value over a human with access to the same information. During the experiment instructions, participants can click a button to reveal example data underlying the machine signal X^M (see Figure EC.1). Participants are also told that the algorithm, trained on a large historical dataset, produces an imperfect binary signal $s_M = f(X^M) \in \{B, R\}$,

Table EC.1 Timeline of Experimental Sessions

Study	Treatment	Session	Date	Time	Participants	Exchange Rate*	Show Up Fee (€)
1	Delegation	1	2025-05-21	11:00	8	0.25	5
1	Delegation	2	2025-05-21	11:00	8	0.25	5
1	Delegation	3	2025-05-21	14:00	8	0.25	5
1	Delegation	4	2025-05-21	14:00	8	0.25	5
1	Delegation	5	2025-05-21	14:00	8	0.25	5
1	Delegation	6	2025-05-28	09:00	8	0.25	5
1	Delegation	7	2025-05-28	09:00	8	0.25	5
1	Delegation	8	2025-05-28	11:00	8	0.25	5
1	Delegation	9	2025-05-28	14:00	8	0.25	5
1	Delegation	10	2025-05-28	14:00	8	0.25	5
1	NoDelegation	1	2025-05-21	14:00	4	0.25	5
1	NoDelegation	2	2025-05-28	09:00	5	0.25	5
1	NoDelegation	3	2025-05-28	11:00	8	0.25	5
1	NoDelegation	4	2025-05-28	14:00	8	0.25	5
2	Delegation	1	2023-05-26	12:00	30	0.25	8
2	Delegation	2	2023-05-30	12:00	30	0.25	8
2	Delegation	3	2023-07-21	13:00	34	0.25	8
2	Delegation	4	2023-07-25	10:00	30	0.25	8
2	Delegation_3P	1	2023-06-16	12:00	20	0.25	4
2	Delegation_3P	2	2023-07-03	16:00	18	0.25	4
2	Delegation_3P	3	2023-07-20	10:00	16	0.25	4
2	DelegationObs	1	2023-07-21	10:00	32	0.25	8
2	DelegationObs	2	2023-07-25	13:00	28	0.25	8
2	NoDelegation	1	2023-06-02	16:00	8	0.125	8
2	NoDelegation	2	2023-06-14	16:00	16	0.125	8
2	NoDelegation	3	2023-06-23	16:00	14	0.125	8
2	NoDelegation	4	2023-07-26	10:00	24	0.125	8
3	Delegation	1	2025-05-20	09:00	8	0.50	5
3	Delegation	2	2025-05-20	09:00	8	0.50	5
3	Delegation	3	2025-05-20	09:00	8	0.50	5
3	Delegation	4	2025-05-20	11:00	8	0.50	5
3	Delegation	5	2025-05-20	11:00	8	0.50	5
3	Delegation	6	2025-05-20	11:00	8	0.50	5
3	Delegation	7	2025-05-20	14:00	8	0.50	5
3	Delegation	8	2025-05-20	14:00	8	0.50	5
3	Delegation	9	2025-05-20	14:00	8	0.50	5
3	Delegation	10	2025-05-21	09:00	8	0.50	5
3	Delegation	11	2025-05-21	09:00	8	0.50	5
3	NoDelegation	1	2025-05-20	09:00	4	0.25	5
3	NoDelegation	2	2025-05-20	11:00	8	0.25	5
3	NoDelegation	3	2025-05-20	14:00	7	0.25	5
3	NoDelegation	4	2025-05-21	09:00	4	0.25	5
3	NoDelegation	5	2025-05-21	11:00	5	0.25	5

Notes: *The exchange rate in *NoDelegation* is half of the *Delegation* when the source bonus is manager's own payoff.

with $s_M = f\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^M\right)$. We describe in the following in more detail a data-generating mechanism that produces unbiased signals with the desired precision q_{s_M}, and generates instances of data X_M and signals s_M that are consistent across the experiment.

The state of the world is a binary variable, $Y \in \{R, B\}$, which is upfront unknown with a prior distribution, $Pr(y = R) = 1 - Pr(y = B) = \mu_0$. A machine/algorithm has access to a set of information relevant for the state of the world, denoted by X^M. Given the information, the machine/algorithm generates an imperfect binary signal, $S_M = f(X^M) \in \{R, B\}$, such that $s_M = R$ (resp. $s_M = B$) is indicative of a red (resp. blue) state. The precision of the machine/algorithm is *known* and denoted by q_{s_M}, such that $Pr(s_M = R|y = R) = Pr(s_M = B|y = B) = q_{s_M} > 1/2$.

The machine generates the signal as follows:

$$S_M = f(\beta_0 + \sum_{i=1}^{100} \beta_i \cdot x_i^M) \in \{B, R\} \tag{EC.63}$$

This simple model can be modified so that the features represent meaningful identifiers, such as, e.g., in a medical domain, age, gender, underlying conditions, etc. Each feature is an i.i.d random variable. The first 90 features follow a uniform distribution $X_i^M \sim U(0,1), \forall i \in \{1, 2, \ldots, 90\}$, while the remaining 10 follow a Bernoulli distribution $X_i^M \sim \text{Bernoulli}(0.5), \forall i \in \{91, 92, \ldots, 100\}$. The signal is constructed as follows:

$$s_M = \begin{cases} y^\perp & \text{,if } \sum_{i=1}^{100} x_i^M > \Gamma \\ y & \text{,if } \sum_{i=1}^{100} x_i^M \leq \Gamma \end{cases} \tag{EC.64}$$

The threshold, Γ, is chosen such that $\mathbb{P}\left(\sum_{i=1}^{100} x_i^M \leq \Gamma\right) = \mathbb{P}(s_M = R|y = R) = \mathbb{P}(s_M = B|y = B) = q_{s_M}$, and $\mathbb{P}\left(\sum_{i=1}^{100} x_i^M > \Gamma\right) = \mathbb{P}(s_M = B|y = R) = \mathbb{P}(s_M = R|y = B) = 1 - q_{s_M}$. By this, we ensure that the machine generates $s_M = R$ (resp. $s_M = B$) with a probability corresponding to the chosen machine signal precision.

We know that the sum of 90 independent and identically distributed $U(0,1)$ random variables plus 10 independent and identically distributed Bernoulli(0.5) random variables, according to the Central Limit Theorem, is approximately a random variable following a normal distribution. Therefore, we choose a Γ that satisfies the following:

$$\mathbb{P}\left(\sum_{i=1}^{90} x_i^U + \sum_{i=1}^{10} x_i^B \leq \Gamma\right) \approx \mathbb{P}\left(N(50, \sqrt{10}) \leq \Gamma\right) = \Phi\left(\frac{\Gamma - 50}{\sqrt{10}}\right) = q_{s_M} \tag{EC.65}$$

Here, the sum of 90 $U(0,1)$ random variables and 10 Bernoulli(0.5) random variables is approximated by a normal distribution $N(50, \sqrt{10})$, where 50 is the mean and $\sqrt{10}$ is the standard deviation of the total sum, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. From this, we get:

$$\Gamma = 50 + \sqrt{10} \cdot \Phi^{-1}(q_{s_M}) \tag{EC.66}$$

Based on this, we randomly generated 1000 instances of the state variable (Y) following the prior μ_0, the Machine data X^M, and the corresponding Machine signal for each instance. We then also randomly generated a Human signal for each instance by the application of Bayes' law.

B.3. On-screen Participant Experience.

In this section, we provide screenshots of the experimental instructions and the main decision screens shown to participants across studies. These materials illustrate the on-screen experience and the key design features implemented in each treatment.

B.3.1. General Instructions. We first present the general instructions common across studies and treatments, followed by study- and treatment-specific materials.

Common Instructions

Screen: Prior about the True State.

True State and Signals

The true state is blue with *50%* probability, and green with *50%* probability.

Before making a decision, the decision maker has access to *informative* but *imperfect* signals that are either 🔵 or 🟢.

Screen: Machine Signal Introduction.

Machine Signal

The machine signal is either ◗ or ◗, and it is based on a statistical machine-learning *algorithm* that analyzes *data*.

The *data* is high dimensional and includes many variables. In each round, a new case with new values for these variables is available.

hide machine data

In particular, each data set includes 100 variables:

0.9	0.95	0.71	0.77	0.15	0.11	0.69	0.44	0.12	0.68
0.2	0.18	0.97	0.0	0.16	0.29	0.79	0.95	0.94	0.72
0.37	0.85	0.67	0.31	0.48	0.56	0.78	0.99	0.48	0.25
0.57	0.22	0.88	0.72	0.9	0.89	0.99	0.59	0.4	0.1
0.49	0.63	0.53	0.79	0.75	0.5	0.06	0.72	0.37	0.41
0.64	0.95	0.86	0.02	0.17	0.77	0.47	0.98	0.91	0.28
0.84	0.78	0.11	0.5	0.25	0.13	0.11	0.43	0.08	0.29
0.45	0.13	0.26	0.6	0.58	0.7	0.21	0.57	0.1	0.46
0.28	0.81	0.04	0.34	0.27	0.43	0.44	0.19	0.59	0.49
0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

The *algorithm* has been trained on a large amount of historical data. Based on this algorithm, the machine generates a machine signal (◗ or ◗).

hide machine algorithm

Specifically, the algorithm adeptly integrates complex features, assigning weights and generating meaningful signals from historical data.

$$S_M = f(\beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i^M)$$

Screen: Machine Signal Precision.

Machine Signal

The machine signal is either ◗ or ◗, and it is based on a statistical machine-learning *algorithm* that analyzes *data*.

The **machine signal has 70% precision**:

- If the true state is blue, the machine signal will more likely return as ◗ (**with 70%**) but it could falsely return as ◗ (**with 30%**)
- If the true state is green, the machine signal will more likely return as ◗ (**with 70%**) but it could falsely return as ◗ (**with 30%**)

Screen: Human Signal Precision.

Human Signal

The Human signal is either ◖ or ◖, and reflects information that is not in the data that the machine analyzes.

In each task, there is a **50%** chance that the Human Signal has **higher precision** than the Machine Signal, and a **50%** chance that it has **lower precision**.

When the Human Signal has **higher precision** than the Machine Signal, **the Human Signal has 80% precision**.

- If the true state is blue, the Human Signal will more likely return as ◖ (**with 80%**) but it could falsely return as ◖ (**with 20%**)
- If the true state is green, the Human Signal will more likely return as ◖ (**with 80%**) but it could falsely return as ◖ (**with 20%**)

When the Human Signal has **lower precision** than the Machine Signal, **the Human Signal has 60% precision**.

- If the true state is blue, the Human Signal will more likely return as ◖ (**with 60%**) but it could falsely return as ◖ (**with 40%**)
- If the true state is green, the Human Signal will more likely return as ◖ (**with 60%**) but it could falsely return as ◖ (**with 40%**)

Study-Specific Instructions

Screen: Study 1, Delegation.

18 / 21

Summary of events

In each period of the experiment:

True State

The computer randomly determines the true state:

blue (50% likely)
green (50% likely)

The state is not revealed yet

Machine Signal

Employee and manager know the precision of the Machine signal.

Employee observes the Machine signal (◗ or ◗).

At the end of the period, Manager has the option to observe the Machine signal (◗ or ◗) by clicking a button.

Human Signal

Employee observes the Human signal (◖ or ◖) and its precision.

Manager does NOT observe the Human signal (◖ or ◖), nor its precision.

Decision and Payoff

Employee chooses BLUE or GREEN

The true state (blue or green) is revealed

Manager receives the payoff from decision and true state

	blue	green
BLUE	$100	-$50
GREEN	-$50	$100

Bonus

Manager pays bonus to their employee and to a random other employee

Bonus is paid out of a separate budget of $60. Manager **does not** keep the remaining budget!

Total income

Manager: Payoffs from decision

Employee: bonus

Screen: Study 1, NoDelegation.

16 / 17

Summary of events

In each period of the experiment:

True State

The computer randomly determines the true state:

blue (50% likely)
green (50% likely)

The state is not revealed yet

Machine Signal

You observe the Machine signal (or).

Human Signal

You observe the Human signal (or) and its precision.

Decision and Payoff

You choose BLUE or GREEN

The true state (blue or green) is revealed

you receive the payoff from decision and true state

	blue	green
BLUE	$100	-$50
GREEN	-$50	$100

Income

You receive the the payoff from the decision and the true state as your income.

Screen: Study 2, Delegation.

17 / 20

Decisions are Delegated

In this experiment, you are working for a company.

Half of you will assume the role of the Employee, the other half will assume the role of the Manager.

The Manager has delegated the decision (RED or BLUE) to the Employee.

In each period of the experiment:

The computer randomly determines the true state:

red (50% likely)
blue (50% likely)

The state is not revealed yet

Both employee and manager observe the machine signal (or)

Only the employee observes the human signal signal (or)

Employee chooses RED or BLUE

The true state (red or blue) is revealed

Manager receives the payoffs from decision and true state

	red	blue
RED	$100	$30
BLUE	$30	$100

Manager pays bonus to employee

Total income

Manager: Payoffs from decision **MINUS** bonus

Employee: bonus

Screen: Study 2, DelegationObs.

17 / 20

Decisions are Delegated

In this experiment, you are working for a company.

Half of you will assume the role of the Employee, the other half will assume the role of the Manager.

The Manager has delegated the decision (RED or BLUE) to the Employee.

In each period of the experiment:

The computer randomly determines the true state:

red (50% likely)
blue (50% likely)

The state is not revealed yet

Both employee and manager observe the machine signal (or)

Both employee and manager observe the human signal (or)

Employee chooses RED or BLUE

The true state (red or blue) is revealed

Manager receives the payoffs from decision and true state

	red	blue
RED	$100	$30
BLUE	$30	$100

Manager pays bonus to employee

Total income

Manager: Payoffs from decision **MINUS** bonus

Employee: bonus

Screen: Study 2, Delegation_3P.

17 / 20

Decisions are Delegated

In this experiment, you are working for a company.

Half of you will assume the role of the Employee, the other half will assume the role of the Manager.

The Manager has delegated the decision (RED or BLUE) to the Employee.

The manager controls a $50 budget which they can assign as a payment to their employee and a random other person in the room (an employee from another group). The manager does not keep for themselves any money from the $50 budget that they do not pay out to their employee and the random other person.

In each period of the experiment:

The computer randomly determines the true state:

red (50% likely)
blue (50% likely)

The state is not revealed yet

Both employee and manager observe the machine signal (or)

Only the employee observes the human signal signal (or)

Employee chooses RED or BLUE

The true state (red or blue) is revealed

Manager receives the payoffs from decision and true state

	red	blue
RED	$100	$30
BLUE	$30	$100

Manager pays bonus to their employee and to a random other employee

Bonus is paid out of a separate budget of $50. Manager **does not** keep the remaining budget!

Total income

Manager: Payoffs from decision

Employee: bonus

Screen: Study 2, NoDelegation

16 / 17

Sequence

In each period of the experiment:

The computer randomly determines the true state:

red (50% likely)
blue (50% likely)

The state is not revealed yet

You observe the machine signal (or) and the human signal (or)

You choose RED or BLUE

The true state (red or blue) is revealed

You receive the payoffs from decision and true state

	red	blue
RED	$100	$30
BLUE	$30	$100

Total income

Payoffs from decision

Screen: Study 3, Delegation

18 /

Summary of events

In each period of the experiment, there are 5 decision tasks.

In each task:

True State

The computer randomly determines the true state:

blue (50% likely)
green (50% likely)

The state is not revealed yet

Machine Signal

Employee and manager know the precision of the Machine signal.

Employee observes the Machine signal (or).

At the end of the period, Manager has the option to observe the Machine signal (or) by clicking a button.

Human Signal

Employee observes the Human signal (or) and its precision.

Manager does NOT observe the Human signal (or), nor its precision.

Decision and Payoff

Employee chooses BLUE or GREEN

The true state (blue or green) is revealed

The payoffs from decision and true state are calculated

	blue	green
BLUE	$100	-$50
GREEN	-$50	$100

...

At the end of the period:

Total Payoff and Bonus

Manager receives the average of the payoffs from all 5 tasks

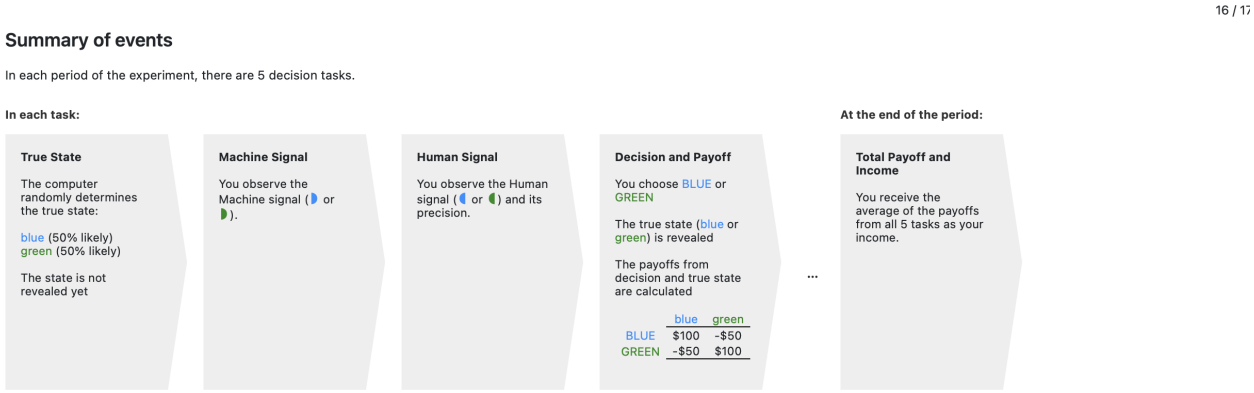
Manager pays bonus to employee

Total income

Manager: Average of the payoffs from all tasks **MINUS** bonus

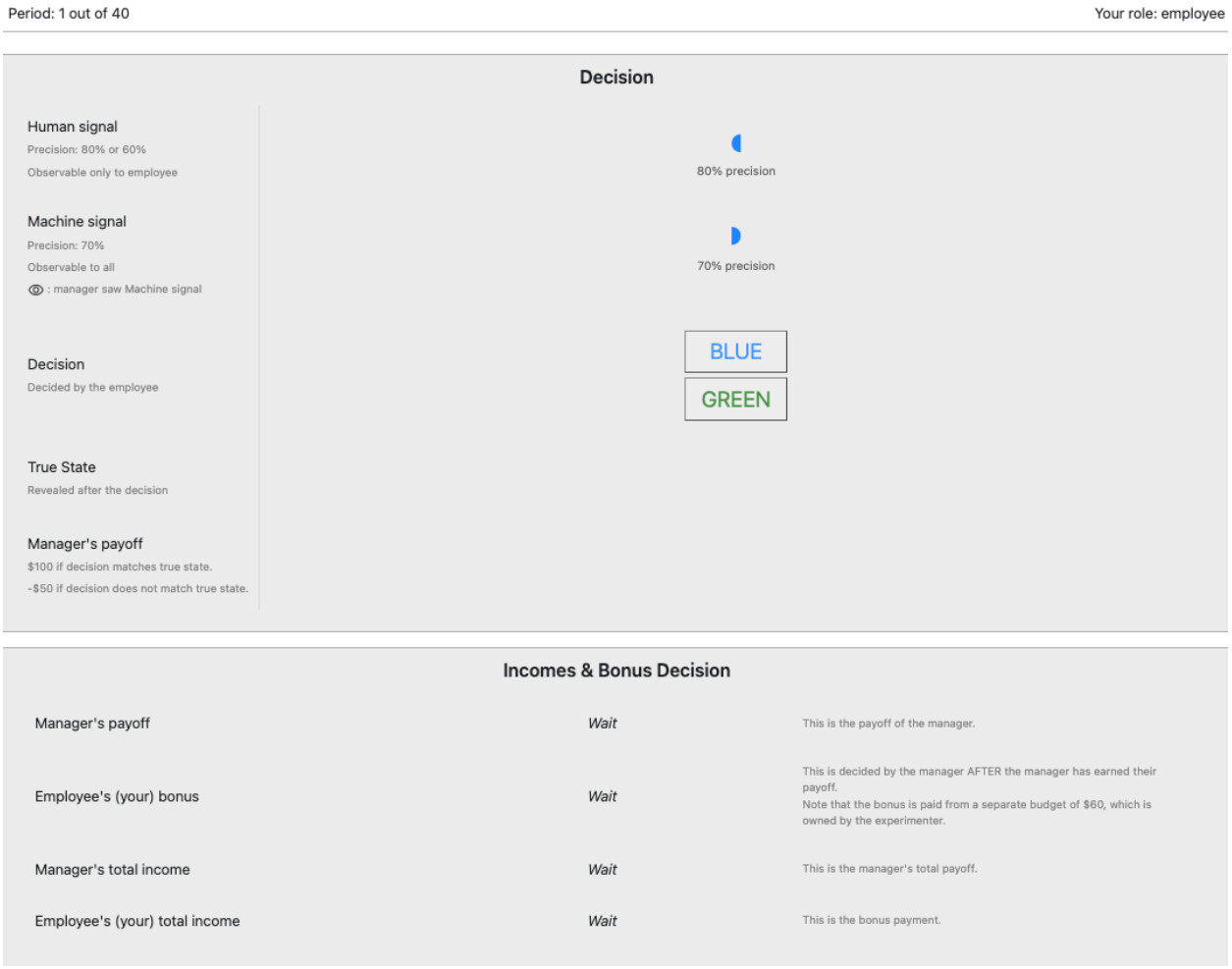
Employee: bonus

Screen: Study 3, NoDelegation

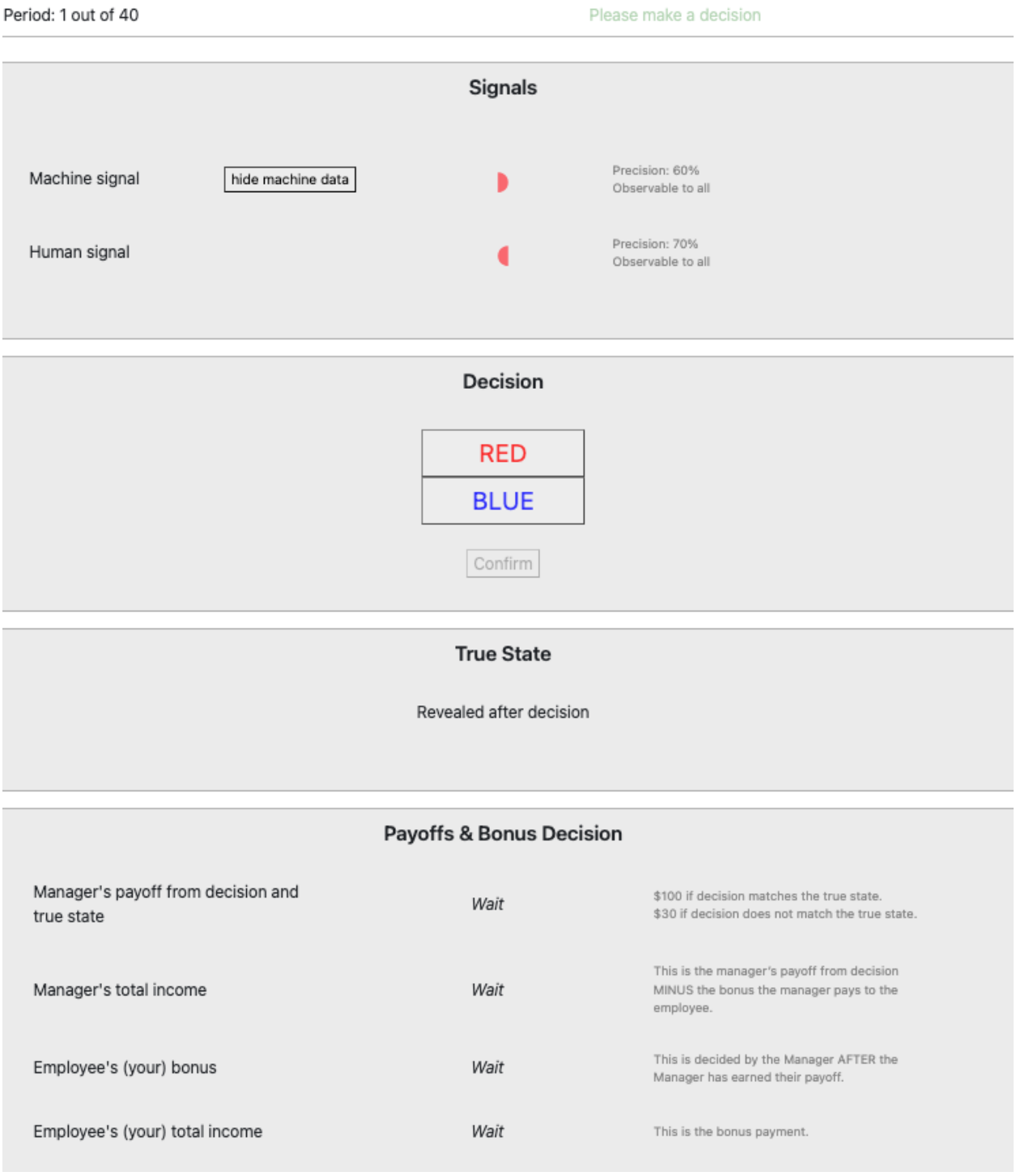


B.3.2. Employee's (DM's) Game Screen. The game interface of participants in the role of a DM is shown below, for each study.

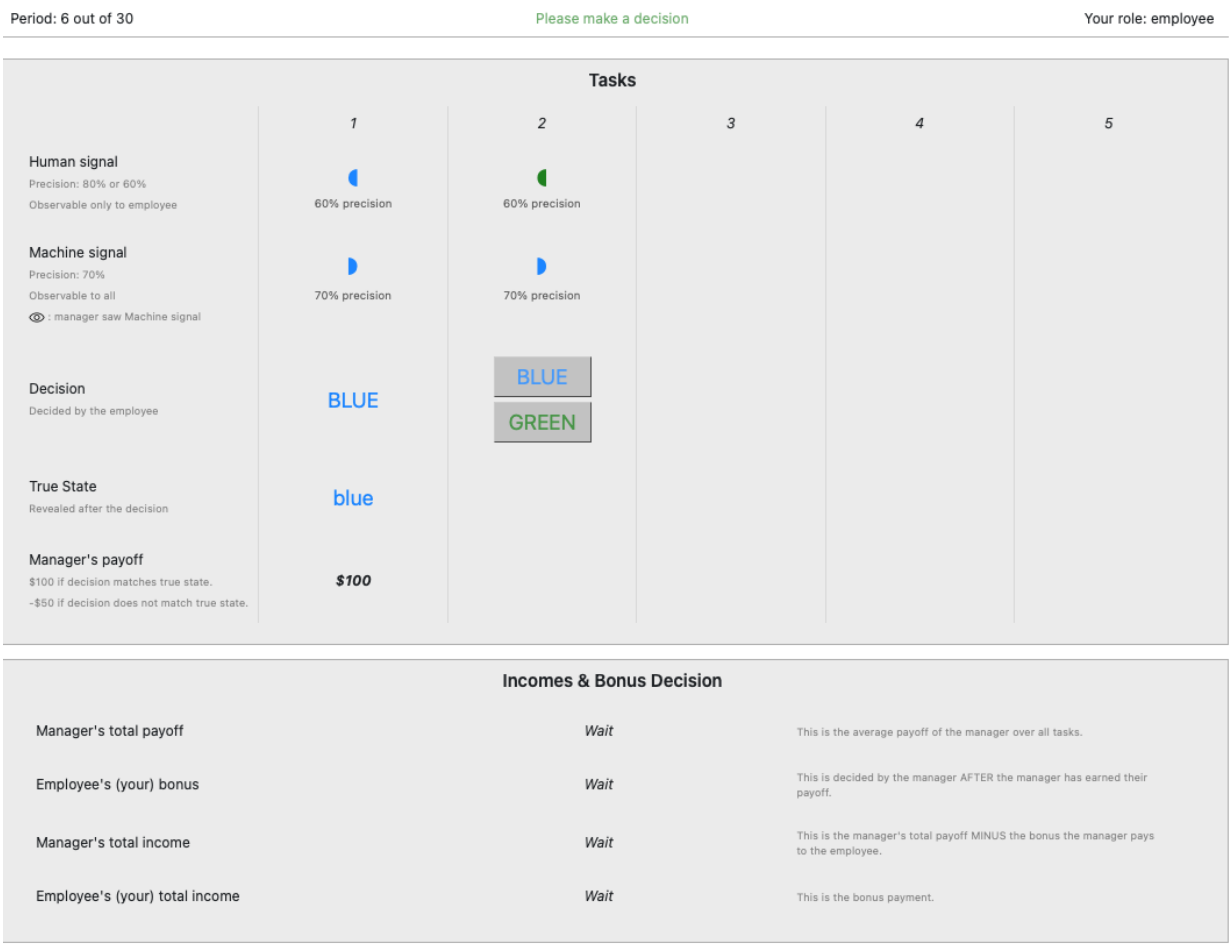
Screen: Study 1, Delegation.



Screen: Study 2, Delegation.

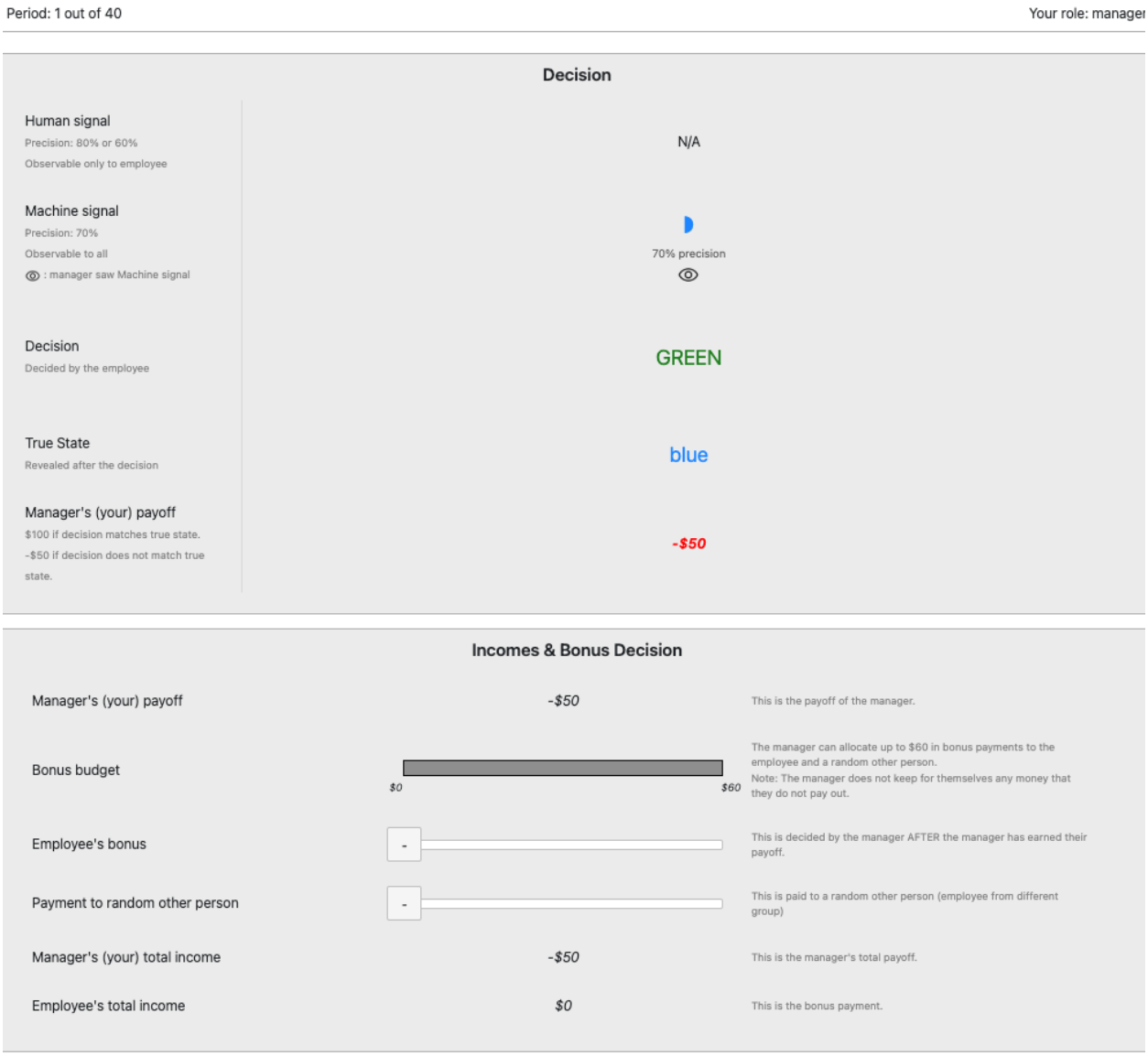


Screen: Study 3, Delegation.



B.3.3. Manager's Game Screen. The game interface of participants in the role of a manager is shown below, for each study.

Screen: Study 1, Delegation.



Screen: Study 2, Delegation.

Period: 1 out of 40

Signals

Machine signal	hide machine data		Precision: 60% Observable to all
Human signal		N/A	Precision: 70% Observable only to employee

Decision

BLUE

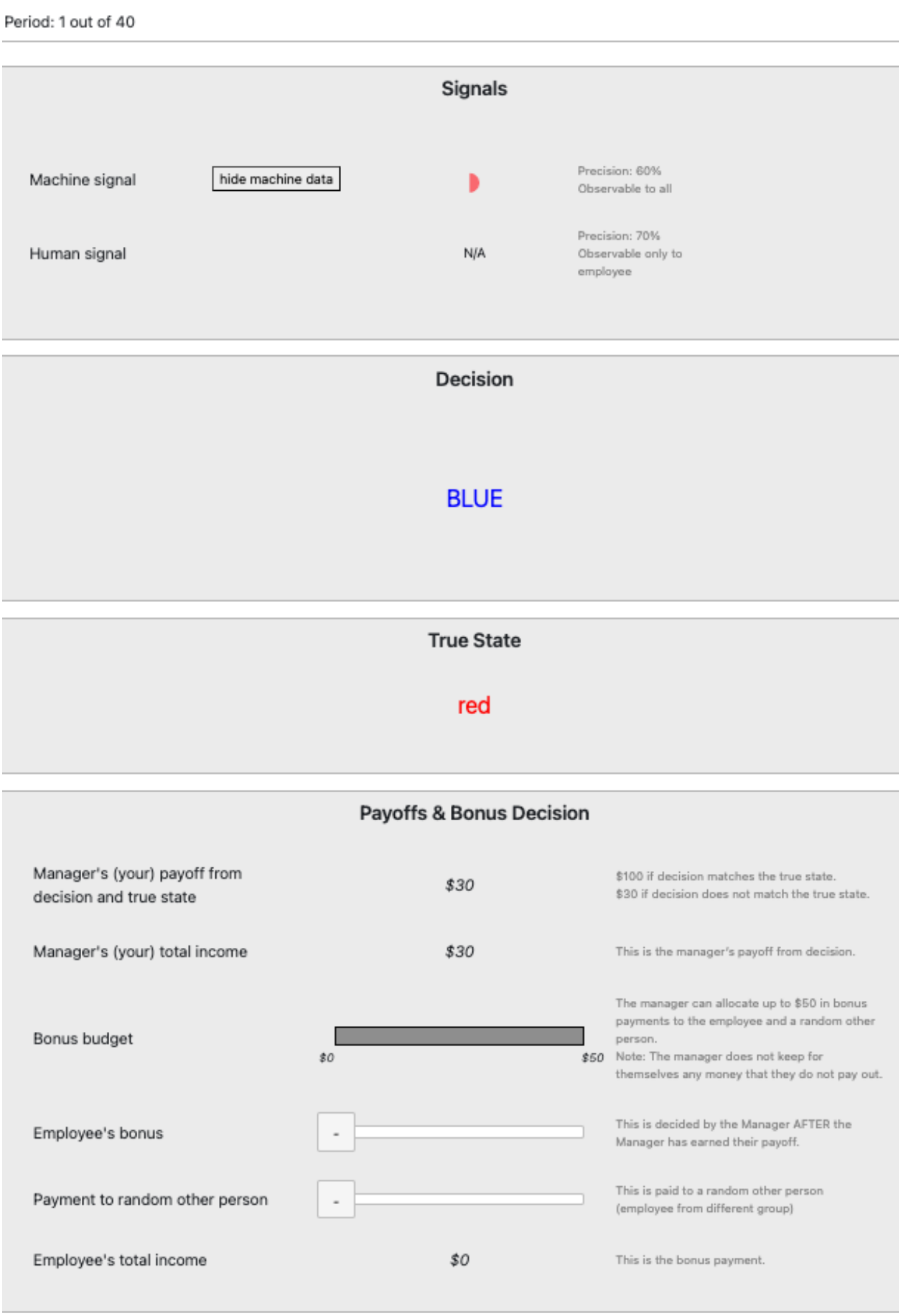
True State

red

Payoffs & Bonus Decision

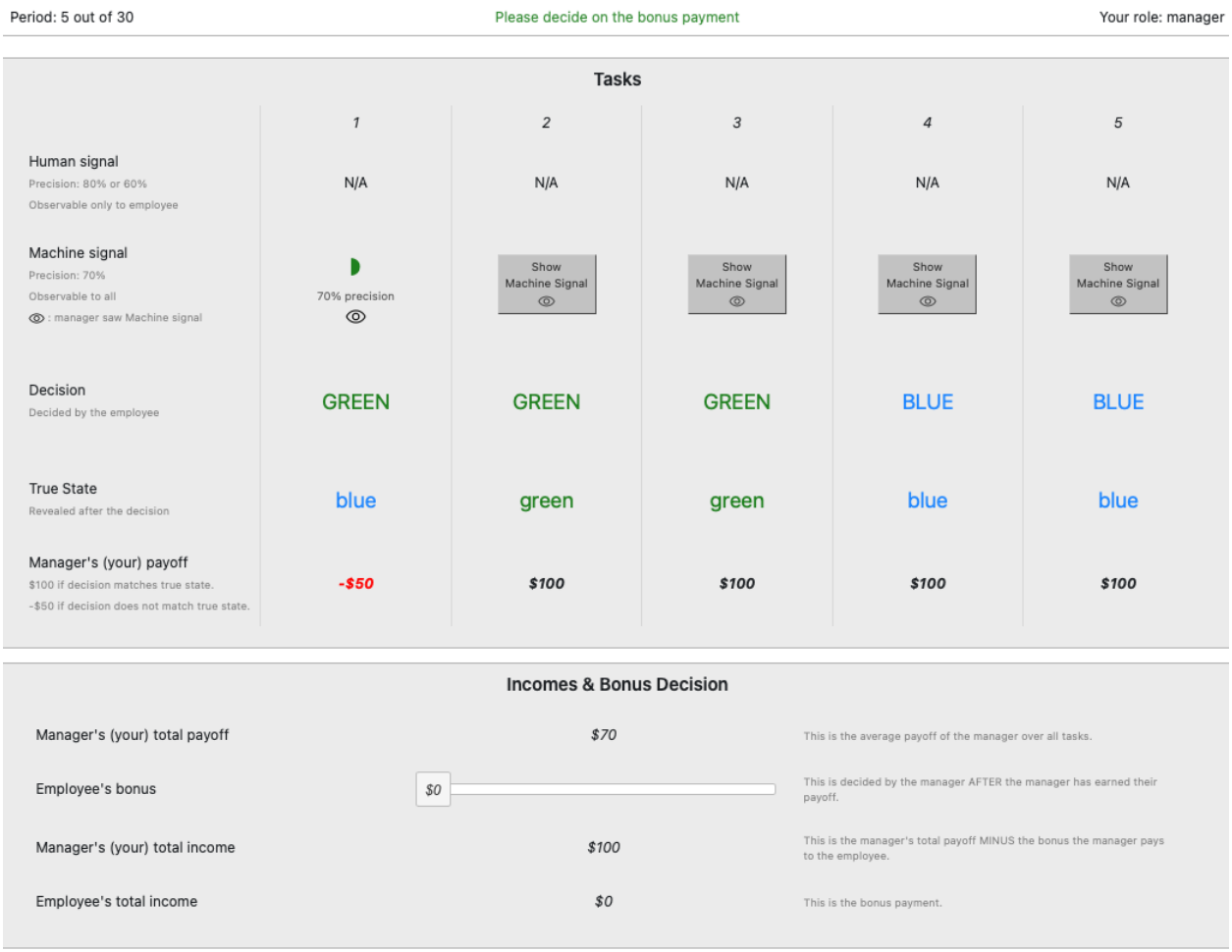
Manager's (your) payoff from decision and true state		*$30*	$100 if decision matches the true state. $30 if decision does not match the true state.
Manager's (your) total income		*$30*	This is the manager's payoff from decision MINUS the bonus the manager pays to the employee.
Employee's bonus	-		This is decided by the Manager AFTER the Manager has earned their payoff.
Employee's total income		*$0*	This is the bonus payment.

Screen: Study 2, Delegation_3P.



Screen: Study 2, DelegationObs.

Screen: Study 3, Delegation.



C. Supplementary Statistical Analyses

This section presents additional analyses that complement and further substantiate the main findings across the studies. Unless otherwise specified, standard errors are two-way clustered at the participant and session levels.[28] Standard errors are reported in parentheses. We report exact p-values in brackets and, for consistency with the main text, indicate conventional significance levels: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

C.1. Study 1

C.1.1. DM: Algorithm Adherence and Delegation. To directly test if delegation increases algorithm reliance when signals conflict, we fit a simple model, $FollowM = \beta_0 + \beta_1 Delegation + \beta_2 Round$. For $Conflict = 1$, the treatment effect of $Delegation$ is $\beta_1 = 0.123$ ($p = 0.01$).

Table EC.2 Effect of Delegation on Following the Machine ($Conflict = 1$, Study 1)

DV: $FollowM$		Estimate
β_0	*Constant*	0.507^{***} (0.025) [0.000]
β_1	*Delegation*	0.123^{**} (0.042) [0.011]
β_2	*Round*	-0.0016 (0.0021) [0.457]
Observations		1,114
Participants (clusters)		65
Sessions (clusters)		14

C.1.2. Managers: Bonus. A potential concern when analyzing managers' bonus payments conditional on viewing the machine signal is endogeneity, since both $Bonus$ and $ViewM$ are manager choice variables. Unobserved factors that affect a manager's propensity to view may also influence bonus payment behavior, leading to biased estimates. Ideally, this concern could be addressed using an instrumental variable that affects viewing behavior but does not directly influence bonus decisions. However, no suitable instrument is available in our setting. We therefore address this issue by adopting a correlated random-effects (CRE) specification following Mundlak (1978) and Wooldridge (2019). This approach relaxes the strict assumption of standard random-effects models that individual-specific effects are uncorrelated with the observed regressors by explicitly modeling their correlation through the inclusion of group means (Mundlak terms). The joint statistical significance of these terms provides a specification test for the presence of unobserved heterogeneity. This specification identifies within-manager effects while controlling for time-invariant heterogeneity that may jointly affect viewing and bonus decisions, and it allows the estimation of coefficients on time-invariant variables, which is not feasible in a standard fixed-effects framework. Formally, let i index managers and t rounds. The bonus assigned by manager i in round t is

$$\begin{aligned}\text{Bonus}_{it} = {} & \beta_0 + \beta_1 \text{BadOutcome}_{it} + \beta_2 \text{ViewedFollow}_{it} + \beta_3 \text{ViewedOverride}_{it} \\ & + \beta_4(\text{ViewedFollow}_{it} \times \text{BadOutcome}_{it}) + \beta_5(\text{ViewedOverride}_{it} \times \text{BadOutcome}_{it}) \\ & + \beta_6 \text{Round}_{it} + \gamma_1 \overline{\text{View}}_i + u_i + \varepsilon_{it},\end{aligned} \tag{EC.67}$$

[28] In specifications with too few session clusters, we cluster standard errors at the participant level only.

where u_i is the manager-specific random intercept capturing unobserved, time-invariant heterogeneity, and $\overline{\text{View}}_i$ is the manager's average viewing frequency (the Mundlak term). The variables ViewedFollow_{it} and $\text{ViewedOverride}_{it}$ indicate rounds in which manager i viewed and observed that the decision followed or overrode the machine signal, respectively. The parameter γ_1 represents the *between-manager* effect associated with managers' average propensity to view the machine signal.

Table EC.3 Bonus Payments (CRE – Mundlak Specification)

DV:		$Bonus^{DM}$	$Bonus^{3p}$
β_0	Constant	40.789***	9.115***
		(4.150) [0.000]	(2.885) [0.002]
β_1	$BadOutcome$	-23.754***	17.803***
		(1.142) [0.000]	(1.057) [0.000]
β_2	$ViewedFollow$	-0.659	-0.144
		(0.999) [0.509]	(0.925) [0.877]
β_3	$ViewedOverride$	3.027**	-2.158*
		(1.414) [0.032]	(1.309) [0.099]
β_4	$ViewedFollow \times BadOutcome$	4.118***	-4.731***
		(1.426) [0.004]	(1.320) [0.000]
β_5	$ViewedOverride \times BadOutcome$	-5.270***	3.163*
		(2.014) [0.009]	(1.864) [0.090]
β_6	$Round$	0.029	-0.000
		(0.024) [0.230]	(0.023) [0.996]
γ_1	$\overline{View}$	11.221**	-2.194
		(5.550) [0.043]	(3.889) [0.573]
$\beta_3 - \beta_2$		3.686***	-2.014*
		(1.222) [0.003]	(1.132) [0.075]
$\beta_2 + \beta_4$		3.459**	-4.874***
		(1.414) [0.014]	(1.309) [0.000]
$\beta_3 + \beta_5$		-2.243	1.005
		(1.674) [0.180]	(1.549) [0.516]
$(\beta_3 + \beta_5) - (\beta_2 + \beta_4)$		-5.701***	5.880***
		(1.389) [0.000]	(1.286) [0.000]
Observations		1,600	1,600
Managers		40	40
Random-intercept SD		12.03	8.27

Table EC.3 reports the CRE estimates with $Bonus^{DM}$ and $Bonus^{3p}$ as dependent variables. The main findings remain robust. Managers pay their DMs significantly higher bonuses after a good outcome than after a bad outcome ($\beta_1 = -23.754$, $p < 0.01$), while paying the third party significantly more after a bad outcome ($\beta_1 = 17.803$, $p < 0.01$). Conditional on a bad outcome, managers reduce the DM's bonus when they view that the DM overrode the machine, relative to when the machine was followed ($(\beta_3 + \beta_5) - (\beta_2 + \beta_4) = -5.701$, $p < 0.01$). Symmetrically, they increase the third party's bonus in this case ($(\beta_3 + \beta_5) - (\beta_2 + \beta_4) = -5.880$, $p < 0.01$).

Finally, the estimated between-manager effect provides evidence of systematic manager heterogeneity: managers with a higher average propensity to view the machine signal tend to pay higher bonuses to their DMs ($\gamma_1 = 11.221$, $p = 0.04$), whereas no corresponding relationship emerges for bonuses paid to third parties ($\gamma_1 = -2.194$, $p = 0.573$). Thus, managers who view less frequently pay lower bonuses specifically to their own DMs, rather than lower bonuses in general. This selective pattern is consistent with the moral-wiggle-room interpretation discussed in the main text: remaining uninformed may provide managers with greater

latitude to justify lower bonuses to their own DMs. Turning to within-manager effects, when the outcome is good, managers pay their DM a higher bonus when they view that the DM overrode the machine signal than when they do not view the machine signal ($\beta_3 = 3.027$, $p = 0.03$). In contrast, bonuses are not significantly different when they view that the DM followed the machine ($\beta_2 = -0.659$, $p = 0.51$). When the outcome is bad, managers pay their DM a higher bonus when they view that the DM followed the machine signal than when they do not view the machine signal ($\beta_2 + \beta_4 = 3.459$, $p = 0.01$). However, bonuses are not significantly different when they view that the DM overrode the machine ($\beta_3 + \beta_5 = -2.243$, $p = 0.18$).

C.1.3. Managers: Mechanism Evidence (from $ViewM$ data). We fit a linear probability model, $ViewM = \beta_1 BadOutcome + \beta_2 Round + \beta_3(BadOutcome \times Round)$ (Table EC.4), and find that managers are on average more likely to view the machine signal after a bad outcome ($\beta_1 = 0.067, p = 0.12$). Further, while the tendency to view decreases after good outcomes ($\beta_2 = -0.005, p = 0.04$), it remains at a high level after bad outcomes ($\beta_2 + \beta_3 = 0.002, p = 0.62$).

Table EC.4 Managers' Viewing Behavior over Time and Outcome Type.

DV: $ViewM$		Estimate
β_0	$Constant$	0.6474*** (0.0571) [0.000]
β_1	$BadOutcome$	0.0665 (0.0391) [0.123]
β_2	$Round$	-0.0053** (0.0022) [0.041]
β_3	$BadOutcome \times Round$	0.0068 (0.0040) [0.127]
$\beta_2 + \beta_3$		0.0015 (0.0029) [0.622]
Observations		1,600
Participants (clusters)		40
Sessions (clusters)		10

C.1.4. Does Algorithm Adherence Pay off? To assess whether DMs receive higher bonuses for overriding the machine signal when the human signal is more precise and in conflict with the machine ($\theta = hi$ and $Conflict = 1$), we estimate the regression $BonusReceive_{it} = \beta_0 + \beta_1 FollowM_{it} + \varepsilon_{it}$, with standard errors two-way clustered at the DM and session levels (Table EC.5). We find that the bonus for following the machine signal is lower than for overriding ($\beta_1 = -9.240$, $p = 0.09$).

Table EC.5 Effect of Following the Machine on DM Bonuses ($\theta = hi$ and $Conflict = 1$).

DV: $BonusReceive$		Estimate
β_0	$Constant$	40.664*** (1.799) [0.000]
β_1	$FollowM$	-9.240* (4.871) [0.090]
Observations		292
Participants (clusters)		40
Sessions (clusters)		10

C.2. Study 2

C.2.1. DM: Algorithm Adherence and Delegation To examine whether learning attenuates the higher adherence induced by delegation, we estimate models allowing for treatment-specific time trends. The results (Table EC.6) show no significant time trend in *DelegationObs* ($\beta_3 + \beta_4 = -0.001$, $p = 0.17$ when $Conflict = 1$, and $\beta_3 + \beta_4 = -0.001$, $p = 0.11$ when $Conflict = 0$) or in *NoDelegation* ($\beta_3 = -0.002$, $p = 0.39$ when $Conflict = 1$, and $\beta_3 = 0.001$, $p = 0.20$ when $Conflict = 0$). In contrast, algorithmic adherence increases over time in *Delegation* ($\beta_3 + \beta_5 = -0.004$, $p < 0.01$ when $Conflict = 1$, and $\beta_3 + \beta_5 = -0.002$, $p = 0.01$ when $Conflict = 0$). This pattern suggests that the observed behavior does not reflect a bias that DMs learn to correct with experience, but rather a sustained response to bonus incentives that reward adherence to the algorithm.

Table EC.6 Round Trends under Delegation and Observed Delegation.

DV: *OptDecision*		$Conflict = 1$	$Conflict = 0$
β_0	Constant	0.8436***	0.9563***
		(0.0134) [0.000]	(0.0149) [0.000]
β_1	*DelegationObs*	0.0246	0.0092
		(0.0487) [0.626]	(0.0305) [0.769]
β_2	*Delegation*	-0.1498***	-0.0661
		(0.0363) [0.003]	(0.0452) [0.177]
β_3	Round	-0.0012	0.0007
		(0.0013) [0.388]	(0.0005) [0.202]
β_4	*DelegationObs* × Round	0.0005	-0.0013*
		(0.0014) [0.724]	(0.0006) [0.069]
β_5	*Delegation* × Round	-0.0023	-0.0022**
		(0.0016) [0.191]	(0.0007) [0.013]
$\beta_3 + \beta_5$		-0.0035***	-0.0015**
		(0.0010) [0.006]	(0.0005) [0.013]
$\beta_3 + \beta_4$		-0.0007	-0.0005
		(0.0005) [0.166]	(0.0003) [0.108]
Observations		2,862	3,298
Participants (clusters)		154	154
Sessions (clusters)		10	10

C.2.2. Does Algorithm Adherence Pay Off? To assess whether DMs receive higher bonuses for overriding the machine signal in conflict, we estimate the regression $BonusReceive_{it} = \beta_0 + \beta_1 \cdot DelegationObs_{it} + \beta_2 \cdot FollowM_{it} + \beta_3 \cdot DelegationObs_{it} \cdot FollowM_{it} + \beta_4 \cdot Round + \epsilon_{it}$, with standard errors clustered at the DM and session levels (Table EC.7). We find that when signals conflict, incorrectly following the less precise machine signals reduces the DM's average bonus ($\beta_2 = -3.905$, $p < 0.01$). Importantly, *DelegationObs* reduces DM's average bonus for incorrect machine-adherence more strongly ($\beta_2 + \beta_3 = -7.444$, $p < 0.01$) because it eliminates the key driver for algorithm adherence in *Delegation*: managers, when observing a bad outcome, blame the DM for (correctly!) overriding the machine signal.

Table EC.7 Effect of Following the Machine on DM Bonuses.

DV: *BonusReceive*		$Conflict = 1$	$Conflict = 0$
β_0	Constant	14.1208*** (2.7765) [0.004]	8.3643*** (1.7698) [0.005]
β_1	*DelegationObs*	7.5861* (3.6215) [0.090]	-3.4170 (1.7294) [0.105]
β_2	*FollowM*	-3.9047*** (0.9241) [0.008]	9.1802*** (1.9907) [0.006]
β_3	*DelegationObs* × *FollowM*	-3.5392* (1.3861) [0.051]	9.4687** (3.6587) [0.049]
β_4	Round	-0.0576 (0.0297) [0.110]	-0.0893** (0.0250) [0.016]
$\beta_2 + \beta_3$		-7.4439*** (1.0734) [0.001]	18.6490*** (3.0902) [0.002]
Observations		1,706	1,974
Participants (clusters)		92	92
Sessions (clusters)		6	6

C.3. Study 3

C.3.1. Manager: Bonus Decisions. As discussed in Section C.1.2, we address potential endogeneity between *Bonus* and $ViewM$ using a correlated random-effects (CRE) specification following Mundlak (1978) and Wooldridge (2019), which identifies within-manager effects while controlling for time-invariant heterogeneity.

Particularly, to examine whether managers reward DMs differently when the DM *overrides* the machine recommendation rather than *follows* it, we aggregate each round into four mutually exclusive counts observed by the manager after viewing:

- $ViewedFollowGood$ denotes the number of tasks in a round for which $ViewM = 1$, $FollowM = 1$, and $BadOutcome = 0$.
- $ViewedOverrideGood$ denotes the number of tasks in a round for which $ViewM = 1$, $FollowM = 0$, and $BadOutcome = 0$.
- $ViewedFollowBad$ denotes the number of tasks in a round for which $ViewM = 1$, $FollowM = 1$, and $BadOutcome = 1$.
- $ViewedOverrideBad$ denotes the number of tasks in a round for which $ViewM = 1$, $FollowM = 0$, and $BadOutcome = 1$.

Each variable captures the number of tasks within a round that fall into the corresponding outcome–behavior category. We also include the total number of bad outcomes in the round, *BadTotal*, to control for the manager's realized payoff. We estimate the following correlated random-effects (Mundlak) model:

$$Bonus_{ir} = \beta_0 + \beta_1\, BadTotal_{ir} + \sum_k \beta_k\, X_{kir} + \beta_6\, Round_{ir} + \gamma_1 \overline{View}_i + \mu_i + \varepsilon_{ir},$$

where i indexes managers and r rounds. The vector X_{kir} contains the four *good/bad* × *follow/override* counts for round r, and $\overline{View}_i$ is the manager's average viewing frequency per round (the Mundlak term). The inclusion of $\overline{View}_i$ implements the Mundlak correction, allowing the random intercept μ_i to be correlated

with the regressors. The coefficients β_k capture within-manager effects (how deviations from a manager's usual pattern affect bonuses), whereas the γ_1 term captures between-manager differences.

We test whether managers reward or penalize DMs differently depending on whether the DM followed or overrode the machine, conditional on the realized outcome of the task:

$$H_0^{\text{good}} : \beta_3 = \beta_2,$$
$$H_0^{\text{bad}} : \beta_5 = \beta_4.$$

Rejecting these null hypotheses indicates that managers systematically differentiate between overrides and follows once outcome and overall performance are held constant.

When the number of bad outcomes in a round exceeds three ($BadTotal \in \{4, 5\}$), the manager's payoff becomes negative. Since bonuses are financed from this payoff, bonus payments are mechanically zero in these cases. Accordingly, the following analysis focuses on rounds with $BadTotal \leq 3$, where the payoff is positive and bonuses can be paid.

Table EC.8 reports the regression results for pooled data and when we stratify the data by the number of bad task outcomes.

Table EC.8 Bonus Level: Follow vs. Override by Realized Bad Outcomes

DV: $Bonus^{DM}$		(1)	(2)	(3)	(4)
		Pooled CRE	$BadTotal = 1$ CRE	$BadTotal = 2$ CRE	$BadTotal = 3$ CRE
β_0	*Constant*	18.659*** (2.235) [0.000]	11.704*** (2.455) [0.000]	5.787*** (1.405) [0.000]	1.347*** (0.400) [0.001]
β_1	*BadTotal*	-6.378*** (0.301) [0.000]	–	–	–
β_2	*ViewedFollowGood*	0.879*** (0.186) [0.000]	-0.244 (0.300) [0.416]	0.088 (0.444) [0.844]	-0.135 (0.271) [0.619]
β_3	*ViewedOverrideGood*	1.236*** (0.428) [0.004]	0.092 (0.531) [0.862]	-1.078 (0.754) [0.152]	0.087 (0.370) [0.814]
β_4	*ViewedFollowBad*	-1.633*** (0.393) [0.000]	1.227 (1.084) [0.258]	1.116* (0.653) [0.087]	0.244 (0.185) [0.187]
β_5	*ViewedOverrideBad*	-2.107*** (0.490) [0.000]	0.716 (1.215) [0.556]	-0.052 (0.774) [0.946]	-0.252 (0.206) [0.220]
β_6	*Round*	-0.134*** (0.025) [0.000]	-0.140*** (0.028) [0.000]	-0.092*** (0.033) [0.005]	-0.013 (0.014) [0.362]
γ_1	$\overline{View}$	1.435* (0.839) [0.087]	1.830* (0.939) [0.051]	0.681 (0.553) [0.218]	0.027 (0.161) [0.868]
$\beta_3 - \beta_2$		0.357 (0.470) [0.447]	0.336 (0.529) [0.526]	-1.166 (0.719) [0.105]	0.222 (0.332) [0.504]
$\beta_5 - \beta_4$		-0.474 (0.543) [0.383]	-0.510 (0.816) [0.532]	-1.168** (0.557) [0.036]	-0.496*** (0.186) [0.008]
Observations		1,246	499	324	148
Managers		44	44	44	43
Random-intercept SD		9.16	10.16	5.59	1.44

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We find that (Table EC.8), each additional bad task reduces the bonus (Column (1), $\beta_1 = -6.378$, $p < 0.01$). More importantly, for each bad task outcome, managers lower the bonus payments more after they observe an unsuccessful machine override, when the round overall resulted in one bad outcome (Column (2), $\beta_5 - \beta_4 = -0.510$, $p = 0.53$), two bad outcomes (Column (3), $\beta_5 - \beta_4 = -1.168$, $p = 0.04$), and three bad outcomes (Column (4), $\beta_5 - \beta_4 = -0.496$, $p < 0.01$).

Differences in absolute payoff size can mechanically cap how much a manager can pay as bonus; expressing the payment as a *fraction* of the round's budget (manager's payoff) removes that mechanical link. We therefore, as a robustness check, re-estimate the model with

$$BonusShare_{ir} = \frac{Bonus_{ir}^{DM}}{Payoff_{ir}} \in [0,1]$$

as the dependent variable. The regression results are reported in Table EC.9.

Table EC.9 Bonus Share: Follow vs. Override by Realized Bad Outcomes

DV: *BonusShare*		(1)	(2)	(3)	(4)
		Pooled CRE	$BadTotal = 1$ CRE	$BadTotal = 2$ CRE	$BadTotal = 3$ CRE
β_0	*Constant*	0.192*** (0.035) [0.000]	0.167*** (0.035) [0.000]	0.145*** (0.035) [0.000]	0.135*** (0.040) [0.001]
β_1	*BadTotal*	-0.023*** (0.004) [0.000]	–	–	–
β_2	*ViewedFollowGood*	0.008*** (0.003) [0.001]	-0.003 (0.004) [0.416]	0.002 (0.011) [0.844]	-0.013 (0.027) [0.619]
β_3	*ViewedOverrideGood*	0.005 (0.006) [0.356]	0.001 (0.008) [0.862]	-0.027 (0.019) [0.152]	0.009 (0.037) [0.814]
β_4	*ViewedFollowBad*	0.004 (0.005) [0.466]	0.018 (0.015) [0.258]	0.028* (0.016) [0.087]	0.024 (0.018) [0.187]
β_5	*ViewedOverrideBad*	-0.026*** (0.007) [0.000]	0.010 (0.017) [0.556]	-0.001 (0.019) [0.946]	-0.025 (0.021) [0.220]
β_6	*Round*	-0.002*** (0.0003) [0.000]	-0.002*** (0.0004) [0.000]	-0.002*** (0.0008) [0.005]	-0.001 (0.0014) [0.362]
γ_1	$\overline{View}$	0.021 (0.013) [0.115]	0.026* (0.013) [0.051]	0.017 (0.014) [0.218]	0.003 (0.016) [0.868]
$\beta_3 - \beta_2$		-0.003 (0.006) [0.658]	0.005 (0.008) [0.526]	-0.029 (0.018) [0.105]	0.022 (0.033) [0.504]
$\beta_5 - \beta_4$		-0.030*** (0.007) [0.000]	-0.007 (0.012) [0.532]	-0.029** (0.014) [0.036]	-0.050*** (0.019) [0.008]
Observations		1,246	499	324	148
Managers		44	44	44	43
Random-intercept SD		0.146	0.145	0.140	0.144

We find, in Table EC.9, that our main results are robust. Each additional bad task reduces the bonus share (Column (1), $\beta_1 = -0.023$, $p < 0.01$). More importantly, managers reduce bonus shares more after observing an *override* than after observing a *follow* that results in a bad outcome. This pattern holds in the pooled specification (Column (1), $\beta_5 - \beta_4 = -0.030$, $p < 0.01$), and when conditioning on rounds with one bad outcome (Column (2), $\beta_5 - \beta_4 = -0.007$, $p = 0.53$), two bad outcomes (Column (3), $\beta_5 - \beta_4 = -0.029$, $p = 0.04$), and three bad outcomes (Column (4), $\beta_5 - \beta_4 = -0.050$, $p < 0.01$).

Notably, when the data are stratified by the number of bad task outcomes in a round, the qualitative pattern mirrors that in Table EC.8, where the dependent variable is the absolute bonus level ($Bonus^{DM}$). However, the pooled effect is statistically significant here but not in Table EC.8. This difference suggests that mechanical constraints on the absolute bonus level, due to variation in total payoffs across rounds, may have attenuated the aggregate effect in the earlier specification.

C.3.2. Manager: Viewing Decisions. We fit a linear probability model, $ViewM = \beta_1 BadOutcome + \beta_2 Round + \beta_3(BadOutcome \times Round)$ (Table EC.10), and find that managers are on average more likely to view the machine signal after a bad outcome ($\beta_1 = 0.082, p = 0.09$). Further, while the tendency to view decreases after good outcomes ($\beta_2 = -0.007, p = 0.02$), it remains at a high level after bad outcomes ($\beta_2 + \beta_3 = -0.005, p = 0.16$).

Table EC.10 Principal Viewing Behavior over Rounds

DV: $ViewM$		Estimate
β_0	Constant	0.389*** (0.054) [0.000]
β_1	$BadOutcome$	0.082* (0.043) [0.087]
β_2	$Round$	-0.0068** (0.0025) [0.021]
β_3	$BadOutcome \times Round$	0.0022 (0.0019) [0.291]
$\beta_2 + \beta_3$		-0.0047 (0.0031) [0.161]
Observations		6,600
Participants (clusters)		44
Sessions (clusters)		11
R^2		0.018
Within R^2		0.018

C.3.3. Does Algorithm Adherence Pay Off? To assess whether DMs receive higher bonuses for overriding the machine signal when the human signal is more precise and in conflict with the machine ($\theta = hi$ and $Conflict = 1$), we estimate the regression $BonusReceive_{it} = \beta_0 + \beta_1 FollowM_{it} + \varepsilon_{it}$, with standard errors two-way clustered at the DM and session levels (Table EC.11). We find that the bonus for following the machine signal is lower than for overriding ($\beta_1 = -3.034$, $p = 0.08$).

Table EC.11 Effect of Following the Machine on DM Bonuses ($\theta = hi$ and $Conflict = 1$).

DV: $BonusReceive$		Estimate
β_0	Constant	12.789*** (1.384) [0.000]
β_1	$FollowM$	-3.034* (1.556) [0.080]
Observations		1,251
Participants (clusters)		44
Sessions (clusters)		11

C.4. Overall Performance.

Table EC.12 reports performance comparisons between the *Delegation* and *NoDelegation* treatments across all studies, and additionally compares each treatment to a *NoHuman* benchmark. Performance is measured using the ex-ante expected payoff of the decision, rather than the realized payoff, in order to abstract from outcome noise and isolate decision quality. Between-treatment p-values are based on Welch two-sample t-tests. Benchmark p-values are based on one-sample t-tests against the *NoHuman* benchmark.

Table EC.12 Between-Treatment Expected Payoff Comparisons (Participant-Level Means)

Study	Condition	*NoDelegation*	*Delegation*	p	*NoHuman*	vs. *NoDelegation*	vs. *Delegation*
1	q_{s_H} = 60%, Conflict = 0	62.12	62.65	0.843	55	0.002	< 0.001
	q_{s_H} = 60%, Conflict = 1	36.21	38.74	0.084	55	< 0.001	< 0.001
	q_{s_H} = 60% (Overall)	49.16	50.70	0.412	55	< 0.001	0.002
	q_{s_H} = 80%, Conflict = 0	84.88	79.60	0.101	55	< 0.001	< 0.001
	q_{s_H} = 80%, Conflict = 1	40.46	34.47	0.026	55	< 0.001	< 0.001
	q_{s_H} = 80% (Overall)	62.67	57.04	0.023	55	< 0.001	0.359
	Conflict = 0 (Overall)	73.50	71.13	0.378	55	< 0.001	< 0.001
	Conflict = 1 (Overall)	38.33	36.61	0.335	55	< 0.001	< 0.001
	Overall	55.92	53.87	0.297	55	0.371	0.501
2†	Conflict = 0	82.72	80.10	0.023	72	< 0.001	< 0.001
	Conflict = 1	70.21	67.98	0.001	72	< 0.001	< 0.001
	Overall (high q_{s_H})	76.46	74.04	0.002	72	< 0.001	0.003
2‡	Conflict = 0	82.72	83.21	0.531	72	< 0.001	< 0.001
	Conflict = 1	70.21	70.44	0.730	72	< 0.001	< 0.001
	Overall (high q_{s_H})	76.46	77.26	0.581	72	< 0.001	< 0.001
3	q_{s_H} = 60%, Conflict = 0	66.21	58.69	0.001	55	< 0.001	0.095
	q_{s_H} = 60%, Conflict = 1	38.26	35.62	0.059	55	< 0.001	< 0.001
	q_{s_H} = 60% (Overall)	52.24	47.15	0.003	55	< 0.001	< 0.001
	q_{s_H} = 80%, Conflict = 0	85.11	76.85	0.007	55	< 0.001	< 0.001
	q_{s_H} = 80%, Conflict = 1	40.04	37.04	0.137	55	< 0.001	< 0.001
	q_{s_H} = 80% (Overall)	62.57	56.94	0.007	55	< 0.001	0.292
	Conflict = 0 (Overall)	75.66	67.77	0.003	55	< 0.001	< 0.001
	Conflict = 1 (Overall)	39.15	36.33	0.043	55	< 0.001	< 0.001
	Overall	57.41	52.05	0.003	55	< 0.001	0.073

Notes: † *Delegation* treatment, ‡ *DelegationObs* treatment.

C.5. Belief Elicitation.

At the end of the experiments for Studies 1 and 3, we administered a short incentivized survey that elicited from participants (regardless of their role in the experiment) their beliefs about key objects central to our model. First in part A, from the DM's perspective, we elicited beliefs about the posterior probability of the true state after observing signals $\mathbf{s}$, namely $\mu_{\mathbf{s}}$. Second in part B, from the manager's perspective, we elicited beliefs about the DM's private information at the time of the decision, namely $\Pr(s_H, \theta \mid s_M, a, y)$.

For each question, participants reported a probability (in percent) using a slider. Payments were determined using a quadratic scoring rule. Let $g \in [0,1]$ denote the reported probability and let $p \in [0,1]$ denote the correct benchmark value for that question.[29] For each question we computed a Score $= \$1 - (g-p)^2$; final payments from the survey were based on the average score across all questions in Parts A and B.

C.5.1. Part A: DM perspective. Table EC.13 presents summary statistics for the first belief elicitation task, where participants guessed the probability of the true state given signals $\mathbf{s}$. Informational states varied by whether the human and machine signals aligned ($s_M = s_H$) or conflicted ($s_M \neq s_H$), and whether the human signal was of high precision ($q^{hi}_{s_H} > q_{s_M}$) or low precision ($q^{lo}_{s_H} < q_{s_M}$).

[29] The benchmark p corresponded to the theoretical Bayesian value in the first task and to the empirical benchmark (within session) in the second task.

Table EC.13 Belief Elicitation: Stated Probability that the State Matches the More Precise Signal

				$s_M = s_H$, $q^{hi}_{s_H} > q_{s_M}$	$s_M \neq s_H$, $q^{hi}_{s_H} > q_{s_M}$	$s_M = s_H$, $q^{lo}_{s_H} < q_{s_M}$	$s_M \neq s_H$, $q^{lo}_{s_H} < q_{s_M}$
Study	Treatment	Role	N	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
1	Delegation	Manager	40	82.35 (12.53)	56.60 (19.60)	76.40 (14.85)	55.18 (13.43)
		DM	40	74.75 (14.89)	60.27 (17.43)	66.00 (13.75)	56.68 (14.45)
	NoDelegation	–	25	80.24 (13.18)	64.04 (13.37)	72.68 (13.19)	52.52 (10.20)
3	Delegation	Manager	44	79.70 (11.34)	57.79 (13.94)	74.29 (12.37)	57.32 (13.43)
		DM	44	82.95 (11.34)	62.95 (13.82)	71.86 (14.90)	52.75 (16.12)
	NoDelegation	–	28	77.71 (14.22)	57.07 (13.32)	68.68 (10.61)	57.82 (10.76)
True Value (Bayesian)			–	90	63	78	61

Notes: Participants were shown the realizations of $\mathbf{s}$ using color labels (as in the experimental interface) and were asked to report the probability that the true state was "blue." Responses are transformed so that all entries correspond to the probability that the state matches the more precise signal.

To assess how participants integrate human and machine signals, we exploit the log-odds representation of Bayesian updating. For a binary signal with precision p, the log-likelihood ratio is given by $\log\left(\frac{p}{1-p}\right)$. Under Bayesian updating, and since $\mu_0 = 1/2$, the posterior log-odds equal the sum of the human and machine log-likelihood ratios. We therefore estimate the following specification:

$$\text{logit}(Belief_{i\mathbf{s}}) = \beta_H \, LLR_{H,\mathbf{s}} + \beta_M \, LLR_{M,\mathbf{s}} + \varepsilon_{i\mathbf{s}}, \tag{EC.69}$$

where $LLR_{H,\mathbf{s}}$ and $LLR_{M,\mathbf{s}}$ denote the log-likelihood ratios implied by the two signals in scenario $\mathbf{s}$. Under Bayesian updating, the coefficients should satisfy $\beta_H = \beta_M = 1$. Deviations from unity indicate that participants under- or overweight the respective signal relative to the Bayesian benchmark. We therefore additionally test the hypotheses $H_1 : \beta_H = 1$ and $H_1 : \beta_M = 1$. The results are reported in Table EC.14.

Table EC.14 Signal Weights by Study, Treatment, and Role

DV: logit($Belief$)	Study 1			Study 3		
	NoDelegation	Delegation		NoDelegation	Delegation	
		DM	Manager		DM	Manager
β_H	0.726***	0.868***	0.501***	0.725***	0.787***	0.511***
	(0.172) [0.000]	(0.244) [0.001]	(0.139) [0.001]	(0.136) [0.000]	(0.118) [0.000]	(0.140) [0.001]
β_M	0.680***	0.593***	1.188***	0.742***	0.744***	0.925***
	(0.118) [0.000]	(0.105) [0.000]	(0.193) [0.000]	(0.118) [0.000]	(0.081) [0.000]	(0.135) [0.000]
$\beta_H = 1$	-0.274	-0.132	-0.499***	-0.275*	-0.213*	-0.489***
	(0.172) [0.125]	(0.244) [0.594]	(0.139) [0.001]	(0.136) [0.054]	(0.118) [0.080]	(0.140) [0.001]
$\beta_M = 1$	-0.320**	-0.407***	0.188	-0.258**	-0.256***	-0.075
	(0.118) [0.012]	(0.105) [0.000]	(0.193) [0.336]	(0.118) [0.038]	(0.081) [0.003]	(0.135) [0.583]
$\beta_H - \beta_M$	0.046	0.275	-0.687***	-0.017	0.044	-0.414**
	(0.140) [0.747]	(0.215) [0.207]	(0.244) [0.008]	(0.107) [0.873]	(0.114) [0.704]	(0.182) [0.028]
Observations	100	160	160	112	176	176
Participants (clusters)	25	40	40	28	44	44

Table EC.14 shows that participants generally place weights below one on both the human and the machine signals, indicating attenuation relative to the Bayesian benchmark. However, this underweighting is not uniform across roles and treatments. Importantly, a clear role-based asymmetry emerges. In both studies, managers in the *Delegation* treatment place significantly greater weight on the machine signal than on the human signal. In contrast, DMs and participants in the *NoDelegation* treatments do not exhibit such a relative overweighting of the machine signal. The results makes intuitive sense, given that the survey

was administered immediately after participants completed the main experiment. During the experiment, managers did not observe the human signal or its precision, although this information was provided in the survey task. The pattern therefore suggests that occupying the manager role systematically shifts beliefs in favor of the machine signal's precision, consistent with a role-induced bias toward machine-based information.

C.5.2. Part B: Manager perspective. To understand the cognitive mechanisms behind managers' evaluation of DMs, we presented them with a second belief elicitation task. A central object in our model, which underlies the manager's assessment of the *ex-ante* quality of the DM's decision, is the belief about the DM's private information at the time of the decision making, $\Pr(s_H, \theta \mid s_M, a, y)$. In the survey, we elicit these beliefs separately: first, regarding the realization of the human signal, $\Pr(s_H, \cdot \mid s_M, a, y)$, and second, regarding the precision of the human signal, $\Pr(\cdot, \theta \mid s_M, a, y)$.[30]

Table EC.15 presents summary statistics. We construct an *Empirical* benchmark for each scenario (s_M, a, y) which equals the observed frequency (within a given treatment) with which (i) the DM's decision matches her private human signal and (ii) the human signal is more precise. We interpret these frequencies as proxies for the corresponding undistorted beliefs, $\Pr(s_H, \cdot \mid s_M, a, y)$ and $\Pr(\cdot, \theta \mid s_M, a, y)$. In addition, we construct a *Theoretical* benchmark that corresponds to the Bayesian posterior of fully rational actors.

A central implication of our model is that managers should sharply differentiate across decision types and realized outcomes. In particular, an override relative to a follow constitutes a strong indication that the DM (i) acted on their private human signal and (ii) did so because that signal was more precise than the machine signal. Under the theoretical benchmark with a fully rational (non-noisy) DM, an override perfectly reveals high precision and adherence to the human signal, implying a belief of 100% for both probabilities, regardless of whether the outcome is ex-post good or bad. With decision noise, however, overrides remain more informative than follows, but no longer perfectly revealing. As reflected in the empirical benchmark, overrides are still stronger indicators that the DM relied on a high-precision human signal, yet this inference is weaker after bad outcomes than after good outcomes. This naturally raises the question of whether participants' stated beliefs are systematically biased relative to the empirical benchmark. Moreover, do their beliefs track the variation in the benchmark across scenarios (follow/override machine, good/bad outcome)? To address these questions, we estimate the following fixed-effects regression by treatment and role:

$$\mathit{Belief}_{is}^{(k)} = \alpha_i + \beta^{(k)}\, \mathit{Empirical}_{s}^{(k)} + \varepsilon_{is}^{(k)}, \qquad k \in \{\mathit{match}, \mathit{prec}\}, \tag{EC.70}$$

where $\mathit{Belief}_{is}^{(k)}$ denotes participant i's stated belief in scenario s for question k, with $k = \mathit{match}$ referring to the belief that the human signal matches the DM's decision, and $k = \mathit{prec}$ referring to the belief that the human signal is more precise. The term α_i captures participant fixed effects, and $\mathit{Empirical}_{s}^{(k)}$ denotes the corresponding empirical benchmark probability in scenario s for question k. The coefficient $\beta^{(k)}$ measures the extent to which stated beliefs align with $\mathit{Empirical}_{s}^{(k)}$. Results are reported in Table EC.16.

Two main findings emerge. First, participants' stated beliefs are systematically biased relative to the empirical benchmark. In all treatments and roles, beliefs track the directional variation in the benchmark

[30] An alternative would have been to elicit the joint belief that the DM had a more precise human signal and followed it. However, we opted to separate the elicitation into two simpler questions to reduce cognitive complexity.

Table EC.15 Belief Elicitation from Manager's Perspective

Question (i): Probability that the human signal matches the DM decision, $\Pr(s_H = a \mid s_M, a, y)$

Study	Treatment	Role/Benchmark	N	Follow M, Good	Override M, Good	Follow M, Bad	Override M, Bad
1	*Delegation*	Manager	40	26.65 (22.13)	67.32 (28.14)	54.18 (28.01)	59.38 (34.08)
		DM	40	44.07 (22.39)	73.85 (20.70)	44.68 (21.44)	51.18 (30.34)
		Empirical	–	25.18	94.90	51.90	73.33
		Theoretical	–	22.2	100.0	50.0	100.0
1	*NoDelegation*	Participant	25	32.80 (14.76)	75.76 (20.81)	49.60 (18.97)	55.76 (30.64)
		Empirical	–	24.48	96.18	44.80	88.57
		Theoretical	–	22.2	100.0	50.0	100.0
3	*Delegation*	Manager	44	29.34 (20.87)	74.45 (10.26)	42.62 (17.01)	52.84 (26.98)
		DM	44	29.20 (22.15)	76.04 (23.33)	42.59 (21.75)	56.86 (33.83)
		Empirical	–	22.48	92.21	52.91	66.09
		Theoretical	–	22.2	100.0	50.0	100.0
3	*NoDelegation*	Participant	28	30.54 (13.75)	73.04 (17.38)	54.64 (19.72)	53.64 (27.85)
		Empirical	–	21.18	99.79	51.13	96.84
		Theoretical	–	22.2	100.0	50.0	100.0

Notes: In the survey, participants were shown the realizations of (s_M, a, y) using color labels (as in the experimental interface) and were asked to report the probability that the human signal was "blue." Responses are transformed so that all entries correspond to the probability that the human signal matches the DM decision.

Question (ii): Probability that the human signal is more precise, $\Pr(\theta = hi \mid s_M, a, y)$

Study	Treatment	Role/Benchmark	N	Follow M, Good	Override M, Good	Follow M, Bad	Override M, Bad
1	*Delegation*	Manager	40	36.35 (22.05)	68.00 (21.90)	60.07 (19.17)	55.00 (27.66)
		DM	40	49.57 (20.98)	64.35 (26.48)	48.62 (19.43)	58.55 (23.88)
		Empirical	–	48.91	92.99	28.28	69.63
		Theoretical	–	44.4	100.0	16.7	100.0
1	*NoDelegation*	Participant	25	44.56 (17.31)	64.36 (26.51)	47.04 (17.35)	62.44 (26.48)
		Empirical	–	47.73	89.31	22.40	61.90
		Theoretical	–	44.4	100.0	16.7	100.0
3	*Delegation*	Manager	44	45.34 (25.18)	74.77 (19.24)	48.09 (22.27)	59.04 (26.57)
		DM	44	48.66 (27.55)	71.48 (27.17)	52.04 (25.36)	59.98 (32.21)
		Empirical	–	47.11	81.44	26.89	63.05
		Theoretical	–	44.4	100.0	16.7	100.0
3	NoDelegation	Participant	28	46.07 (10.37)	65.54 (21.34)	48.67 (20.35)	55.78 (22.48)
		Empirical	–	45.63	92.34	22.80	80.13
		Theoretical	–	44.4	100.0	16.7	100.0

Table EC.16 Directional Alignment Regressions with Benchmark Tests

DV: Belief$^{(k)}$	Study 1			Study 3		
	NoDelegation	Delegation		NoDelegation	Delegation	
		DM	Manager		DM	Manager
$\beta^{(match)}$	0.451**	0.371***	0.488***	0.362**	0.454***	0.439***
	(0.191) [0.027]	(0.120) [0.004]	(0.157) [0.003]	(0.138) [0.014]	(0.118) [0.000]	(0.093) [0.000]
$\beta^{(prec)}$	0.303**	0.262***	0.206***	0.237**	0.380***	0.513***
	(0.135) [0.034]	(0.078) [0.002]	(0.062) [0.002]	(0.104) [0.030]	(0.102) [0.001]	(0.106) [0.000]
$\beta^{(match)} = 1$	-0.549***	-0.629***	-0.512***	-0.638***	-0.546***	-0.561***
	(0.191) [0.008]	(0.120) [0.000]	(0.157) [0.002]	(0.138) [0.000]	(0.117) [0.000]	(0.093) [0.000]
$\beta^{(prec)} = 1$	-0.697***	-0.738***	-0.795***	-0.763***	-0.620***	-0.487***
	(0.135) [0.000]	(0.078) [0.000]	(0.062) [0.000]	(0.103) [0.000]	(0.102) [0.000]	(0.106) [0.000]
Observations	100	160	160	112	176	176
Participants (clusters)	25	40	40	28	44	44

across scenarios ($\beta^{(k)} > 0$, $p < 0.01$), implying that participants adjust their beliefs in the correct direction when the empirical benchmark increases. In particular, beliefs systematically respond to whether the DM followed or overrode the machine and whether the outcome was good or bad, indicating that participants incorporate these key informational cues when forming their assessments. However, beliefs are insufficiently sensitive to the variation in the empirical probabilities ($\beta^{(k)} < 1$, $p < 0.01$).