When Thinking Pays Off: Incentive Alignment for Human-AI Collaboration

JOSHUA HOLSTEIN*, Karlsruhe Institute of Technology, Germany
PATRICK HEMMER*, Karlsruhe Institute of Technology, Germany
GERHARD SATZGER, Karlsruhe Institute of Technology, Germany
WEI SUN, IBM Research, USA

Collaboration with artificial intelligence (AI) has improved human decision-making across various domains by leveraging the complementary capabilities of humans and AI. Yet, humans systematically overrely on AI advice, even when their independent judgment would yield superior outcomes, fundamentally undermining the potential of human-AI complementarity. Building on prior work, we identify prevailing incentive structures in human-AI decision-making as a structural driver of this overreliance. To address this misalignment, we propose an alternative incentive mechanism designed to counteract systemic overreliance. We empirically evaluate this approach through a behavioral experiment with 180 participants, finding that the proposed mechanism significantly reduces overreliance. We also show that while appropriately designed incentives can enhance collaboration and decision quality, poorly designed incentives may distort behavior, introduce unintended consequences, and ultimately degrade performance. These findings underscore the importance of aligning incentives with task context and human-AI complementarities, and suggest that effective collaboration requires a shift toward context-sensitive incentive design.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Human computer interaction (HCI)**; **HCI theory, concepts and models**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Overreliance, Mechanism Design, AI-Assisted Decision-Making, Appropriate Reliance, Human-AI Collaboration, Human-AI Decision-Making

ACM Reference Format:
Joshua Holstein, Patrick Hemmer, Gerhard Satzger, and Wei Sun. 2026. When Thinking Pays Off: Incentive Alignment for Human-AI Collaboration. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/3805689.3812250

1 Introduction

Artificial intelligence (AI) is being increasingly introduced to support human decision-making across diverse domains, promising higher decision performance and efficiency. Such performance improvements, however, depend on the complementary capabilities of both partners [41, 50]: While AI excels at processing vast amounts of data and identifying complex patterns, humans contribute contextual understanding, ethical reasoning, and the ability to handle novel or ambiguous situations [54]. However, realizing this complementarity potential in human-AI collaboration fundamentally hinges on the ability of human decision-makers to appropriately rely on

*These authors contributed equally to this work.

Authors' Contact Information: Joshua Holstein, joshua.holstein@kit.edu, Karlsruhe Institute of Technology, Germany; Patrick Hemmer, patrick.hemmer@partner.kit.edu, Karlsruhe Institute of Technology, Germany; Gerhard Satzger, gerhard.satzger@kit.edu, Karlsruhe Institute of Technology, Germany; Wei Sun, sunw@us.ibm.com, IBM Research, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812250

AI advice [22, 61], i.e., knowing when to follow AI advice and when to override it based on their own judgment and expertise.

Yet, achieving this appropriate reliance proves challenging in practice as human decision-makers frequently rely on AI advice, even in situations where their independent judgment would lead to better outcomes [72], i.e., they "overrely". The implications of this overreliance are concerning for high-stakes decisions in domains such as medical diagnosis or criminal justice, where overreliance can have profound consequences for individuals, organizations, and society overall [57]. For example, when decision-makers follow incorrect AI advice, decision quality suffers, and human expertise may gradually erode [38], a pattern supported by evidence that immediate AI collaboration reduces brain activity compared to working independently before collaborating [36]. To address these challenges, researchers have explored various approaches, yet overreliance remains a persistent problem [26].

While existing interventions have shown promise, they primarily focus on individual-level cognitive factors [56, 65] and information provision like explanations and confidence [6, 11], thereby often overlooking the broader environmental factors that shape reliance behavior, as research on human oversight effectiveness highlights that individual-level interventions may be insufficient without addressing the broader socio-technical environment [66]. This oversight is problematic given that real-world AI-assisted decision-making typically occurs within organizational contexts where performance metrics, rewards, and accountability structures may inadvertently shape reliance behavior. Indeed, meta-analytic evidence from traditional environments shows that monetary incentives significantly enhance task performance by influencing both intrinsic motivation and effort allocation [13]. Although the role of incentive designs on human behavior is well-established in organizational behavior research [53], their role in human-AI collaboration remains largely unexplored, where the asymmetric cost structure between effortless AI advice without reciprocal obligations and costly independent effort creates novel dynamics. This leads to our research questions that progress from a theoretical analysis to empirical evaluation:

RQ1: *What is the role of incentive mechanisms on human reliance in AI-assisted decision making?*
RQ2: *How do different incentive mechanisms affect human reliance and performance in AI-assisted decision making?*

To address these research questions, we draw on an established theoretical framework in human-AI decision-making by Bansal et al. [4] and identify that systematic overreliance on AI advice cannot be attributed exclusively to the human, but rather stems from an inherent structural cause—a misalignment of incentives embedded in traditional AI-assisted decision-making designs. Even when humans are perfectly rational decision-makers operating with complete information, this structural misalignment systematically biases them toward overreliance, suggesting that the solution lies not in training humans or improving AI explanations, but in modifications to the underlying incentive mechanism. Based on our analysis, we extend this framework by designing an alternative incentive mechanism that introduces monetary rewards for independent decision-making, directly targeting the system-induced overreliance behavior. We then evaluate these incentive mechanisms through a between-subjects behavioral experiment with 180 participants. Our results show that incentive mechanisms significantly reduce overreliance, resulting in higher human-AI team performance. Overall, this work makes three contributions to the field of human-AI collaboration: First, we identify a misalignment of incentives in AI-assisted decision-making as a systemic source for overreliance on AI advice. Second, we propose a new incentive mechanism specifically counteracting the structural misalignment, aiming to reduce human overreliance. Third, we empirically investigate how different instantiations of this incentive mechanism affect human reliance behavior and performance in AI-assisted decision making. We show that dynamic incentives reduce overreliance and improve human-AI team performance, whereas static incentives introduce unintended gaming behaviors that prioritize reward acquisition over performance.

2 Related Work

Research on human-AI collaboration has extensively investigated the challenge of appropriate reliance, where humans must discern when to follow or override AI advice [61]. A substantial body of work has documented the persistent problem of overreliance, where humans systematically accept AI advice even when their independent judgment would yield superior outcomes [26, 60], particularly problematic in high-stakes decision-making.

To address overreliance, researchers have pursued several intervention strategies. Information-based approaches provide additional contextual information, such as AI confidence scores [6, 35, 81], specialized training [51, 64], or explanations of AI reasoning through feature-based [55, 62], example-based [2, 74], or natural language descriptions [11, 35]. Complementary human-centered interventions focus on the development of trust [48] and its calibration [32, 33] as well as improving mental models of AI capabilities [5, 73] or even predicting human behavior [39], while cognitive interventions employ forcing functions [10], confidence alignment techniques [43], and prompts encouraging consideration of alternatives [11].

Although these approaches have demonstrated promising results, they predominantly focus on the AI system or the humans, neglecting the broader socio-technical decision-making environment [66], e.g., incentive structures in organizational settings. Incentive structures are a well-established tool in organizations to empower employees toward better outcomes [53]. Conceptually, it is assumed that monetary incentives can increase effort, which can translate into higher task performance [9]. For example, Stone and Ziebart [69] show that performance-contingent incentives induce a longer time to make a decision by examining more information and pursuing decision strategies that ultimately lead to more accurate choices compared to randomly distributed incentives. However, research also highlights the difficulty for incentive mechanisms in capturing all relevant behavioral aspects that can lead to dysfunctional behavioral responses [53]. In multi-dimensional tasks, employees may optimize for one dimension at the expense of the other if the incentive mechanism only targets a single dimension [29].

These dynamics have also been extensively studied in crowdwork settings, where online platforms such as MTurk or Prolific provide a natural laboratory for studying payment structures and their behavioral consequences. A foundational distinction pertains to the type of motivation: Mao et al. [45] show systematic differences between volunteer and paid crowdworkers in both motivation and contribution quality, whereas Liang et al. [40] demonstrate that intrinsic motivation interacts with extrinsic incentives. Specifically, they show that the effect of financial rewards on effort is moderated by workers' intrinsic interest in the task. Regarding the design of payment schemes, evidence suggests that bonuses do not uniformly improve performance. Ho et al. [28] establish that performance-based payments raise the quality only for effort-responsive tasks, which aligns with expectancy theory [75], which explains that workers will only exert additional effort if they believe it can be converted into better performance and, in turn, into higher reward. Similarly, Yin and Chen [80] show that the decision of whether and how to offer bonuses must be tailored to task characteristics. Beyond effort, incentive structures also influence participation. Patel et al. [49] find that monetary reward structures induce self-selection effects in crowdsourcing contests, altering the composition of contributors and creating trade-offs between participation breadth and contribution appropriateness, resembling an instance of the multi-dimensional incentive problem noted above [29]. Finally, financial incentives appear most effective when combined with social mechanisms: Shaw et al. [63] find that prompting workers to consider peer responses, alongside monetary payment, yields more accurate judgments than financial incentives alone—a result particularly relevant to the judgment tasks common in AI-assisted decision-making experiments.

Taken together, this body of work suggests that the effectiveness of incentive schemes is highly context-dependent, shaped, for example, by task characteristics, worker motivation, and participation structure. Yet how these dynamics operate when humans decide alongside AI advice remains largely unexplored [17]. Notably, Wu et al. [78] find that while human-AI collaboration often enhances task performance, it can also simultaneously undermine workers' intrinsic motivation, suggesting that the introduction of AI advice carries incentive-relevant

consequences that standard payment schemes do not account for. To the best of our knowledge, only Wang et al. [77] consider the decision-making environment in their work, by varying the rewards associated with each correct decision. However, neither do they study AI advice under the lens of overreliance, nor do they theoretically underpin the decision-making environment; instead, they focus on predicting human behavior. In contrast, we comprehensively analyze and extend an established AI-assisted decision-making framework [4] and derive a theoretical model explaining human overreliance on AI advice, which we underline empirically through a behavioral experiment.

3 Problem Description

In this section, we introduce a formal model of AI-assisted decision-making, which reveals a fundamental misalignment: due to effort costs, humans may rationally defer to AI, leading to overreliance. To address this issue, we propose an incentive-based intervention that introduces a bonus for independent human decisions, designed to realign incentives and mitigate overreliance.

3.1 Effort-Induced Overreliance

We build on the framework of Bansal et al. [4] to formalize the AI-assisted decision-making setting, where humans receive AI advice but retain control over the final decision. Let $x \in X \subset \mathbb{R}^n$ denote a task instance and Y a finite set of possible decisions. For each x, the AI outputs a probability distribution $h(x)$ over Y, from which the predicted label is $\hat{y}_{\text{AI}} = \arg\max h(x)$, with associated confidence $P_{\text{AI}} = \max h(x)$, assuming calibrated scores.

Given this information, the human makes a meta-decision $m \in \{\text{accept}, \text{solve}\}$: whether to accept the AI's recommendation or solve the task instance independently. If the human solves the task instance, the final decision is $\hat{y}_{\text{H}}$; otherwise, it is $\hat{y}_{\text{AI}}$. The environment then returns a utility U based on the correctness of the final decision.

We model the consequences of decisions using a reward $\gamma > 0$ for correct outcomes and a penalty $\beta > 0$ for incorrect ones. In addition, solving a task instance incurs an effort cost $\lambda > 0$, representing time or cognitive load. The resulting utility outcomes are summarized in Table 1.

	Correct	**Incorrect**
Accept	γ	$-\beta$
Solve	$\gamma - \lambda$	$-\beta - \lambda$

Table 1. Utility outcomes in AI-assisted decision-making.

Let $P_{\text{H}} = P(\hat{y}_{\text{H}} = y \mid m = \text{solve})$ denote the human's probability of being correct when solving the task instance. Assuming the human is a rational agent who trusts the AI, their meta-decision maximizes expected utility. Thus, the human will accept the AI's advice if and only if:

$$\mathbb{E}[U(m = \text{Accept})] \geq \mathbb{E}[U(m = \text{Solve})]$$
$$P_{\text{AI}}\gamma + (1 - P_{\text{AI}})(-\beta) \geq P_{\text{H}}(\gamma - \lambda) + (1 - P_{\text{H}})(-\beta - \lambda)$$

By rearranging the expected utility comparison, the condition under which a human accepts the AI's advice becomes:

$$P_{\text{AI}} \geq P_{\text{H}} - \frac{\lambda}{\gamma + \beta}$$

This condition reveals a structural misalignment: the human is incentivized to defer to the AI whenever its probability of being correct exceeds the human's, adjusted downward by the term $\frac{\lambda}{\gamma+\beta}$. Since $\lambda > 0$, the adjustment

is strictly positive, lowering the threshold for accepting AI advice. As a result, the human may rationally choose to follow the AI even when $P_H > P_{AI}$. While this misalignment decreases with negligible effort or sufficiently high stakes (i.e., large $\gamma + \beta$), it vanishes only when $\lambda = 0$.

In effect, the human does not defer because the AI is more accurate, but because the *effort to override is not worth the marginal gain*. This demonstrates that overreliance can emerge *even under perfect information*, driven not by irrationality but by the incentive structure itself.

3.2 Incentive Alignment via Bonus Mechanism

To address this *misalignment of incentives*, we propose a novel incentive mechanism for human-AI decision-making that aims to *mitigate overreliance* on AI. Specifically, we introduce a new *bonus* term, denoted by θ, awarded to human decision-makers who invest effort and correctly solve a task instance independently. The updated utility structure with the bonus is summarized in Table 2.

	Correct	**Incorrect**
Accept	γ	$-\beta$
Solve	$\gamma - \lambda + \theta$	$-\beta - \lambda$

Table 2. Updated potential outcomes with the bonus θ.

Incorporating this bonus θ, the decision rule becomes:

$$\mathbb{E}[U(m = \text{Accept})] \geq \mathbb{E}[U(m = \text{Solve})]$$
$$P_{AI}\gamma + (1 - P_{AI})(-\beta) \geq P_H(\gamma - \lambda + \theta) + (1 - P_H)(-\beta - \lambda)$$

Rearranging terms yields a new threshold condition:

$$P_{AI} \geq P_H + \frac{\theta P_H - \lambda}{\gamma + \beta}$$

This expression highlights the incentive correction introduced by θ, i.e., when $\theta = \lambda / P_H$, the right-hand term vanishes, eliminating the incentive misalignment. Accordingly, θ should scale with the required effort λ, providing stronger incentives when solving is more costly or when the human success probability P_H is low. By compensating for effort and/or task difficulty, the bonus θ encourages independent problem-solving and mitigates overreliance on AI advice, thereby promoting more balanced and effective human-AI collaboration.

The theoretical framework reveals that effective human-AI collaboration requires reconceptualizing incentive structures beyond simple rewards for correct decisions. Rather than providing only a fixed reward γ, we introduce bonuses θ that allow for the redistribution of the total available compensation between rewards (γ) and bonuses (θ) for independent decision-making, while keeping the maximum payouts fixed.

This enables different bonus allocation strategies ranging from *static bonus* across all instances to *dynamic bonus* that varies the magnitude and availability of bonuses across different task instances based on where the expected utility of human effort is largest. In the following, we present a behavioral experiment to empirically investigate the effectiveness of these mechanisms in mitigating overreliance on AI.

4 Behavioral Experiment Design

To evaluate the effect of our proposed incentive mechanism, we conducted a behavioral experiment with 180 participants performing an image classification task, approved by the university's institutional review board. We employed a between-subjects design with three conditions: a *Baseline* condition replicating standard AI-assisted decision-making, a *Static Bonus* condition providing a fixed bonus for independent correct classifications, and

a *Dynamic Bonus* condition providing a variable bonus targeted at instances where human judgment is most valuable. Below, we describe the experimental design, including the task and dataset, the AI system, and the overall setup of the behavioral study.

Dataset and Task. We selected a task that requires no prior domain knowledge but remains sufficiently challenging to demand meaningful effort. As noted by Fügener et al. [19], such generic tasks often yield more generalizable insights, ensuring that observed effects are attributable to the incentive mechanism rather than task-specific confounds, and establishing a foundation for future studies to build upon with increased complexity. To this end, we used a multi-class image classification task based on the ImageNet-16H dataset [68], which comprises 1,200 images across 16 classes with phase noise distortion applied. This distortion introduces varying levels of difficulty for both human and AI classification. In addition to ground-truth class labels, the dataset includes annotations from six independent human annotators per image. We used annotator disagreement as a proxy for human probability (P_H) to correctly solve the instance: higher disagreement indicates greater difficulty in classification, while lower disagreement suggests the task is easier for humans.

AI System. We fine-tuned a DenseNet-161 [31] pre-trained on ImageNet [58] on the distorted images. Details are left to the Appendix.

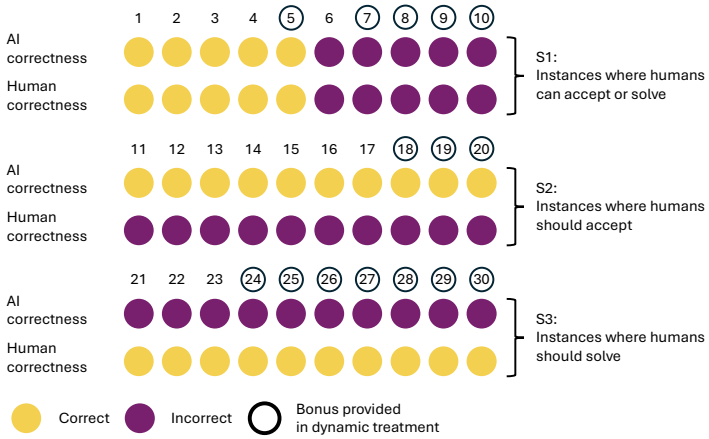

Fig. 1. Overview of the 30-instance selection scheme.

Instance Selection. For the experiment, we aimed to simulate authentic real-world scenarios where the relative strengths of humans and AI vary across task instances. We selected 30 instances representing three complementary scenarios (Figure 1): (S1) both humans and AI perform similarly, (S2) AI outperforms humans, and (S3) humans outperform AI. To determine the relative performance of humans and AI, we used existing human annotator performance from the dataset and compared this with whether the AI prediction was correct or incorrect to strategically sample instances. This yielded 10 instances per scenario (S1-S3), enabling evaluation of how incentive mechanisms perform across different complementarity conditions. Complete selection criteria appear in the Appendix.

Treatments. Based on our theoretical framework, we derived the following allocation of reward (γ) and bonus (θ) across treatments (for full derivations see the Appendix): In the *Baseline* condition, participants received a fixed reward for correct classifications ($\gamma = 0.06$) and no bonus for solving independently ($\theta = 0.00$), replicating the standard incentive structure our framework identifies as misaligned. In the *Static Bonus* condition, the reward was reduced ($\gamma = 0.03$) and a fixed bonus ($\theta = 0.03$) was introduced for correctly solving any instance independently.

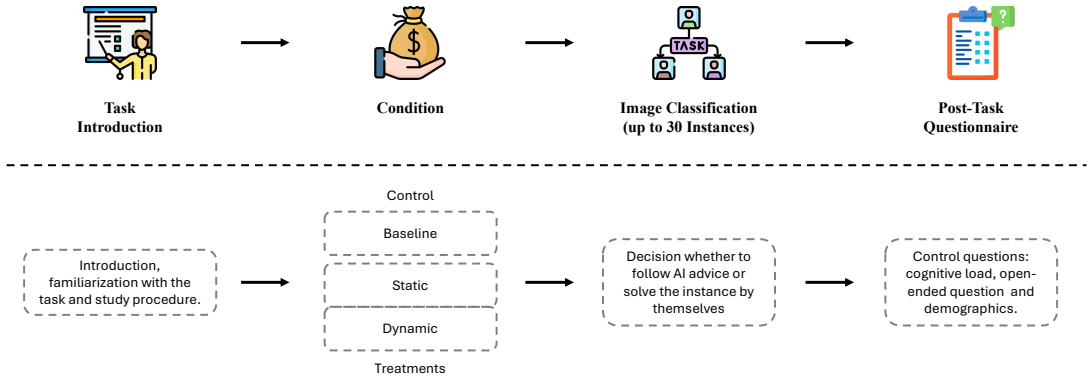

Fig. 2. Procedure of the behavioral experiment.

The bonus amount was adjusted to equalize expected payouts between always following AI advice and always solving independently. Finally, in the *Dynamic Bonus* condition, participants again received the same reduced reward ($\gamma = 0.03$), but the bonus ($\theta = 0.06$) was selectively applied only to instances where the AI was not expected to outperform the human, specifically, those with low calibrated AI confidence. No bonus was given when the AI was sufficiently confident, thereby focusing human effort where it is most valuable. In both bonus conditions, the bonus was awarded only when participants chose to solve independently and answered correctly, even if their answer then matched the AI's advice. All three conditions were calibrated to yield the same maximum payout of £1.80.

To guide effort allocation, AI confidence scores were shown in all conditions, grouped into four bins (very low, low, high, very high). A five-minute time limit was enforced to simulate effort constraints (λ), encouraging strategic decision-making under time pressure, where solving independently consumes time that could otherwise be used to complete additional instances, thereby creating an opportunity cost consistent with our theoretical model and similar to real-world scenarios.

Participants. We recruited 180 participants through Prolific, with 60 participants assigned to each treatment condition. Eligibility was restricted to U.S. residents fluent in English. The sample consisted of 43.6% male participants, with an average age of 39.6 years. The median completion time was 12 minutes, in which all participants received a performance-independent participation payment of £1. In addition to the performance-independent base payment of £1, participants earned an average of £0.81 in performance-based incentives, including both (γ) and bonus (θ) components. This results in a total average compensation of £1.81, which exceeds Prolific's recommended remuneration rate.

Study Procedure. Participants were randomly assigned to one of the three treatment conditions and introduced to the task, including the bonus structure and time limit (Figure 2). A comprehension check ensured understanding of the incentive mechanisms; participants who failed this twice were excluded from the study. Next, participants completed a guided tutorial explaining the task interface, including AI advice, confidence levels, timer, and bonus eligibility. After two training instances, they proceeded to the main task, consisting of 30 instances to be completed within a five-minute global countdown timer visible throughout the task. On each instance, participants were simultaneously presented with the distorted image, the AI's predicted class label, its confidence level, and any applicable bonus (see Figure 7 in the Appendix). Participants then chose to either follow the AI's recommendation or solve the task independently, the latter advancing them to a classification screen with a 4×4 grid of all 16 possible classes. The AI's advice was then no longer displayed on the classification screen. Following

the task, participants completed a post-task questionnaire measuring cognitive load, along with treatment-specific open-ended questions regarding how AI confidence and the payment structure influenced their decisions. An optional feedback section concluded the study.

Metrics. Our analysis examined three primary measures across experimental conditions: task completion rates (i.e., the number of instances processed within the time limit), human-AI team performance, and reliance behavior. Additionally, we measured cognitive load by using the NASA-TLX scale [23].

5 Experimental Results

5.1 Task Completion

We begin by examining the effect of the treatments on task completion, measured as the total number of instances classified within the allotted time. Our analysis reveals differences in task completion rates across experimental conditions, with static bonus participants completing the most instances ($\mu_s = 26.93$), followed by baseline participants ($\mu_b = 25.35$), and dynamic bonus participants completing the fewest instances ($\mu_d = 23.63$) (see Figure 3). As Shapiro-Wilk tests confirmed non-normal distributions and Levene's test revealed heterogeneous variances across conditions, we employ a Kruskal-Wallis test, which confirms statistically significant differences in completion rates ($H = 7.5$, $p = 0.023$), with static bonus participants processing approximately 14% more instances than dynamic bonus participants.

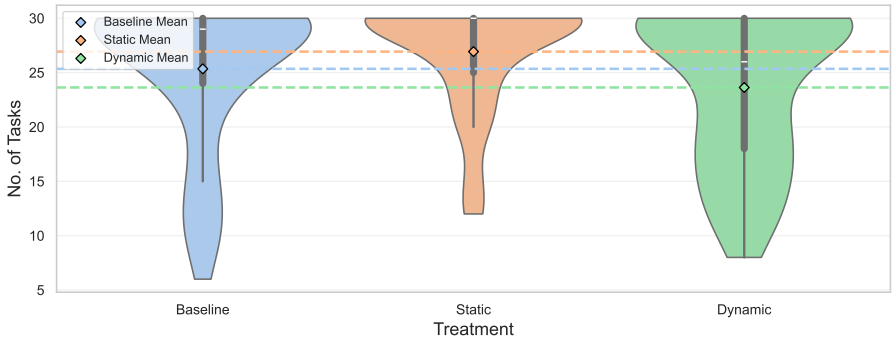

Fig. 3. Distribution of processed instances across experimental conditions.

Mediation Analysis. To explore the mechanisms underlying reduced task completion rates in the dynamic treatment, we conducted a mediation analysis examining whether cognitive load mediates the relationship between incentive structures and task completion. The analysis reveals that dynamic bonuses significantly increase participants' cognitive load compared to baseline ($\beta = 0.403, p = 0.027$), which in turn reduces task completion ($\beta = -1.094, p = 0.023$), with cognitive load accounting for approximately 26% of the total effect (see the Appendix for full details).

These systematic differences in completion rates have implications for measurement precision in our subsequent analyses. Participants who completed fewer instances yield less reliable estimates of individual performance and reliance behavior due to smaller sample sizes. Therefore, an unweighted analysis would treat all participant means as equally precise, which is inappropriate given the heteroskedasticity induced by the design. To account for this heterogeneity, we apply inverse variance weighting [24] in our analyses of human-AI team performance and reliance behavior, assigning greater weight to participants with lower variance and more observations. To further enhance the robustness of these weights, we apply winsorization at the 5th and 95th percentiles to the weight distributions within each treatment group [71] as detailed in the Appendix. All subsequent treatment

comparisons are conducted using weighted independent samples t-tests comparing each treatment against the baseline condition, with Bonferroni correction applied to control for multiple comparisons. This pairwise approach is preferred over omnibus tests as our interest lies in contrasting each incentive condition with the baseline rather than in differences among the incentive conditions themselves. Therefore, we treat the baseline condition as the reference point against which each incentive mechanism is evaluated.

5.2 Human-AI Team Performance

Next, we assess the impact of incentive mechanisms on overall human-AI team performance. As Shapiro-Wilk tests confirmed normality of the performance distributions across all three conditions, we apply the weighted t-tests introduced above. Our analysis reveals contrasting effects: the static and dynamic bonus treatments lead to opposite shifts in accuracy relative to the baseline. More specifically, participants in the static bonus condition achieved lower accuracy on average than those in the baseline group ($\mu_s = 0.632$, $\mu_b = 0.643$; $t = -6.01$, $p < 0.001$, $d = -0.10$). In contrast, participants in the dynamic bonus condition exhibited significantly higher accuracy than baseline ($\mu_d = 0.653$; $t = 5.04$, $p < 0.001$, $d = 0.09$) (see Figure 4). While these aggregate differences are modest, they mask important heterogeneity across instance types, as we detail below.

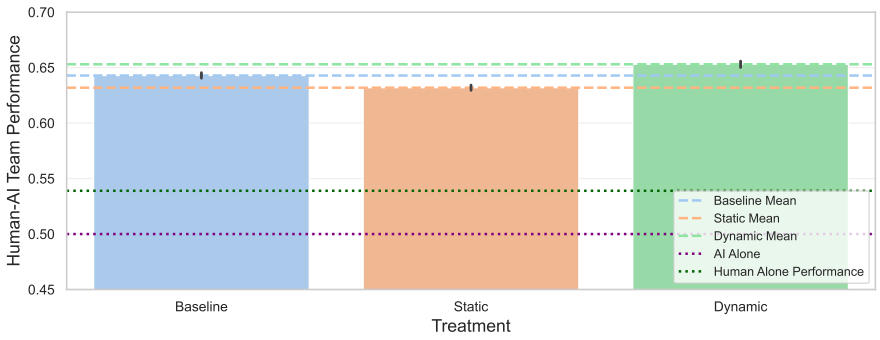

Fig. 4. Weighted average human-AI team performance across treatments with 95% confidence intervals. Reference lines show AI-only performance (purple) and human-only performance (dark green) derived from dataset ground truth labels.

Targeted Accuracy Gains from Bonus Placement. To further understand how bonus placement drives performance differences, we analyze instance-level accuracy within the dynamic treatment by comparing participant performance to the baseline on the *same* instances. When bonuses are available in the dynamic condition, participants significantly outperform baseline counterparts ($\mu_d = 0.537$, $\mu_b = 0.524$; $t = 2.83$, $p = 0.009$, $d = 0.07$). In contrast, when no bonus is available, no significant difference is observed ($\mu_d = 0.796$, $\mu_b = 0.786$; $t = 1.517$, $p = 0.259$, $d = 0.031$).

This pattern suggests that the observed performance gains are not due to a general motivational effect, such as participants trying harder throughout the tasks, but rather stem from the targeted placement of bonuses. If motivation alone were driving the improvements, we would expect accuracy gains even on non-bonus instances. Instead, the effects appear only where bonuses are present, indicating that the mechanism works by directing human effort to where it is most needed.

Table 3 further contextualizes these findings by presenting weighted accuracy across three performance scenarios: (S1) human and AI perform equally well, (S2) AI outperforms humans, and (S3) humans outperform AI. In S1, both treatments slightly but significantly reduce performance ($d = -0.07$ for both), suggesting unnecessary deliberation. In S2, both improve performance, with dynamic bonuses yielding the greatest gains ($d = 0.13$

Instance Type	Baseline	Static	Dynamic
(S1) Human & AI perform equally	0.613	0.601*	0.599*
(S2) AI outperforms humans	0.814	0.841***	0.871***
(S3) Humans outperform AI	0.499	0.482*	0.573***

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 3. Team accuracy across complementarity conditions.

for static and $d = 0.26$ for dynamic). The most interesting and theoretically important pattern emerges in S3, where static bonuses slightly degrade performance ($d = -0.06$), while dynamic bonuses increase accuracy by 7.5 percentage points ($d = 0.25$)—precisely the instances where human input is most valuable. These instance-level patterns represent the core empirical contribution, in which dynamic incentives successfully redirect human effort toward where it yields the greatest collaborative benefit.

Together, these findings demonstrate that while the static bonus applied uniformly can lead to misallocated effort and reduced overall performance, the dynamic bonus improves accuracy by incentivizing effort where human input is most valuable. By selectively aligning incentives with instance-level difficulty, the dynamic bonus enhances human-AI complementarity. Having established these accuracy effects, we now turn to examine how the different incentive mechanisms influence reliance behavior, specifically, whether bonuses shift participants' willingness to accept or override AI advice.

5.3 Reliance Behavior

Finally, we examine participants' reliance behavior to assess whether our incentive interventions effectively reduced overreliance, as suggested by our theoretical framework. We define reliance as the proportion of instances where participants accepted the AI's suggestion without attempting to solve the problem themselves.

Both treatments led to significant reductions in reliance compared to the baseline. Baseline participants exhibited relatively high reliance on AI advice ($\mu_b = 0.505$), while static bonus participants showed the lowest reliance ($\mu_s = 0.130$; $t = -133.48$, $p < 0.001$, $d = -1.59$), and dynamic bonus participants demonstrated intermediate levels of reliance ($\mu_d = 0.199$; $t = -97.06$, $p < 0.001$, $d = -1.34$) (see Figure 5). These results confirm that both incentive mechanisms reduced reliance on AI, encouraging more independent engagement with the task.

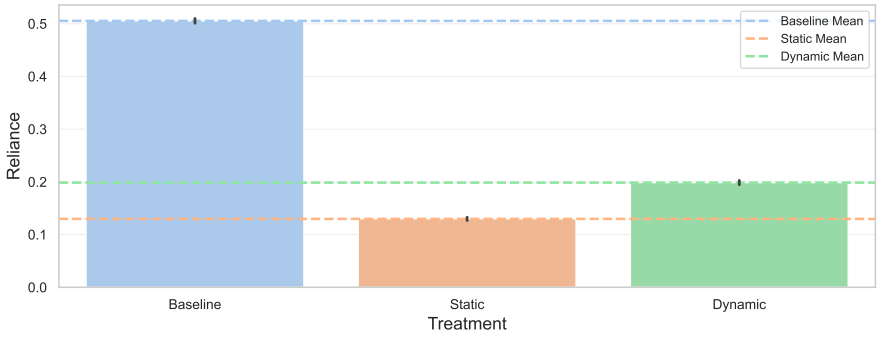

Fig. 5. Weighted average participant reliance across treatments with 95% confidence intervals.

Reliance Mitigation from Bonus Placement. To understand how bonus placement shapes reliance behavior, we first examine whether the reduction in reliance within the dynamic treatment occurs specifically when bonuses are available. Participants showed significantly lower reliance on bonus-eligible instances ($\mu_{bonus} = 0.101$) compared to non-bonus instances ($\mu_{nobonus} = 0.452$; $t = -76.93$, $p < 0.001$, $d = -1.31$). This indicates that reliance reductions are driven by bonus availability, reflecting targeted behavioral adaptation rather than a general treatment effect.

We next examine reliance behavior across instance configurations. Table 4 presents reliance rates for each scenario (S1–S3), showing that both incentive treatments reduced reliance across all instance types. Most importantly, both treatments significantly decreased reliance on instances where humans outperform AI (S3), effectively encouraging human intervention when the AI is likely to be wrong.

However, this shift came at a cost: both treatments also reduced reliance on instances where AI outperforms humans (S2), leading to *underreliance* precisely when following AI advice would have been beneficial. Despite this, our earlier analysis showed that the dynamic bonus still improved overall team performance, suggesting that its targeted nature mitigates this trade-off more effectively than the static approach.

Instance Type	Baseline	Static	Dynamic
(S1) Human & AI perform equally	0.485	0.177^{***}	0.324^{***}
(S2) AI outperforms humans	0.734	0.297^{***}	0.523^{***}
(S3) Humans outperform AI	0.221	0.154^{***}	0.131^{***}

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 4. Reliance rates across complementarity conditions.

Gaming the System: Behavior Distortion. Finally, we observe evidence of arbitrage behavior, where participants opted to "solve" tasks independently but submitted answers that simply mirrored the AI's recommendation, an effortful detour taken solely to collect the bonus. This behavior was most pronounced in the static bonus treatment, where bonuses were awarded indiscriminately, regardless of AI confidence. Participants in this condition exhibited significantly higher rates of such behavior ($\mu_s = 0.454$) than those in the baseline ($\mu_b = 0.173$; $t = 42.80$, $p < 0.001$, $d = 1.68$) or dynamic conditions ($\mu_d = 0.318$; $t = 19.43$, $p < 0.001$, $d = 0.75$), where such gaming provided no advantage.

Qualitative responses further confirm this incentive-driven gaming. Participants explicitly described choosing to "solve" tasks, even when they agreed with the AI, purely to receive the bonus: *"Even if the AI's recommendation was correct, I would tend to solve the task myself in order to receive the bonus." "If I was pretty sure and the AI backed it up, then I would go ahead and take the additional bonus payment."*

These findings reveal a critical downside of the static bonus: by overshooting and rewarding effort indiscriminately, it encourages arbitrary and unnecessary behavior, paradoxically increasing effort without improving outcomes. This underscores the delicate nature of incentive design: to be effective, bonuses must not only encourage engagement but must also be carefully targeted to avoid unintended consequences. The dynamic bonus, by tying rewards to the task-specific value of human input, better preserves this balance.

6 Discussion

Despite the promise of human-AI collaboration, humans often overrely on AI advice, even when their own judgment would lead to better outcomes [60, 72]. In this work, we investigate how incentive structures shape this reliance on AI advice, revealing a fundamental misalignment in the economic incentives underlying human-AI collaboration. Our findings provide empirical validation for dynamic, task-aware incentives as a means to reduce overreliance and promote more effective collaboration.

6.1 Key Findings

Our primary theoretical contribution lies in identifying a previously overlooked structural cause for systematic overreliance by analyzing the framework established by Bansal et al. [4]: the misalignment of incentives in AI-assisted decision-making. While existing research has focused extensively on cognitive factors such as trust and confidence calibration [32, 43] as well as mental models [5, 30, 73], our analysis reveals that the challenge may also be rooted in the economic incentive structures of human-AI collaboration design, positioning incentive misalignment as an additional, complementing driver of overreliance alongside existing cognitive factors [8, 64].

Our behavioral experiment subsequently validates the efficacy of the proposed incentive mechanisms, confirming that both static and dynamic bonuses reduce reliance on AI advice. However, their effects on human-AI team performance diverge. Static bonuses harmed performance by incentivizing independent effort indiscriminately, even when AI advice remains superior, and enabled strategic gaming, in which participants solve tasks independently to collect bonuses while mirroring AI advice. Dynamic bonuses, by contrast, improved performance by focusing incentives on instances where human judgment is most valuable, demonstrating that effective incentive design requires targeting the value of human input rather than independent effort alone.

6.2 Implications

Addressing Overreliance Requires Targeting Structural Causes. Our findings demonstrate that a key driver of overreliance is not inherent to either the human or the AI system, but rather stems from a misalignment in the incentive structures embedded in the system design. When humans are not appropriately incentivized to invest cognitive effort in verifying or overriding AI recommendations, they default to acceptance—even when human judgment would yield better outcomes. This suggests that effective human-AI collaboration cannot be achieved solely through improving AI accuracy or human training [51, 64]. Instead, it requires addressing the economic and structural conditions under which decisions are made. System designers should focus on creating incentive mechanisms that encourage humans to engage their independent judgment selectively and strategically.

Incentives Must Be Context-dependent Rather Than Static. The contrasting outcomes between static and dynamic bonuses highlight the importance of tailoring incentives to instances where human judgment is expected to add the most value. Static bonuses, applied uniformly, created opportunities for gaming, ultimately harming performance. In contrast, dynamic bonuses showed promise by reducing overreliance and improving overall decision quality. However, our findings also show that dynamic bonuses, while more targeted, can also lead to reduced reliance even when the AI is demonstrably superior. This suggests that although context-sensitive incentives are a step in the right direction, further refinements are needed.

Challenges of Designing Optimal Incentives. Even within our simplified framework, designing an optimal bonus mechanism presents several challenges. Key variables such as human effort (λ) and accuracy (P_{H}) are difficult to observe, subjective, and vary across individuals and tasks. This complicates calibration and risks misaligned incentives. Static bonuses may oversimplify, while overly complex dynamic schemes may require reliable real-time estimates that are difficult to acquire. Obtaining well-calibrated AI confidence estimates is a practical prerequisite for the dynamic bonus mechanism, and while established techniques such as temperature scaling or isotonic regression exist for this purpose [76], their effectiveness varies across models and domains, requiring practitioners to verify calibration quality before deployment. Moreover, even well-intentioned bonuses can distort behavior, encouraging strategic gaming or underreliance in unintended situations. However, while we demonstrate that dynamic incentives improve human-AI team performance, we also find that these incentives increase cognitive load, resulting in lower task completion rates. This suggests that more sophisticated incentive mechanisms may have unintended consequences beyond their primary effects, requiring careful evaluation of trade-offs in real-world implementations. These findings underscore that an effective design must balance simplicity, fairness, and adaptability while aligning incentives with the true value of human input.

Deployment and Governance Requirements. Translating our findings into practice requires careful attention to implementation details and governance structures. Organizations must establish transparency mechanisms that inform workers when bonuses are available and how they are determined, enabling informed decisions about effort allocation. This includes transparent communication about the relationship between AI confidence levels and bonus eligibility in dynamic schemes as employed within our study. Additionally, organizations require strategies to identify unintended consequences such as gaming behaviors [14, 37], excessive cognitive burden [34], or systematic disadvantaging of certain worker populations [12, 52]. Worker input into incentive design decisions represents another critical requirement, as ethical deployment demands worker participation in determining acceptable tradeoffs between accuracy, cognitive load, and compensation.

Domain Applicability and Incentive-Compatible System Design. The effects of incentive misalignment vary across domains, and our theoretical model clarifies where overreliance is most problematic. Settings characterized by high cognitive effort costs (λ) or low incentive levels ($\gamma + \beta$) produce the largest misalignment term, meaning rational agents will defer to AI even when their own judgment would yield better outcomes. Medical diagnosis, legal review, and content moderation are particularly concerning in this regard [3, 47], as large caseloads and time pressure keep effort costs high, while performance incentives in these settings may not adequately reflect the marginal value of independent judgment. These domains also tend to have the audit infrastructure needed to implement performance-contingent bonuses, since individual decisions are logged and outcomes become observable over time, providing the basis for awarding θ. By contrast, domains involving subjective quality judgments, real-time decisions under strict latency constraints, or outcomes observable only with long delays present greater implementation challenges, as do traditional employment settings with fixed salary structures. Identifying domains where both the problem and the remedy are most tractable therefore represents a promising direction for future work.

6.3 Ethical Considerations

Our findings raise important questions about the ethics of incentive design in AI-assisted work.

Worker Welfare and Cognitive Burden. While dynamic bonuses improved team performance, they significantly increased cognitive load and reduced task completion rates. This creates a tension that organizations must navigate thoughtfully. On the one hand, further increased cognitive engagement may pose challenges for workers, for example, with cognitive differences [42], or non-native language speakers [1]. On the other hand, incentivizing active cognitive engagement may serve a protective function against skill degradation. Recent evidence suggests that passive reliance on AI systems, particularly generative AI, can lead to deskilling and erosion of expertise over time [38, 70]. By encouraging workers to maintain cognitive involvement in their work, incentive mechanisms may support the long-term preservation and development of human capabilities [15]. The key question becomes not whether to impose cognitive demands, but how to design incentives that balance skill maintenance and task performance with worker capacity and well-being.

Worker Agency and Distributive Fairness. If dynamic bonuses increase organizational effectiveness while increasing workers' cognitive load, fair implementation requires ensuring that workers are adequately compensated for increased cognitive demands. In this context, incentive mechanisms may differentially affect worker populations. Moreover, bonuses tied to task difficulty require workers to accurately assess when their judgment adds value—a metacognitive skill that varies across individuals [18]. For example, workers with less experience, AI literacy, or self-efficacy may struggle to consistently perform accurate self-assessments [16].

Performance-Centric Framing. Our study focused on performance outcomes but did not assess other factors such as worker autonomy [46], job satisfaction [27], or long-term well-being [44]. This focus shaped our core framing, where we characterized excessive reliance on AI advice as problematic when human judgment would be more accurate. While accuracy maximization is a reasonable primary goal, the acceptable trade-off between

accuracy and other goals varies across contexts. In lower-stakes contexts, workers may legitimately prefer some cognitive offloading even at modest accuracy costs. Conversely, in high-stakes domains like healthcare [47] or criminal justice [3], accuracy and fairness must remain paramount regardless of worker preferences.

6.4 Limitations and Directions for Future Work

Our study shows that incentive structures systematically influence reliance behavior in human-AI collaboration. By isolating this mechanism in a controlled setting, we demonstrate a causal relationship between incentive design and the mitigation of overreliance. While this fundamental work was necessarily constrained by some factors that limit generalizability, these factors simultaneously open several promising directions for future research that can build upon and extend our findings.

Extending Across Domains and Stakes. We deliberately selected an image classification task that requires no specialized expertise, allowing us to isolate the effect of incentive mechanisms from domain-specific confounds. This methodological choice, common in behavioral research [19], aims to ensure that observed effects stem from the incentive structure itself rather than task-specific confounds. Having established this baseline effect, future research should examine how incentive mechanisms operate across different domains and stakes. Professional decision-making contexts such as healthcare, criminal justice, or financial services involve specialized knowledge and serious consequences for errors [3, 47]. These factors may amplify or attenuate the effects we observe, and understanding these contextual moderators represents an important next step.

Sample Characteristics. While we aimed for balanced representation in terms of gender and age through our recruitment criteria, our Prolific sample may not represent workplace populations across other dimensions such as educational backgrounds, cognitive abilities, or professional experience. Workers in organizational settings face different motivational structures, such as career advancement [7], peer relationships [21], or intrinsic professional commitment [20, 46], which our monetary incentives do not capture. Future research should investigate how these factors influence responses to incentive mechanisms in organizational contexts and explore how personalizing incentive mechanisms based on these factors affects reliance in human-AI collaboration.

Longitudinal Effects and Adaptation. Our short-term experimental session establishes immediate behavioral responses to incentive mechanisms. However, sustained exposure may produce adaptation effects, skill development or atrophy, or shifting motivations. Longitudinal studies examining whether initial behavioral changes persist, fade, or evolve over time would provide critical insights for practical implementation [67]. Such work could reveal whether incentive mechanisms require periodic recalibration, whether workers develop more sophisticated strategies for effort allocation, or whether prolonged exposure leads to habituation effects that diminish treatment efficacy.

Combining Incentive and Information-Based Interventions. Our work demonstrates that incentive mechanisms can reduce overreliance by addressing structural misalignment in human-AI collaboration. However, prior research has shown that information-based interventions such as explanations [11], confidence calibration [43], or specialized training [51, 64] can also improve reliance behavior. Future research should investigate whether combining incentive mechanisms with these approaches yields synergistic effects.

Theoretical Framework Assumptions. Our formal model assumes rational agents who maximize expected utility through a linear combination of rewards, penalties, bonuses, and effort costs. This simplification isolates incentive misalignment as a driver of overreliance, consistent with Bansal et al. [4], who themselves acknowledge that humans are not purely rational. For example, cognitive biases [8] or varying risk preferences [79] may produce systematic deviations. Additionally, the relationships between γ, β, θ, and λ may interact in more complex, non-linear ways that our formulation does not capture, and future work should explore more complex utility formulations accounting for diminishing returns or interaction effects between effort costs and incentive magnitude.

Boundary Conditions of Dynamic Incentives. Our dynamic bonus mechanism relies on calibrated AI confidence as a practical proxy for the expected value of human effort, which remains observable without ground-truth human accuracy estimates. However, this proxy may be less reliable in settings where human and AI capabilities are unknown or rapidly shifting. Future work should examine how dynamic incentives perform under such conditions, including settings with multiple human decision-makers whose capabilities vary. Additionally, the confidence threshold distinguishing bonus-eligible from non-eligible instances was derived a priori from the theoretical framework, leaving room for future work to examine the sensitivity of the results to threshold variations across domains and AI systems.

7 Conclusion

We investigate how incentive structures influence overreliance in traditional human-AI collaboration setups and propose an alternative mechanism design to address systemic bias toward overreliance on AI advice. Our theoretical analysis uncovers misaligned incentive structures as a structural driver of overreliance. Our behavioral experiment with 180 participants then demonstrates that dynamic incentives simultaneously reduce overreliance and improve human-AI team performance. In contrast, static incentives foster strategic behavior that undermines their intended purpose. While our findings are demonstrated in image classification, the identified bias term may operate differently across domains with varying cognitive effort requirements and decision stakes. Future research should investigate how these mechanisms operate across diverse domains and populations, develop adaptive approaches that balance performance with worker welfare, and explore governance structures that ensure incentive designs serve both organizational goals and worker well-being.

References

[1] Michaela Albl-Mikasa, Maureen Ehrensberger-Dow, Andrea Hunziker Heeb, Caroline Lehr, Michael Boos, Matthias Kobi, Lutz Jäncke, and Stefan Elmer. 2020. Cognitive load in relation to non-standard language input: Insights from interpreting, translationand neuropsychology. *Translation, Cognition & Behavior* 3, 2 (2020), 263–286.

[2] Yotam Amitai, Yael Septon, and Ofra Amir. 2024. Explaining Reinforcement Learning Agents through Counterfactual Action Outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 9 (2024), 10003–10011.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[7] Monique Bernadette van Rijn, Huadong Yang, and Karin Sanders. 2013. Understanding employees' informal workplace learning: The joint influence of career motivation and self-construal. *Career development international* 18, 6 (2013), 610–628.

[8] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91.

[9] Sarah E Bonner and Geoffrey B Sprinkle. 2002. The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, Organizations and Society* 27, 4 (2002), 303–345.

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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 (2021), 1–21.

[11] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[12] Maarten Buyl, Christina Cociancig, Cristina Frattone, and Nele Roekens. 2022. Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1071–1082.

[13] Yuyao Chen, Zhengtang Zhang, Jinfan Zhou, Chuwei Liu, Xia Zhang, and Ting Yu. 2023. A cognitive evaluation and equity-based perspective of pay for performance on job performance: A meta-analysis and path model. *Frontiers in Psychology* 13 (2023), 1039375.
[14] Pascal Courty and Gerald Marschke. 2004. An empirical investigation of gaming responses to explicit performance incentives. *Journal of Labor Economics* 22, 1 (2004), 23–56.
[15] K Anders Ericsson. 2004. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic medicine* 79, 10 (2004), S70–S81.
[16] Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. 2025. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior* (2025), 108779.
[17] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 489–503.
[18] John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34, 10 (1979), 906.
[19] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Mis Quarterly* 45, 3 (2021), 1527–1556.
[20] Barry Gerhart and Meiyu Fang. 2015. Pay, intrinsic motivation, extrinsic motivation, performance, and creativity in the workplace: Revisiting long-held beliefs. *Annu. Rev. Organ. Psychol. Organ. Behav.* 2, 1 (2015), 489–521.
[21] Adam M Grant and Marissa S Shandell. 2022. Social motivation at work: The organizational psychology of effort for, against, and with others. *Annual review of psychology* 73, 1 (2022), 301–326.
[22] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 221–236.
[23] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
[24] Joachim Hartung, Guido Knapp, and Bimal K Sinha. 2011. *Statistical meta-analysis with applications*. John Wiley & Sons.
[25] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2025. Complementarity in human-AI collaboration: concept, sources, and evidence. *European Journal of Information Systems* 0, 0 (2025), 1–24. doi:10.1080/0960085X.2025.2475962
[26] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* 78 (2021), 118.
[27] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 453–463.
[28] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. 419–429.
[29] Bengt Holmstrom and Paul Milgrom. 1991. Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization* 7 (1991), 24–52.
[30] Joshua Holstein and Gerhard Satzger. 2025. Development of Mental Models in Human-AI Collaboration: A Conceptual Framework. In *Proceedings of the 46th International Conference on Information Systems (ICIS) (ICIS 2025 Proceedings, 2)*. Association for Information Systems (AIS), Nashville, USA.
[31] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.
[32] Patricia K. Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. 2024. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice.. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 609–622.
[33] Patricia K Kahr, Gerrit Rooks, Martijn C Willemsen, and Chris CP Snijders. 2023. It seems smart, but it acts stupid: Development of trust in ai advice in a repeated legal decision-making task. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 528–539.
[34] Virpi Kalakoski, Sanna Selinheimo, Teppo Valtonen, Jarno Turunen, Sari Käpykangas, Hilkka Ylisassi, Pauliina Toivio, Heli Järnefelt, Heli Hannonen, and Teemu Paajanen. 2020. Effects of a cognitive ergonomics workplace intervention (CogErg) on cognitive strain and well-being: a cluster-randomized controlled trial. A study protocol. *BMC psychology* 8, 1 (2020), 1.
[35] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.
[36] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint*

arXiv:2506.08872 (2025).
[37] Ian Larkin. 2014. The cost of high-powered incentives: Employee gaming in enterprise software sales. *Journal of Labor Economics* 32, 2 (2014), 199–227.
[38] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–22.
[39] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2024. Decoding AI's Nudge: A Unified Framework to Predict Human Behavior in AI-Assisted Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 9 (2024), 10083–10091.
[40] Huigang Liang, Meng-Meng Wang, Jian-Jun Wang, and Yajiong Xue. 2018. How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: A mediated moderation model. *Computers in Human behavior* 81 (2018), 168–176.
[41] Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. 2025. Human and AI Perceptual Differences in Image Classification Errors. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 13 (2025), 14318–14326.
[42] Rosemary Lysaght, Lynn Shaw, Andrea Almas, Anita Jogia, and Sherrey Larmour-Trode. 2008. Towards improved measurement of cognitive and behavioural work demands. *Work* 31, 1 (2008), 11–20.
[43] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are you really sure?" Understanding the effects of human self-confidence calibration in AI-assisted decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.
[44] Anne Mäkikangas, Ulla Kinnunen, Taru Feldt, and Wilmar Schaufeli. 2016. The longitudinal development of employee well-being: A systematic review. *Work & stress* 30, 1 (2016), 46–70.
[45] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan Schwamb, Chris Lintott, and Arfon Smith. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 1. 94–102.
[46] Youyan Nie, Bee Leng Chua, Alexander Seeshing Yeung, Richard M Ryan, and Wai Yen Chan. 2015. The importance of autonomy support and the mediating role of work motivation for well-being: Testing self-determination theory in a Chinese work organisation. *International journal of psychology* 50, 4 (2015), 245–255.
[47] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[48] Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 546–561.
[49] Chirag Patel, Mariyani Ahmad Husairi, Christophe Haon, and Poonam Oberoi. 2023. Monetary rewards and self-selection in design crowdsourcing contests: Managing participation, contribution appropriateness, and winning trade-offs. *Technological Forecasting and Social Change* 191 (2023), 122447.
[50] Kenny Peng, Nikhil Garg, and Jon Kleinberg. 2025. A No Free Lunch Theorem for Human-AI Collaboration. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 13 (2025), 14369–14376.
[51] Marc Pinski, Martin Adam, and Alexander Benlian. 2023. AI knowledge: Improving AI delegation through human enablement. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–17.
[52] Robert Lee Poe and Soumia Zohra El Mestari. 2024. The conflict between algorithmic fairness and non-discrimination: An analysis of fair automated hiring. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1907–1916.
[53] Canice Prendergast. 1999. The provision of incentives in firms. *Journal of economic literature* 37, 1 (1999), 7–63.
[54] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2023. A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 11. 127–139.
[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
[56] Lara Riefle, Patrick Hemmer, Carina Benz, and Michael Vössing. 2024. The Role of Cognitive Styles for Explainable AI. In *32nd European Conference on Information Systems (ECIS)*. Association for Information Systems (AIS).
[57] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115 (2015), 211–252. Issue 3.
[59] Shlomo S Sawilowsky and R Clifford Blair. 1992. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin* 111, 2 (1992), 352.
[60] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and*

Society. 617–626.

[61] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422.

[62] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Looks Can Mislead: Feature-Based Explanations Do Not Reliably Improve Fairness of AI-Assisted Decisions. *Available at SSRN 5029746* (2024).

[63] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 275–284.

[64] Philipp Spitzer, Joshua Holstein, Patrick Hemmer, Michael Vössing, Niklas Kühl, Dominik Martin, and Gerhard Satzger. 2025. Human Delegation Behavior in Human-AI Collaboration: The Effect of Contextual Information. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–28.

[65] Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger, and Niklas Kühl. 2024. Don't be Fooled: The Misinformation Effect of Explanations in Human-AI Collaboration. *arXiv preprint arXiv:2409.12809* (2024).

[66] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2495–2507.

[67] Mark Steyvers and Aakriti Kumar. 2024. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science* 19, 5 (2024), 722–734.

[68] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

[69] Dan N. Stone and David A. Ziebart. 1995. A Model of Financial Incentive Effects in Decision Making. *Organizational Behavior and Human Decision Processes* 61, 3 (1995), 250–261.

[70] Lev Tankelevitch, Elena L. Glassman, Jessica He, Majeed Kazemitabaar, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, and Hari Subramonyam. 2025. Tools for Thought: Research and Design for Understanding, Protecting, and Augmenting Human Cognition with Generative AI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 804, 8 pages. doi:10.1145/3706599.3706745

[71] John Wilder Tukey et al. 1977. *Exploratory data analysis*. Vol. 2. Springer.

[72] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303.

[73] Karel van den Bosch, Emma van Zoelen, Tjeerd Schoonderwoerd, Anthia Anthia Solaki, Birgit Van der Stigchel, and Ivana Akrum. 2025. Design and Effects of Co-Learning in Human-AI Teams. *Journal of Artificial Intelligence Research* 82 (2025), 1445–1493.

[74] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations. *Artificial Intelligence* 291 (2021), 1–19.

[75] Victor Vroom, Lyman Porter, and Edward Lawler. 2015. Expectancy theories. In *Organizational behavior 1*. Routledge, 94–113.

[76] Cheng Wang. 2023. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222* (2023).

[77] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022*. 1697–1708.

[78] Suqing Wu, Yukun Liu, Mengqi Ruan, Siyu Chen, and Xiao-Yun Xie. 2025. Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports* 15, 1 (2025), 15105.

[79] Cathy Yang, Kevin Bauer, Xitong Li, and Oliver Hinz. 2026. My advisor, her AI, and me: Evidence from a field experiment on human–AI collaboration and investment decisions. *Management Science* 72, 1 (2026), 242–264.

[80] Ming Yin and Yiling Chen. 2015. Bonus or not? learn to reward in crowdsourcing.. In *IJCAI*. 201–208.

[81] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

Appendix

This appendix provides detailed methodological specifications and supplementary analyses supporting our main findings. We present comprehensive implementation details for our behavioral experiment, including AI system training, instance selection criteria, and incentive mechanism derivations, followed by extended statistical analyses and robustness checks. These materials ensure experimental replicability in our analytical approach, particularly regarding our inverse variance weighting methodology for handling heterogeneous measurement precision across treatments.

Experimental Design

AI System Training Details. We implemented a DenseNet-161 architecture pre-trained on ImageNet. The dataset was split into 60% training, 10% validation, and 20% test sets. The model was fine-tuned on distorted images over 40 epochs using SGD optimizer with a learning rate of $1 \cdot 10^{-3}$, weight decay of $5 \cdot 10^{-4}$, and cosine annealing learning rate scheduler. We used a batch size of 32 and applied early stopping on the validation loss. The AI system achieved an accuracy of 67.5% on the test set.

As we later provide the binned confidence rating of the AI system to the humans, we aimed for a calibrated confidence design. Specifically, we binned the confidence in four levels ranging from very low (0-0.25), over low (0.25-0.5) and high (0.5-0.75) to very high (0.75-1).

Instance Selection. Effective evaluation of human-AI collaboration requires understanding when each partner should contribute to decision-making. In real-world scenarios, the relative capabilities of humans and AI systems vary across different task instances, creating distinct strategic considerations for optimal collaboration. When both partners perform similarly on a task instance, efficiency considerations favor following AI advice due to reduced cognitive effort costs. When AI demonstrates superior performance, optimal collaboration requires humans to rely on AI advice. However, at instances where humans outperform AI, humans should invest effort to solve task instances independently to achieve better outcomes rather than (over)relying on AI advice.

Given these theoretical foundations, we strategically selected instances to represent these three complementary scenarios: (S1) instances where both humans and AI perform similarly, (S2) instances where AI outperforms humans, and (S3) instances where humans outperform AI. Our goal was to construct a balanced dataset enabling the evaluation of how different incentive mechanisms perform across the full spectrum of realistic scenarios of human-AI collaboration.

To operationalize these theoretical categories, we leveraged two complementary indicators: human annotator agreement patterns from Steyvers et al. [68] and calibrated confidence ratings:

Annotator Agreement. We reasoned that human annotator disagreement serves as a proxy for human probability for solving a task correctly with high disagreement indicating instances that humans find challenging, while low disagreement suggests instances where humans can reliably perform well.

Calibrated Confidence. We used temperature scaling to calibrate the confidence of the AI. In detail, we optimized the temperature parameter T on the validation set and divided the logits by the selected parameter prior to the softmax to infer the predictions on the test set. This allowed us to target accuracy rates of AI predictions that correspond to the midpoint of each confidence bin: "very low" confidence (0-0.25 confidence range, 12.5% target accuracy), "low" confidence (0.25-0.5 confidence range, 37.5% target accuracy), "high" confidence (0.5-0.75 confidence range, 62.5% target accuracy), and "very high" confidence (0.75-1.0 confidence range, 87.5% target accuracy).

Building on this logic, we categorized instances by selecting instances with high human annotator disagreement (a minimum of 4 out of 6 annotators disagreeing) compared to instances with strong human consensus (a maximum of 3 out of 6 annotators disagreeing). Additionally, instances can be classified based on whether the AI was correct/incorrect in its predictions. Strategically sampling from these options, allows us to select instances

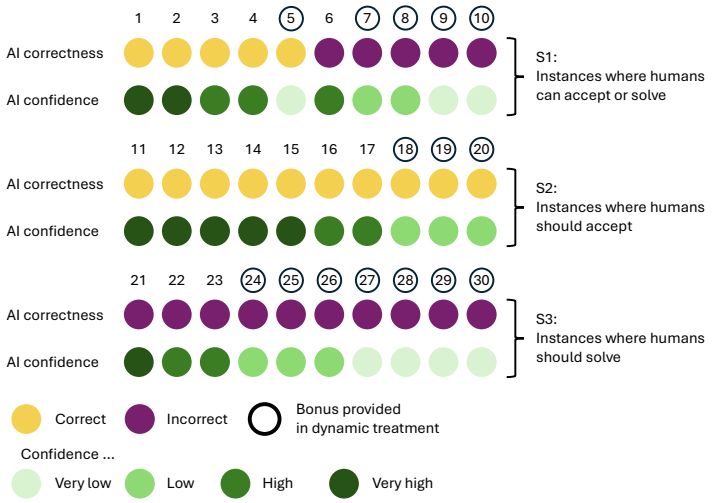

Fig. 6. Distribution of instances with confidences to complementary subsets.

for S1-S3, ensuring that our instance selection aligns with the underlying capability differences that define each complementarity scenario. Figure 6 demonstrates our instance selection, including how confidence levels distribute across our three complementarity subsets, confirming the coherence of our selection methodology.

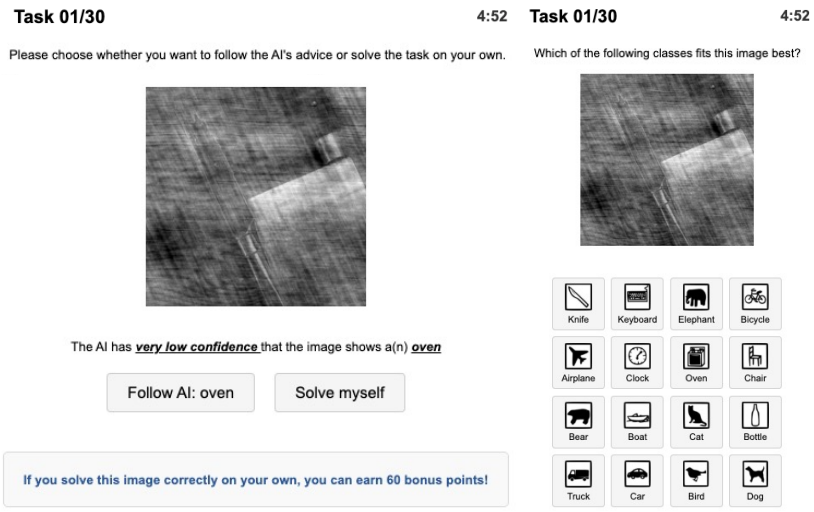

Fig. 7. Task interface showing the main decision screen (left) and the independent classification screen (right). The bonus notification (shown in blue) was absent in the baseline condition. In the static condition, a fixed bonus of 30 points was displayed for all instances; in the dynamic condition, the bonus varied by instance—60 points for low AI confidence instances and no bonus for high AI confidence instances—directing human effort to where it was most valuable, with every 10 points translating to £0.01.

Incentive Mechanism Design. In the following, we outline how we determined the incentive mechanisms across treatments. The overall mechanism was designed to maintain equivalent maximum payouts across all treatments while systematically varying the distribution between rewards (γ) and independent decision-making bonuses (θ).

The maximum variable performance-based payment for each participant can be expressed as:

$$\text{Var Payment}_{\text{max}} = \sum_{i=1}^{n} (\gamma + \theta_i \cdot \mathbb{I}[\text{solve}_i])$$

where $\mathbb{I}[\text{solve}_i]$ is an indicator function equal to 1 if the participant chooses to solve instance i independently and answers correctly, and 0 otherwise.

Based on pilot testing, we anticipated a study duration of approximately 10 minutes, leading to a base participation payment of £1.00 to ensure fair compensation above minimum wage standards. Additionally, our theoretical framework necessitates performance-based incentives to motivate human effort in decision-making. To create meaningful incentives for task engagement while maintaining experimental feasibility, we set the maximum performance-based payment at £1.80, resulting in a maximum total compensation of £2.80 for an estimated task duration of 10 minutes, which corresponds to a high remuneration according to Prolific standards. Following, we outline the theoretical derivation of the distribution of the performance-based payment across (γ) and (θ).

Baseline. In traditional AI-assisted decision-making settings, participants receive rewards based solely on final decision accuracy, regardless of whether they rely on AI advice or solve independently. This represents the standard incentive structure that our theoretical analysis identifies as systematically biased toward overreliance. With $\theta = 0$, rearranging the maximum variable payout results in:

$$\gamma_{\text{b}} = \frac{\text{Var Payment}_{\text{max}}}{n_{\text{instances}}} = \frac{£1.80}{30} = £0.06$$

Static. For the static bonus setting, we construct two comparable scenarios to align the incentives for two extreme behavioral strategies: (E1) solving all instances independently and (E2) always following AI advice. Our goal is to ensure that a rational decision-maker would be indifferent between these two strategies in expectation. To make the comparison meaningful, we incorporate time constraints reflective of our experimental setup. Let t_{max} denote the total time allowed for the task (300 seconds in our experiment), and $t_{\text{per instance}}$ the average time required for a human to solve one instance independently (20 seconds, based on Hemmer et al. [25]). Additionally, let p_{H} denote the average human accuracy and p_{AI} the average AI accuracy, both computed against ground-truth labels from empirical data from Steyvers et al. [68].

For humans always rejecting AI advice and solving all instances independently (E1), we receive:

$$\text{Performance}_{\text{E1}} = \frac{t_{\text{max}}}{t_{\text{per instance}}} \cdot p_{\text{H}}$$

For humans always following AI advice without independent solving (E2):

$$\text{Performance}_{\text{E2}} = n_{\text{instances}} \cdot p_{\text{AI}}$$

Setting equal expected payouts between strategies:

$$(\gamma_s + \theta_s) \cdot \text{Performance}_{\text{E1}} = \gamma_s \cdot \text{Performance}_{\text{E2}}$$

$$\theta_s = \gamma_s \cdot \left(\frac{\text{Performance}_{\text{E2}}}{\text{Performance}_{\text{E1}}} - 1 \right)$$

Substituting our experimental parameters ($t_{\text{max}} = 300$ seconds, $t_{\text{per instance}} = 20$ seconds, and $p_{\text{H}} = p_{\text{AI}} = 0.5$ based on our balanced instance selection) results in:

- $\text{Performance}_{\text{E1}} = 300s/20s \cdot 0.5 = 7.5$ instances
- $\text{Performance}_{\text{E2}} = 30 \cdot 0.5 = 15$ instances

This yields the equivalence of $\theta_{\text{s}} = \gamma_{\text{s}}$. Given the maximum payment constraint, we receive:

$$£1.80 = 30 \cdot (\gamma_{\text{s}} + \theta_{\text{s}})$$
$$\gamma_{\text{s}} = \theta_{\text{s}} = £0.03$$

Dynamic. The dynamic condition strategically allocates bonuses to instances where human judgment provides the highest expected utility. Unlike the static bonus setting with a single bonus level, we implement a dual-bonus structure with two different bonus amounts ($\theta_{\text{d, 1}}$ and $\theta_{\text{d, 2}}$) applied conditionally based on AI confidence levels, while maintaining overall payment equivalence across treatments.

The performance-based payment structure becomes:

$$\text{Var Payment}_{\text{max}} = \sum_{i=1}^{n} (\gamma_d + \theta_{d,1,i} \cdot \mathbb{I}[\text{solve}_i \wedge p_{\text{AI},i} \geq 0.5] + \theta_{d,2,i} \cdot \mathbb{I}[\text{solve}_i \wedge p_{\text{AI},i} < 0.5])$$

For the dynamic treatment, we preserve equivalent maximum variable payment relative to the static condition ($\sum_{i=1}^{30}(\gamma + \theta_i) = £1.80$), with instances distributed such that half satisfy $p_{\text{AI}} < 0.5$ and half satisfy $p_{\text{AI}} \geq 0.5$. To focus the entire bonus budget on instances where human input is most valuable, we set $\theta_{\text{d, 1}} = £0.00$ for high-confidence instances ($p_{\text{AI}} \geq 0.5$) and maintain $\gamma_{\text{d}} = £0.03$ (same as static condition).

$$£1.80 = 30 \cdot \gamma_{\text{d}} + 15 \cdot \theta_{\text{d, 1}} + 15 \cdot \theta_{\text{d, 2}}$$

Substituting $\gamma_{\text{d}} = £0.03$ and $\theta_{\text{d, 1}} = £0.00$:

$$\theta_{\text{d, 2}} = £0.06$$

This yields:

$$\gamma_{\text{d}} = £0.03$$
$$\theta_{\text{d, 2}} = £0.06 \quad \text{(low AI confidence: } p_{\text{AI}} < 0.5\text{)}$$
$$\theta_{\text{d, 1}} = £0.00 \quad \text{(high AI confidence: } p_{\text{AI}} \geq 0.5\text{)}$$

This effectively doubles the incentive for independent effort when human judgment adds the most value.

This design creates an hourly compensation range of £11.40-£13.80 based on individual performance and task completion rates, ensuring fair compensation while maintaining strong incentives for strategic decision-making across all experimental conditions. This range is derived from our empirical dataset analysis and instance selection design. Based on the collected labels [68], we expect humans to correctly solve approximately 50% of instances independently, establishing the lower bound of our compensation structure. Under the different payment structures, this baseline human performance translates to an average minimum payment of £11.40 per hour. The upper bound (£13.80) reflects human-AI complementarity, where participants can potentially solve 25 out of 30 instances correctly. This scenario accounts for the 5 instances in our dataset where neither humans nor AI systems typically perform well.

Experimental Results

In the following, we elaborate our testing methodology in more detail.

Testing Methodology. Due to identified systematic differences in task completion rates across treatments, participants provided varying numbers of observations, creating heterogeneous measurement precision across our two primary outcome measures: human-AI team performance, and reliance behavior. Participants completing fewer instances yield less reliable estimates due to smaller sample sizes, violating the equal-variance assumption underlying standard analytical approaches.

Following Hartung et al. [24], we employed *inverse variance weighting* to address this analytical challenge, i.e., assigning greater weight to observations with higher precision (lower variance) and reduced weight to less reliable measurements (higher variance). For each participant i, we calculated weights based on the inverse of their estimated measurement variance:

$$w_i = \frac{1}{\sigma_i^2}$$

The specific variance calculation depends on the outcome measure, i.e., human-AI team performance and reliance. As both are binary outcomes, both measures, when aggregated, follow a binomial distribution. While instance difficulties vary, randomized presentation ensures that each participant encounters a representative sample of instances, allowing us to model their overall performance with success parameter p_i as a binomial distribution. This results in:

$$w_i = \frac{n_i}{p_i(1-p_i)}$$

where n_i represents the number of instances classified by participant i and p_i is the proportion of instances where the participant followed AI advice. For performance measures, we calculated sample variance based on individual accuracy across completed instances.

This weighting scheme creates computational difficulties when participants exhibit perfect reliance behavior (either $p_i = 0$ or $p_i = 1$), as the denominator $p_i(1-p_i)$ approaches zero, resulting in infinite weights. To address this issue while preserving the high information value, we assigned participants with either $p_i = 0$ or $p_i = 1$ a weight equal to 125% of the maximum finite weight observed among other participants within the same treatment group. Results were robust to alternative scaling factors (100%-200% of maximum finite weight). This approach ensures these participants receive appropriately high but finite influence in the analysis.

To improve robustness and mitigate the influence of extreme weights, we applied winsorization [71] at the 5th and 95th percentiles to the weight distributions within each treatment group before conducting comparative analyses. This procedure caps extreme weights while preserving the overall weighting structure, preventing any single participant from disproportionately influencing results.

Statistical Testing. All comparative analyses for human-AI team performance and reliance rates employed two-sided Welch's t-tests with inverse variance weighting to account for heterogeneous task completion rates across participants unless otherwise specified. Given our sample size of 60 participants per treatment condition, we proceeded with t-tests without formal normality testing, leveraging the robustness of t-tests to departures from normality, particularly with moderate to large sample sizes [59].

Multiple Comparisons Correction. To control for Type I error inflation arising from multiple treatment comparisons against the baseline condition, we applied Bonferroni correction to all p-values. Specifically, for each outcome measure, we conducted two primary comparisons (Static vs. Baseline and Dynamic vs. Baseline), resulting in an adjusted p-values for individual tests.

Task Completion. Task completion rates differed across treatments, with static bonus participants completing the most instances ($\mu_s = 26.933$), followed by baseline ($\mu_b = 25.350$), and dynamic bonus participants ($\mu_d = 23.633$). Shapiro-Wilk tests indicated non-normal distributions across all conditions ($p < 0.001$), while Levene's test revealed heterogeneous variances ($p = 0.012$).

Kruskal-Wallis tests confirmed significant differences in completion rates across conditions ($H = 7.505$, $p = 0.023$). One-sided Kolmogorov-Smirnov tests revealed no significant distributional shift for static vs. baseline ($D = 0.117$, $p = 0.407$), a trend for dynamic vs. baseline ($D = 0.200$, $p = 0.091$), and a significant shift between static vs. dynamic conditions ($D = 0.233$, $p = 0.038$), with static participants completing approximately 14% more instances than dynamic participants.

Mediation Analysis. We conducted mediation analysis to examine whether cognitive load mediates the relationship between treatments and task completion, using bootstrap analysis with 5000 simulations (Table 5). The analysis demonstrates a significant Average Causal Mediation Effect (ACME = -0.441, 95% CI [-1.129, 0.000], p = 0.048), with cognitive load accounting for approximately 26% of the dynamic treatment effect, indicating that dynamic bonuses operate substantially through increased cognitive burden.

Human-AI Team Performance. Performance analyses employed the same inverse variance weighting methodology for both overall treatment comparisons across the full dataset (Table 6) and subset analyses across three complementarity scenarios (S1-S3, Table 8).

Influence of Bonus Availability on human-AI team performance. To examine the targeted effects of bonus placement, we conducted conditional analyses comparing dynamic treatment participants against baseline participants on identical instance subsets (Table 7). Specifically, we partitioned instances based on bonus availability in the dynamic condition and compared performance between treatments within each partition. This within-instance comparison controls for instance-specific difficulty effects and isolates the impact of bonus availability.

Reliance Behavior. Reliance analyses employed the same inverse variance weighting methodology for overall treatment comparisons across the full dataset (Table 9) and subset analyses across complementarity scenarios (S1-S3, Table 11).

Influence of Bonus on Reliance Behavior. To examine whether bonus availability directly influences reliance behavior, we conducted within-subjects analyses comparing dynamic treatment participants' reliance on bonus-eligible versus non-bonus instances (Table 10). This paired design controls for individual differences while isolating the effect of bonus availability. For participants with different completion patterns across bonus and non-bonus instances, weights were combined using the harmonic mean.

Arbitrage Behavior Analysis. We also examined instances where participants chose to solve independently but submitted answers matching AI recommendations—a form of strategic gaming to collect bonuses without genuine independent effort. Specifically, we compared static treatment against both baseline and dynamic treatment followed by appropriate correction of p-values (Table 12).

Dependent variable	Cognitive Load		Task Completion	
	coeff	se	coeff	se
Intercept	4.178***	0.128	29.920***	2.157
Treatment:				
- *Baseline*				
- *Static*	0.019	0.181	1.605	1.152
- *Dynamic*	0.403*	0.181	-1.276	1.168
Cognitive Load			-1.094*	0.478
R^2	0.034		0.071	
F-statistic	3.146*		4.483**	

	Bootstrap Mediation Effects (5000 simulations)			
	Effect	Boot SE	Boot LLCI	Boot ULCI
Dynamic versus Baseline:				
ACME	-0.441*	0.288	-1.129	0.000
ADE	-1.276	1.255	-3.702	1.210
Total Effect	-1.717	1.247	-4.114	0.760
Prop. Mediated	0.257	0.983	-1.541	2.250

Note: *$p < .05$; ** $p < .01$; *** $p < .001$.

Table 5. Mediation analysis: Effect of incentive structures on efficiency through cognitive load.

Comparison	**Treatment Mean**	**Baseline Mean**	**t**	**p**	**d**
Static vs. Baseline	0.632	0.643	-6.008	<0.001	-0.101
Dynamic vs. Baseline	0.653	0.643	5.042	<0.001	0.087

Table 6. Overall treatment effects on human-AI team performance.

Condition	**Dynamic Mean**	**Baseline Mean**	**t**	**p**	**d**
Bonus Available	0.537	0.524	2.832	0.009	0.069
No Bonus Available	0.796	0.786	1.517	0.259	0.031

Table 7. Effect of bonus availability on human-AI team performance.

Instance Type	Baseline	Static	Dynamic	Static vs. Baseline			Dynamic vs. Baseline		
	Mean	Mean	Mean	t	p	d	t	p	d
(S1) Human & AI perform equally	0.613	0.601	0.599	-2.371	0.036	-0.065	-2.330	0.040	-0.066
(S2) AI outperforms humans	0.814	0.841	0.871	6.137	<0.001	0.127	12.788	<0.001	0.264
(S3) Humans outperform AI	0.499	0.482	0.573	-2.305	0.042	-0.059	9.692	<0.001	0.251

Table 8. Human-AI team performance across complementarity conditions.

Comparison	Treatment Mean	Baseline Mean	t	p	d
Static vs. Baseline	0.130	0.505	-133.48	<0.001	-1.586
Dynamic vs. Baseline	0.199	0.505	-97.06	<0.001	-1.336

Table 9. Overall treatment effects on reliance behavior.

Condition	With Bonus Mean	Without Bonus Mean	t	p	d
Dynamic Treatment	0.101	0.452	-76.93	<0.001	-1.309

Note: Comparison within dynamic treatment participants.

Table 10. Effect of bonus availability on reliance behavior.

Instance Type	Baseline	Static	Dynamic	Static vs. Baseline			Dynamic vs. Baseline		
	Mean	Mean	Mean	t	p	d	t	p	d
(S1) Human & AI perform equally	0.485	0.177	0.324	-47.31	<0.001	-1.160	-21.04	<0.001	-0.519
(S2) AI outperforms humans	0.734	0.297	0.523	-63.59	<0.001	-1.356	-27.50	<0.001	-0.635
(S3) Humans outperform AI	0.221	0.154	0.131	-13.14	<0.001	-0.273	-20.32	<0.001	-0.447

Table 11. Reliance behavior across complementarity conditions.

Comparison	Static Mean	Comparison Group Mean	t	p	d
Static vs. Baseline	0.454	0.173	42.80	<0.001	1.682
Static vs. Dynamic	0.454	0.318	19.43	<0.001	0.749

Table 12. Strategic gaming (arbitrage) behavior: Static treatment comparisons.

Acknowledgements

Generative AI tools were utilized throughout this work. Specifically, Claude and Github Copilot were employed to generate code for visualizations. Additionally, ChatGPT, DeepL Write, and Grammarly were used to enhance the writing quality of tutorials and explanations provided to participants during the experiments, as well as to improve the language throughout this paper.