NET WORK LISER

Discussion Paper Series

IZA DP No. 18678

May 2026

Algorithm Aversion in Prosocial Tasks: Evidence from AI-Based Performance Evaluation

Martin Abel
Bowdoin College, J-PAL and IZA@LISER

Raghad S. Dawi
Bowdoin College

Tyler Lenk
Bowdoin College

Aidan Singer
Bowdoin College

ISSN: 2365-9793

IZA DP No. 18678 May 2026

Algorithm Aversion in Prosocial Tasks: Evidence from AI-Based Performance Evaluation*

Abstract

How do workers respond when artificial intelligence replaces human judgment in evaluating prosocial work? Partnering with a non-profit addressing food insecurity, we recruit 1,491 U.S. volunteers to write fundraising messages and cross-randomize evaluation by humans versus AI and the presence of performance pay. AI evaluation reduces effort by 11–14 percent among volunteers with low commitment to the cause, while having no effect on those strongly aligned with the mission. Performance pay fails to mitigate these adverse effects. Workers perceive AI as less effective at identifying quality, which appears to be the primary mechanism, and as less fair and transparent than human evaluation. Introducing an AI algorithm that explicitly applies human evaluation criteria does not mitigate these negative effects, suggesting that resistance to AI evaluation reflects deeper skepticism about machines' capacity for subjective judgment.

JEL classification

J24, M54

Keywords

algorithm aversion, algorithmic management, artificial intelligence, intrinsic motivation, worker effort

Corresponding author

Martin Abel
m.abel@bowdoin.edu

* The experiment was registered under registry number AEARCTR-0016411 and IRB approval was obtained from Bowdoin College. All errors and omissions are our own. Declarations of interest: none.

1 Introduction

Organizations are increasingly integrating artificial intelligence (AI) systems into managerial functions such as hiring, worker monitoring, and performance evaluation (Kellogg et al., 2020; Acemoglu et al., 2022). Although these systems promise lower costs and greater efficiency, their use may change how workers allocate effort and how they perform when assessments are conducted by an algorithm rather than a human. An emerging literature shows that algorithmic evaluations are often perceived as less fair or legitimate (Lavanchy et al., 2023; Germann and Merkle, 2023), but causal evidence on whether (and for whom) these perceptions translate into lower effort or lower-quality output remains limited. These potential adverse effects may be especially relevant in settings that rely on intrinsically motivated workers and provide little or no extrinsic incentive.

We partner with a nonprofit organization to study the effect of AI integration in a prosocial task. We use Prolific to recruit 1,491 U.S.-based participants who volunteer to write fundraising messages for the organization.[1] We cross-randomize whether messages are evaluated by human raters or an AI system and whether participants receive performance-based pay. This design causally identifies the effects of AI evaluation and its interaction with extrinsic incentives on workers' perceptions of the evaluation process, effort provision, and message quality.

We present four main results. First, performance pay increases the time spent on the main fundraising writing task without reducing effort in an optional hashtag-writing task or donations to the organization, suggesting that monetary incentives do not crowd out intrinsic motivation in our setting. However, conditional on performance pay, AI evaluation reduces writing time relative to human evaluation. Without pay, the effect of AI is qualitatively similar but smaller and statistically insignificant.

Second, following our pre-registered analysis plan, we examine heterogeneity by mission alignment, proxied by participants' support for the government prioritizing food shortages over other social causes. We find that among participants with below-median mission alignment, AI evaluation reduces writing time by 36 to 40 seconds (12–14%) in both the performance-pay and no-pay conditions. By contrast, among mission-aligned participants, AI evaluation has

[1] Prolific has been shown to consistently provide high-quality data for research. Recent work by Celebi et al. (2025) finds that 90% of respondents passed all main data quality checks.

no detectable effect on either writing time or message quality. We observe a similar pattern for the optional task, with AI reducing effort only among less mission-aligned participants.

Third, we examine potential mechanisms underlying the negative effects of AI evaluation. Across both pay and no-pay conditions, participants rate AI evaluators as less fair, less transparent, and less capable of identifying effective messages than human evaluators. Mediation analysis indicates that perceived effectiveness is the primary driver of the behavioral response among the tested perception measures, explaining a much larger share of the effort reduction than either fairness or transparency. This suggests that skepticism about AI capability is an important mediator of the response to algorithmic oversight.

One proposed explanation in the literature for negative responses to AI is that algorithms are perceived to apply evaluative criteria that differ from human judgment (Castelo et al., 2019; Newman et al., 2020). To test whether this perception drives the negative response, we introduced a treatment arm in which participants were (truthfully) told that the AI evaluator was instructed to apply criteria commonly associated with human judgment, such as whether messages are "sympathetic, emotionally impactful, and relatable". This intervention increased volunteers' belief that the AI values "emotional language and relatable anecdotes" in its assessment. However, it did not improve perceptions of the AI evaluation process, nor did it mitigate the adverse effects of AI on effort, suggesting that aligning AI criteria more closely with human evaluation standards may be insufficient to restore trust or motivation.

Last, we study downstream responses to AI evaluation and performance feedback. Specifically, one week after the initial task, participants learned whether their message was rated above or below the median and whether it was selected as one of 20 messages for potential use in a fundraising drive. Afterwards, participants were invited to complete a new, unincentivized task: writing campaign slogans. We find that participants continue to view AI evaluations as less fair and less effective than human evaluations, even after receiving positive feedback. AI evaluation reduces effort on the follow-up task among participants who received negative feedback (statistically significant only in the no-pay condition), while we observe no significant differences following positive feedback.

We contribute to understanding behavioral responses to algorithmic management by identifying perceived evaluator effectiveness as a primary mechanism through which AI evaluation affects worker effort. While existing work emphasizes concerns about algorithmic fairness and bias (Cowgill and Tucker, 2019; Newman et al., 2020; Dargnies et al., 2026),

we show that workers reduce effort primarily because they doubt AI's ability to accurately assess subjective quality.[2] Our results suggest that there are two distinct mechanisms: an instrumental channel, where workers reduce effort because they perceive the AI as a noisy or imprecise evaluator, and a motivational channel reflecting the broader psychological costs associated with algorithmic oversight.

An emerging literature explores interventions to mitigate these types of negative responses to AI. A common strategy is to make algorithms appear more human-like, such as using anthropomorphic design features (Blut et al., 2021) or demonstrating that algorithms can learn from mistakes (Berger et al., 2020).[3] However, most studies measure trust, stated preferences, or adoption decisions rather than how algorithmic oversight affects actual effort and output quality. Our study tests a distinct approach: we inform participants that the AI is instructed to apply human-centric evaluation criteria. Although participants update their beliefs about the AI's criteria, this does not improve perceptions of evaluator effectiveness, fairness, or transparency, nor does it prevent effort reduction.

Our third contribution addresses an important practical question: can monetary incentives offset the motivational costs of AI evaluation? The literature on intrinsic motivation documents that performance pay can either crowd out or complement prosocial motivation (Bénabou and Tirole, 2003; Gneezy and Rustichini, 2000; Mellström and Johannesson, 2008), while emerging work suggests that the integration of AI may reduce worker effort (Margalit and Raviv, 2024). To our knowledge, our study is the first to experimentally test whether monetary incentives can offset the motivational costs of AI evaluation. We find that while performance pay significantly increases baseline effort, it does not attenuate the negative effects of AI evaluation on either perceptions or effort. This suggests that extrinsic incentives are insufficient to compensate for the specific psychological and instrumental costs workers associate with algorithmic oversight.

[2]This relates to a small set of papers examining how providing information about algorithmic reliability or accuracy affects trust and adoption (Prahl and Van Swol, 2017; Yin et al., 2019). In contrast to our paper, these studies typically focus on choice behavior rather than worker effort.

[3]These studies build on a broader literature on how to mitigate algorithm aversion. For example, Dietvorst et al. (2018) show that allowing people to modify algorithm forecasts reduces algorithm aversion. Dargnies et al. (2026) find that removing gender profiling from a hiring algorithm increases workers' willingness to be evaluated by it, but providing details on how the algorithm works does not. Similarly, Kizilcec (2016) shows that greater transparency about algorithmic processes does not improve acceptance.

2 Study Design

2.1 Recruitment and Sample

We recruit 2,280 study participants via the online platform Prolific in July and August 2025. Participants are informed that the study investigates drivers of prosocial behavior and that the study takes around 10 minutes, for which they receive a base payment of 2 USD (see Online Appendix B1 for details). After giving informed consent, participants complete a baseline survey with questions about demographic characteristics, prosocial behavior, and attitudes towards technology. We next ask if they are interested in spending "a few minutes of their time to help with a prosocial task" (Figure B3). The 1,491 participants (65.4%) who agreed to volunteer form the evaluation sample.

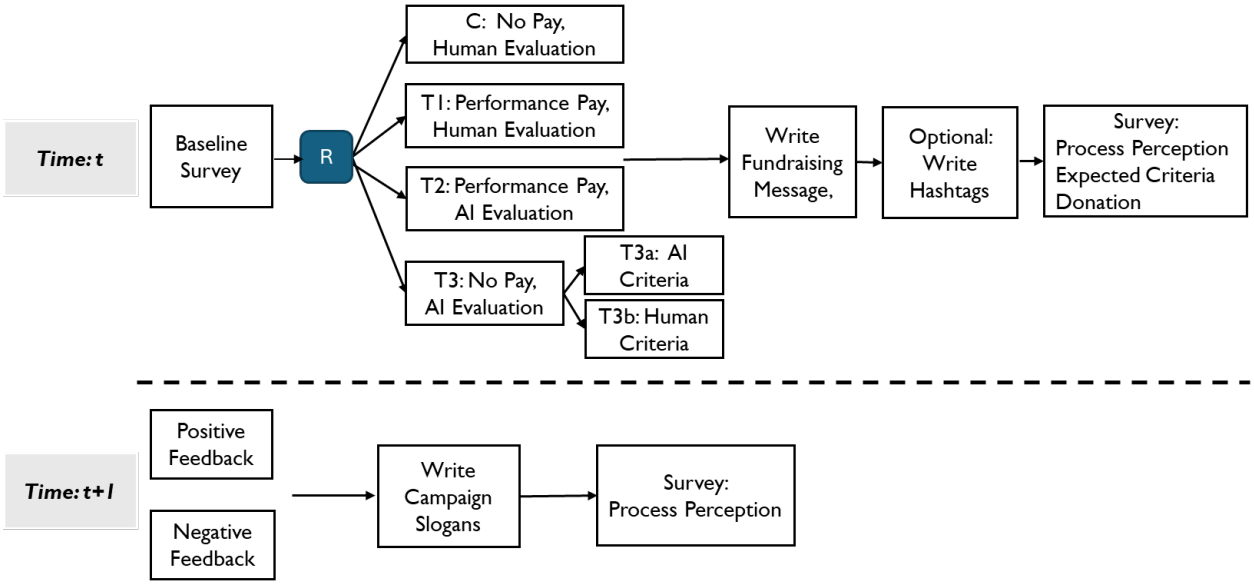
Our initial sample is representative of the US adult population with regard to age, race, and gender. However, as in the real world, selection into groups willing to volunteer is not random.[4] As shown in Table A1, volunteering rates are significantly higher among women (9.1 percentage points (pp), $p < 0.01$), college graduates (6.6 pp, $p < 0.05$), and individuals who have previously volunteered for charity (19.1 pp, $p < 0.01$). Conversely, individuals identifying as conservative or moderate are significantly less likely to volunteer than liberals (8.6 and 11.5 pp respectively, $p < 0.01$). The characteristics of our evaluation sample can be found in Table A2.

2.2 Writing Task and Treatment Variation

Figure 1 describes the experimental design. Participants who agree to volunteer first learn that the task is to "write a message to encourage people to donate to a local food bank". To help them with the task, they watch a 72 second video, which provides background information, including statistics on food shortages (see Figure B4 for the transcript). The sample size of 1,491 is equally divided into five groups.

[4]Importantly, we initially did not mention the specific prosocial cause. We therefore expect variation in levels of mission alignment for fighting food shortages, which we will explore in our analysis.

Figure 1: Experimental Design



We employ a cross randomization design, varying whether volunteers receive payment based on message quality and whether they are evaluated by humans or AI. Specifically, the performance pay group receives a 1 cent bonus for every evaluation point, for a maximum of USD 1.00. In addition, they will receive a USD 10.00 bonus if they write one of the 20 most highly rated messages, which will be used in the fundraising campaign.

For the second dimension of variation, they are informed that either an "Artificial Intelligence (AI) algorithm" or "a group of people" will evaluate the effectiveness of their message (Figures B5 and B6). Within the AI evaluation group without performance pay, we introduce a fifth group, which is (truthfully) told that the AI's algorithm is "instructed to be "human-like" and makes assessments based on how sympathetic, emotionally impactful, and relatable" the message is (Figure B7).[5] This arm is restricted to the no-pay condition to preserve statistical power within the main 2×2 design while still allowing a direct test of whether aligning AI criteria with human judgment mitigates negative responses. To increase the salience of these treatment variations, an image of a group of people, an AI symbol with

[5]The AI evaluation was implemented using GPT-4. For the standard AI treatments, the model was provided with the exact scoring instructions used by human raters (see Figure B12) . In the 'human criteria' treatment arm, the prompt was augmented with the instruction: "Make assessments based on how sympathetic, emotionally impactful, and relatable the message is". Messages were ranked separately within the AI and human groups.

code, or an AI symbol with a humanoid is presented under the message.

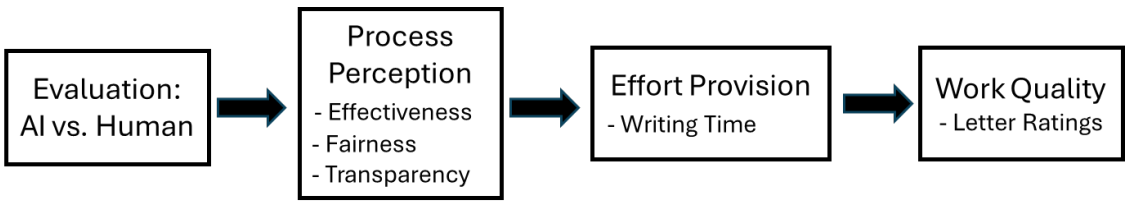
This design allows us to test the causal impact of being evaluated by AI versus humans, and whether this effect varies depending on performance pay. It is motivated by an extensive literature documenting the role of pay in either preserving or crowding out intrinsic motivation (see Gneezy et al. (2011) for a review). The introduction of "human criteria" within the AI / no-pay group is motivated by recent efforts to improve perceptions of AI by making it appear more human. If negative responses to AI stem from beliefs that algorithms apply fundamentally different evaluative criteria than humans, then explicitly instructing the AI to use human-centric criteria should attenuate these effects. If, however, the aversion reflects deeper skepticism about machines' capacity for subjective judgment, this intervention should have little impact.

Approximately one week after the first task, we re-invite participants to a second survey in which they receive feedback about the quality ratings of their messages. Specifically, they learned whether their message was above or below the median and whether it had been selected as one of the top 20 messages (Figure B11). They are then invited to write an additional campaign slogan for which no monetary incentive is provided. Lastly, we again measure their perception of the evaluation process along the same dimensions as before.

2.3 Theory of Change and Outcomes

Figure 2 depicts our theory of change, describing potential channels through which AI evaluations may affect the quality of work. First, AI vs. human evaluation may affect how the process is perceived. Specifically, it may affect whether the evaluation process is seen as "transparent", "fair and unbiased", and will "identify the most effective messages". We measure these process perceptions in random order on a zero to ten scale (see Figure B9 for details).

Figure 2: Theory of Change



Effects on process perceptions may influence volunteers' effort provision, which we capture through three complementary measures. First, without announcing it to participants, we measure the time they spend crafting the fundraising message. Because we do not communicate a target time, writing time provides a direct proxy for intrinsic motivation. Second, we measure the time spent writing hashtags, an additional task that is explicitly described as optional (Figure B8). This outcome provides an additional direct measure of effort and captures the willingness to exert effort on a voluntary task.[6] Importantly, writing time and process perceptions are measured before participants receive any feedback and thus measure the effect of the *expectation* of algorithmic evaluation.

Third, we use message quality ratings, as higher effort may translate into more effective fundraising messages. To measure message quality, we recruit a separate sample of 1,000 Prolific participants to evaluate the messages. Each evaluator reviews ten messages in random order and rates each on a 0 to 100 scale based on its effectiveness in raising funds (see Figure B12 for details).[7] In addition to humans, we also have a large language model score messages on the same scale. However, following our pre-analysis plan, we focus on human ratings as our primary outcome, as humans ultimately make donation decisions.

This theory of change regarding the impact of AI evaluations may also interact with the effects of performance pay. In particular, performance pay could moderate the effect of AI evaluation at multiple points: by sustaining effort despite negative process perceptions, by shifting attention from the evaluation process to the monetary reward, or by signaling organizational commitment to the evaluation process, even when it is delegated to AI. These interactions lead to different predictions for whether performance pay attenuates the effect of AI on process perceptions and/or effort conditional on perceptions.

2.4 Empirical Strategy

For our main specification, we estimate:

$$y_i = \alpha + \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \delta X_i + \epsilon_i \tag{1}$$

[6] Other forms of discretionary effort may also respond to algorithmic oversight: Granulo et al. (2024) find that workers managed by algorithms offer less help to colleagues.

[7] We also compute the share of spelling errors as an objective measure of message quality.

We use OLS to estimate how the main outcomes y for participant i vary by treatment group T, where $T1_i$ indicates 'Performance Pay/Human Evaluation' and $T2_i$ indicates 'Performance Pay/AI Evaluation'. Importantly, $T3_i$ is a pooled indicator for all participants in the 'No Pay/AI Evaluation' condition, representing the average effect of algorithmic oversight without extrinsic incentives across both standard and human-criteria sub-arms. Coefficients β measure the causal effect of AI evaluation and pay compared to the group that does not receive performance pay and is evaluated by humans. We report results with and without controlling for covariate vector X_i.[8]

To analyze the message quality ratings of participant i, rated by human-rater j, we estimate the following specification:

$$Quality_{ij} = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \delta X_i + \omega_j + \epsilon_{ij} \tag{2}$$

The dependent variable is the message quality of participant i rated by human-rater j. Our preferred specification reports results with rater fixed effects ω_j to account for variation in grading schemes between evaluators. Furthermore, because each message i is rated by multiple evaluators and each evaluator j rates multiple messages, the error term ϵ_{ij} may be correlated within messages and within evaluators. We therefore use two-way clustering of standard errors ϵ_{ij}.

Lastly, for our heterogeneity analysis, we will estimate:

$$y_i = \alpha + \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 S_i + \gamma_1 T1_i * S_i + \gamma_2 T2_i * S_i + \gamma_3 T3_i * S_i + \delta X_i + \epsilon_i \tag{3}$$

As registered in our PAP, we will create indicator variables for moderators S_i and interact them with treatment dummies. Coefficients γ measure whether treatment effects vary for subgroup S_i. Other details follow the specifications described above.

[8] Our PAP included regressions with interaction terms for performance pay and AI evaluations (i.e., $Pay \times AI$). We do not separately report these interaction coefficients because the indicator-based model in Specification 1 is mathematically equivalent and more directly identifies the causal effect of AI evaluation within each incentive condition.

3 Main Results

Following our theory of change (Figure 2), this section begins by reporting the treatment effects on effort provision and work quality (Section 3.1). We then examine how these effects vary by participants' baseline levels of mission-alignment (Section 3.2).

3.1 Effects on Effort and Work Quality

Table 1 reports the treatment effects on writing time and the quality ratings of messages. While writing time is a significant predictor of output quality ($p < 0.001$), the relationship is relatively attenuated: a one standard deviation (162 seconds) increase in writing time is associated with a 0.08 standard deviation increase in quality ratings. This noise is expected in creative tasks where the mapping between work time and quality is often non-linear. However, we argue that writing time, as a non-contracted measure of effort, serves as a primary indicator of worker engagement and motivation.

We find that performance pay increases effort. In the human-evaluation group, performance pay raises writing time by approximately 42 seconds (14%), and in the AI-evaluation group by about 26 seconds (8.8%, $p < 0.05$) (Cols. 1–2). These results indicate that performance pay did not crowd out intrinsic motivation, in contrast to evidence from other contexts (Gneezy and Rustichini, 2000; Bénabou and Tirole, 2006; Mellström and Johannesson, 2008).

Among participants receiving performance pay, AI evaluation reduces writing time by 23 seconds (Col. 1–2). In the no-pay group, we also observe a small reduction of about 7 seconds, although this difference is not statistically significant. These changes in effort translate into corresponding differences in message quality, but the effects are noisier, given the attenuated relationship between writing time and quality. Only in the no-pay group does AI evaluation reduce message quality, and the effect is modest, approximately 2 points (0.075 SD), and statistically significant only at the 10 percent level when covariates are included (Col. 4).

So far, we have shown the effects of AI on people's motivation to follow through on their stated commitment to provide effort in support of raising funds for a specific cause. A related question is whether they would be willing to go beyond their commitment and exert effort in tasks that are optional. We find that volunteers in the AI no pay group spend about 6

Table 1: Message Quality and Time Investment

	Writing Time		Quality	
	(1)	(2)	(3)	(4)
Human/Pay	41.68***	42.49***	0.70	0.49
	(13.02)	(12.80)	(1.34)	(1.32)
AI/Pay	19.10	18.48	-0.13	-0.09
	(13.20)	(13.08)	(1.40)	(1.35)
AI/No Pay	-7.27	-6.67	-1.81	-2.06*
	(10.99)	(10.98)	(1.18)	(1.17)
Observations	1491	1491	9688	9688
Control Mean	294.3	294.3	56.7	56.7
Control SD	162.1	162.1	26.8	26.8
Adj R-Square	0.01	0.03	0.43	0.43
Demographic Controls	N	Y	N	Y
AI/Pay vs. Human/Pay	0.092	0.070	0.546	0.678
AI/Pay vs. AI/No Pay	0.021	0.027	0.187	0.130

Notes: The dependent variable in Cols 1-2 is the time (in seconds) people spent writing messages (winsorized at the 95th percentile) and in Cols 3-4 is human-rated messaged quality. The independent variables are indicators for the four treatment groups with the Human/No Pay control group left out. Participant Demographic Controls are included in even columns. Standard errors in parentheses are robust and in Cols 3-4 are clustered by participant and rater. Cols 3-4 include rater fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

seconds (9.7%) less completing the optional task of writing hashtags for the campaign (Table A10). These findings are important, as firms cannot contractually specify or observe many activities.

3.2 Heterogeneity by Mission Alignment

These pooled estimates may mask heterogeneity in how workers respond to AI evaluation. We pre-registered participants commitment to the organization's specific mission as a key moderator for treatment effects.[9] The relationship between AI evaluation and mission

[9] Bénabou and Tirole (2003) distinguish between intrinsic motivation for the task itself (deriving utility from the process of working) and prosocial motivation for the outcome (valuing the cause being served). Our pre-specified measure — support for prioritizing food insecurity — captures the latter. In our pre-analysis

alignment is ex ante ambiguous, as economic incentives and social preferences can be either substitutes or complements (Bénabou and Tirole, 2003; Bowles and Polania-Reyes, 2012). On one hand, the use of AI algorithms may be perceived as impersonal, reducing the meaningfulness of the task and potentially undermining motivation (Frey and Jegen, 2001; Gneezy et al., 2011). Highly mission-aligned individuals may also identify more strongly with human beneficiaries, and AI oversight could "dehumanize" the cause. On the other hand, motivation among this group may be sufficiently strong to sustain effort even when evaluation is conducted by AI.

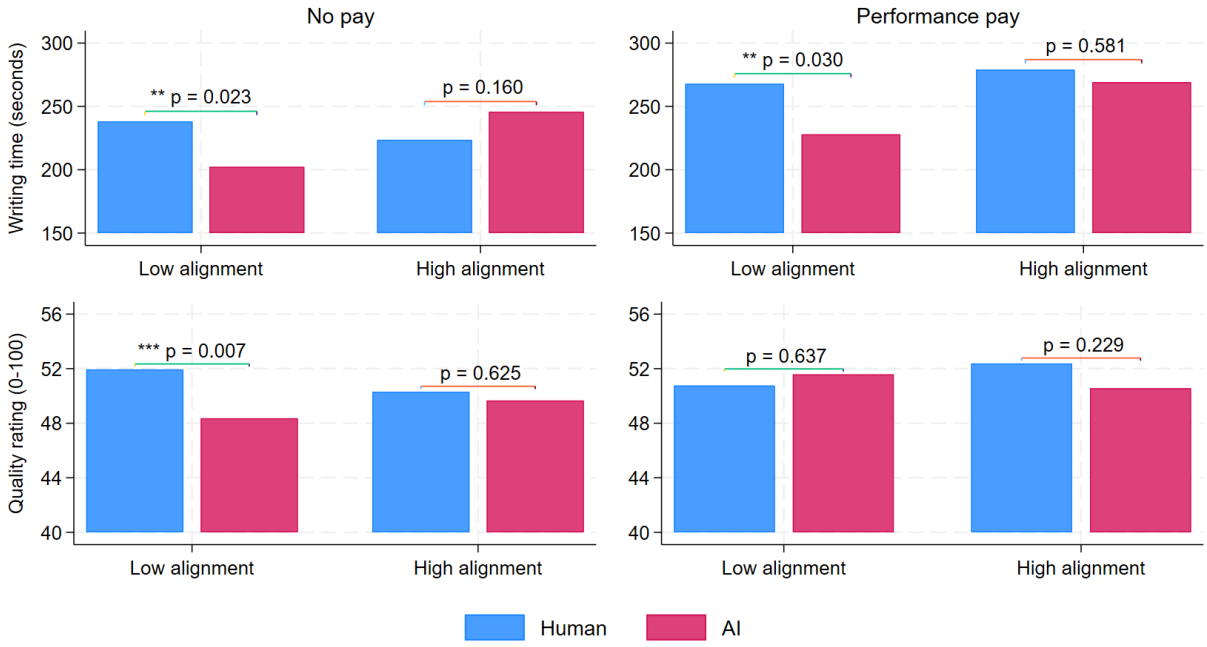
To measure baseline mission alignment, we asked participants: "In your opinion, how much should the U.S. government prioritize spending resources on addressing food shortage over other important social problems?" (0-100 scale). While this question references government spending, it captures participants' prioritization of food insecurity as a cause. Donation decisions support this interpretation: participants scoring above the median on this measure donated 27% more of their bonus to the food bank ($p < 0.01$), demonstrating revealed preference for the cause beyond stated government spending views. Following our analysis plan, we split participants at the median value of 80 and test whether treatment effects differ between high and low mission-aligned groups.

Figure 3 shows that the negative effects of AI evaluation are concentrated among workers with lower mission alignment. For this group, AI evaluation reduces writing time by 36 seconds without performance pay and by 40 seconds with performance pay (top row), and lowers message quality by 3.5 points (0.13 SD) in the no-payment condition (bottom row). Performance pay does not attenuate the negative effect on effort, suggesting that monetary incentives cannot compensate for the motivational costs of AI evaluation. The difference in AI's effect by alignment is significant at the 5% level for writing time in both pay conditions, but not for message quality.[10] For mission-aligned workers, AI evaluation has no detectable effect on either outcome. The same pattern appears in the optional hashtag task: low-alignment volunteers in the AI/no-pay condition spend 9.7 seconds (16%) less writing hashtags than their human-evaluated counterparts, while mission-aligned volunteers show no negative effect (Appendix Table A10, Col. 4). The next section examines mechanisms behind

plan, we labeled this moderator "intrinsic motivation," but "mission alignment" captures more precisely participants' commitment to the organization's cause rather than enjoyment of the writing task.

[10]One concern with our proxy is that it may conflate mission alignment with political ideology. Appendix Table A3, Col. 2, shows that the heterogeneity results are robust to controlling for political views and other demographics.

Figure 3: Effect by Mission Alignment



Bars are adjusted means with demographic controls. Brackets show Human vs AI p-values within each alignment cell.
Significance: * p<0.10, ** p<0.05, *** p<0.01.

]

these heterogeneous effects.

4 Mechanisms

This section examines potential mechanisms underlying participants' responses to AI evaluation. Section 4.1 reports treatment effects on perceived fairness, transparency, and evaluator effectiveness. Section 4.2 tests whether framing the AI as applying "human-like" evaluation criteria improves process perceptions or mitigates the effort response. Finally, Section 4.3 examines whether baseline attitudes toward technology and familiarity with AI moderate the effects of AI evaluation. Outcomes and subgroup analyses were pre-specified in our analysis plan.

4.1 Perception of Evaluation Process

Perceptions of the evaluation process are a potentially important mechanism shaping how individuals respond to different forms of assessment.[11] This mechanism is particularly relevant in the context of the rapid integration of technology into workplace evaluation systems. Empirical evidence finds that algorithmic assessments are frequently regarded as less fair or legitimate, in part because they are perceived as impersonal or opaque (Cowgill and Tucker, 2019).

Table 2 shows that participants in our study setting also hold negative views of the AI evaluation process. In the no-pay group, AI evaluations are rated 0.35 sd less fair (Col. 1), 0.61 sd less effective (Col. 3), and 0.17 sd less transparent (Col. 5) than human evaluations. Under performance pay, these negative effects are somewhat attenuated but remain similar in magnitude and statistically significant. Importantly, perceptions do not differ significantly by participants' baseline intrinsic motivation (Cols. 2, 4, 6).

Next, we examine whether the negative effect of AI oversight on effort among participants with low intrinsic motivation is mediated by perceptions of the evaluation process. Across performance pay conditions, AI's negative effect of 41 seconds shrinks by about ten seconds (25%) when we control for perceived effectiveness (Table A4, Col. 1-2). Controlling for perceived fairness or transparency leads to smaller reductions (Col. 3–4). When all three process perceptions are included jointly, only perceived effectiveness remains a significant predictor of effort (Col. 5). Moreover, treatment effects in this joint model resemble those when only effectiveness is controlled for. While mediation analysis has limitations in isolating causal mechanisms given that mediators are endogenous (Celli, 2022), these findings provide evidence for an *instrumental* channel: workers exert less effort because they view the AI as a "noisy" grader that fails to reliably reward quality. However, the fact that a large majority of the treatment effect remains unexplained by process perceptions suggests that AI evaluation also imposes a motivational cost.

[11] For example, hiring managers are reluctant to use algorithms because they perceive them as less fair (Hoffman et al., 2018). Likewise, workers often view criticism as inaccurate, which helps explain their negative reactions to feedback (Abel, 2024).

Table 2: Process Perception

	Fair		Effective		Transparent	
	(1)	(2)	(3)	(4)	(5)	(6)
Human/Pay	0.08	-0.12	0.07	-0.04	0.19	0.24
	(0.17)	(0.24)	(0.14)	(0.21)	(0.21)	(0.29)
AI/Pay	-0.55***	-0.77***	-1.00***	-1.26***	-0.34	-0.42
	(0.17)	(0.23)	(0.16)	(0.23)	(0.21)	(0.29)
AI/No Pay	-0.73***	-0.75***	-1.11***	-1.22***	-0.46**	-0.37
	(0.15)	(0.20)	(0.14)	(0.18)	(0.18)	(0.25)
High Alignm. x Human/Pay		0.39		0.22		-0.12
		(0.35)		(0.29)		(0.41)
High Alignm. x AI/Pay		0.42		0.48		0.11
		(0.34)		(0.32)		(0.42)
High Alignm. x AI/No Pay		0.03		0.21		-0.21
		(0.30)		(0.27)		(0.36)
High Alignment		0.23		0.26		0.60**
		(0.25)		(0.22)		(0.30)
Observations	1488	1488	1483	1483	1484	1484
Control Mean	7.51	7.51	7.90	7.90	6.64	6.64
Control SD	2.17	2.17	1.86	1.86	2.65	2.65
Adj R-Square	0.11	0.12	0.16	0.17	0.10	0.11
Demographic Controls	Y	Y	Y	Y	Y	Y
Low: AI/Pay vs. Human/Pay	0.000	0.007	0.000	0.000	0.009	0.026
Low: AI/No Pay vs. AI/Pay	0.210	0.908	0.423	0.867	0.495	0.837
High: AI/Pay vs. Human/Pay		0.010		0.000		0.124
High: AI/No Pay vs. AI/Pay		0.081		0.237		0.273

Notes: The dependent variables are process perceptions of fairness (Cols 1-2), effectiveness (Cols 3-4), and transparency (Cols 5-6), each measured on a 0 to 10 scale. High Alignm. is an indicator variable for individuals with above-median baseline support for government food shortage efforts. The independent variables are indicators for the treatment groups, with the Human/No Pay group serving as the omitted control category. Participant Demographic Controls are included in all specifications. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.2 AI with Human Evaluation Criteria

Organizations adopting AI have begun implementing strategies to mitigate algorithm aversion and concerns about fairness and effectiveness. A common approach is to incorporate "human

elements" into AI systems, such as emphasizing that algorithms apply similar evaluation criteria to humans or using anthropomorphic design features. These strategies aim to elicit trust by reducing the psychological distance between humans and machines (Blut et al., 2021). However, it remains unclear whether such efforts meaningfully increase the perceived fairness or legitimacy of these systems.

To test the effect of such interventions, half of the participants in the AI/no-pay group were (truthfully) told that the algorithm assessing their messages uses human evaluation criteria. Table A5 shows that adding this framing does not significantly alter any treatment effects, including perceptions of process efficiency or fairness (Cols. 2, 4, 6, 8).

A key question for interpreting these results is whether participants internalized our treatment variation - specifically, whether they believed that the selection criteria differed across groups. To assess this, we asked participants: "What do you think is more important to include to receive a positive assessment from the people who evaluate your message?" Response options were "emotional language and relatable anecdotes," "numbers and statistics," or "equally important." Table A6 shows large differences in perceived evaluation criteria across treatment arms. Within the AI/no-pay group, participants are 9.5 pp (28%) less likely to mention anecdotes compared to the control group (Col. 2). By contrast, in the AI with human criteria group, participants are 22 pp (65%) more likely to respond that anecdotes and emotional language lead to positive assessments (Col. 2). These results are robust to alternative coding of responses (Col. 4).

In sum, these results suggest that the underlying reason for aversion to AI evaluation is not necessarily the evaluative criteria themselves, but rather a perceived "algorithmic reductionism" (Newman et al., 2020). Even when informed that the AI utilizes human-centric benchmarks like sympathy and relatability, participants may believe that a machine fundamentally lacks the capacity for the subjective judgment required to evaluate such qualities (Castelo et al., 2019). Consequently, they may view the AI's application of these criteria as a superficial and ineffective proxy for human judgment. This conclusion is consistent with prior evidence showing that greater transparency about how algorithms operate does not necessarily improve their acceptance or use (Kizilcec, 2016; Dargnies et al., 2026).

4.3 Attitude towards Technology and Familiarity with AI

Finally, we explore whether the negative response to AI evaluation is (partly) driven by a general lack of familiarity or skepticism toward technology, factors often cited as underlying reasons for "algorithm aversion" (Burton et al., 2020; Dargnies et al., 2026). Table A7 reports heterogeneity by participants' baseline attitudes toward technology. Specifically, we asked respondents if recent advances in technology had a more positive or more negative impact on society. We then construct an indicator for above-median attitudes and interact it with treatment assignments, as specified in our PAP. As expected, technology-optimistic respondents, who believe AI has had a mostly positive impact, perceive AI evaluations as significantly more effective than their skeptical counterparts (Col. 5–6). However, even this group views human evaluators as more effective overall. It is also notable that more positive views toward technology do not appear to mitigate concerns regarding fairness or bias (Col. 7–8).

Furthermore, a positive attitude toward technology is not associated with more positive behavioral responses to AI oversight. In fact, the interaction coefficients for writing time and quality are directionally negative (Col. 1–4). Overall, these results indicate that general attitudes toward technology do not significantly moderate the aversion to AI evaluation, though a specific measure of attitudes toward AI might yield different results.[12]

We observe a similar pattern regarding AI usage. High-frequency users, defined as those using AI at least several times per week, report significantly more positive views of AI effectiveness (Table A8, Col. 5–6). However, this optimism does not translate into improved performance. The interaction terms for writing time and quality are negative, indicating that frequent users do not respond more positively to AI oversight. While the exact interpretation of this moderator remains somewhat ambiguous, these findings suggest that increased familiarity alone is insufficient to make people more receptive to AI-based evaluations.

Lastly, we want to acknowledge important caveats to the interpretation of these results. Unlike the previous subsections, which relied on experimental manipulation, this analysis depends on worker characteristics that are not randomly assigned. For example, high frequency users of AI or those with more positive views towards technology likely differ along

[12]We intentionally elicited general technology attitudes to avoid priming participants regarding the study's focus on AI. However, responses to this question are highly correlated with AI-specific attitudes collected at endline.

many unobservable characteristics. Nonetheless, the consistent lack of a positive behavioral response across both moderators suggests that increased familiarity alone is unlikely to be sufficient to overcome resistance to AI-based evaluation.

5 Downstream Consequences of AI Integration

Our previous analysis has focused on how AI evaluation affects immediate effort provision on the assigned task. However, the integration of algorithmic oversight may have broader implications for the worker-organization relationship. Specifically, if AI evaluation leads workers to perceive the organization as less fair or effective, this may erode commitment and diminish their willingness to support the organization beyond the scope of the initial agreement. This section examines these downstream consequences through two outcomes: participants' willingness to donate a portion of their earnings back to the non-profit organization (Section 5.1) and their response to performance feedback provided one week later (Section 5.2).

5.1 Donations

First, we asked participants who evaluated fundraising messages whether they would be more or less likely to donate to an organization that uses AI to identify effective messages. While the majority (67%) reported that it would have no effect on their donation decision, 30.3% indicated they would be less likely to donate compared to only 2.7% who would be more likely. To understand the mechanisms underlying donation reluctance, we asked those who would donate less to explain their reasoning in an open-ended question.[13] The most frequently cited reason was that the use of AI signals inauthenticity and a lack of genuineness (40.4% of responses; Table A11). One respondent wrote: "I prefer to support organizations where I feel a real human connection, not one where my emotions are being calculated and tested by a machine." Other common concerns included perceptions that AI signals low organizational effort (18.8%), trust and manipulation concerns (15.4%), general ideological opposition to AI (15.1%), AI's lack of emotional understanding (11.6%), and doubts about AI's ability to identify effective messages (10.3%).

[13] Open-ended responses were initially categorized using Claude (Sonnet 4.5), which suggested nine distinct categories. The research team reviewed the categorizations and validated the final coding scheme. Multiple categories could be assigned to each response.

Going beyond hypothetical questions, we give participants in our experimental sample the option to donate part of their bonus to the organization running the fundraising campaign. We find that participants in the AI groups donate 4.5% to 6% less (Table A10, Col. 1). While this reduction in actual donations is economically meaningful, it is imprecisely estimated ($p > 0.10$). However, the directional consistency between hypothetical intent and incentivized behavior suggests a non-trivial reputational risk to organizations that could offset the efficiency gains of AI integration.[14]

5.2 Responses to Performance Feedback

Performance feedback is a fundamental component of the evaluation process and has been shown to affect job satisfaction and future effort (Kluger and DeNisi, 1996). Participants who are evaluated by AI may respond particularly negatively to performance feedback since the process is perceived as ineffective and less fair. Table 3 reports the impact of performance feedback on the time participants invest in the unincentivized campaign slogan task (Col. 1-2) and process perceptions (Col. 3-8).

We first examine the impact on writing time. While treatment coefficients in the pooled sample are small and statistically insignificant (Col. 1), this aggregate result masks substantial heterogeneity by feedback content and baseline motivation.[15] Within the no-pay condition, participants who received negative feedback from an AI reduced their writing time by 12.7 seconds (20%) relative to those who received identical negative feedback based on human ratings (Table A9, Col. 2). A similar, though less pronounced, pattern emerges in the AI/pay group. Furthermore, when exploring treatment effects by baseline intrinsic motivation, the negative impact of criticism on effort appears to be again concentrated among low-motivation individuals (Table A9, Cols. 3-4). However, these interaction effects are imprecisely estimated and not statistically significant.

[14] These results do not necessarily imply a divergence of incentivized and hypothetical evaluations, as documented in other studies of AI (Abel and Johnson, 2025). A 6% drop in donations could, for example, be the result of 30% of participants donating 20% less to organizations using AI.

[15] One important caveat for interpreting these results is that all participants received feedback, meaning the content (positive vs. negative) was not randomly assigned. Consequently, while the treatment effects within feedback type remain well-identified, the observed differences in responses to positive versus negative feedback within a treatment arm should be interpreted as correlational. A related concern is that the treatment may have affected the content of the feedback, which is addressed by the fact that praise or criticism was determined based on performance distributions within treatment arms.

Table 3: Effect of Performance Feedback

	Writing Time		Fair		Effective		Transparent	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human/Pay	9.12	1.15	-0.03	0.03	-0.09	0.37	-0.03	0.15
	(6.53)	(9.33)	(0.21)	(0.32)	(0.20)	(0.31)	(0.26)	(0.37)
AI/Pay	3.28	-5.01	-0.63***	-0.49	-0.64***	-0.44	-0.26	-0.26
	(6.47)	(9.17)	(0.22)	(0.34)	(0.21)	(0.32)	(0.26)	(0.36)
AI/No Pay	-1.63	-12.57	-0.67***	-0.58**	-0.57***	-0.37	-0.14	0.06
	(5.45)	(7.65)	(0.19)	(0.29)	(0.18)	(0.27)	(0.22)	(0.31)
Pos Feedb x Human/Pay		16.28		-0.11		-0.91**		-0.34
		(12.88)		(0.40)		(0.38)		(0.51)
Pos Feedb x AI/Pay		16.19		-0.33		-0.46		-0.05
		(12.83)		(0.43)		(0.41)		(0.50)
Pos Feedb x AI/No Pay		22.09**		-0.18		-0.41		-0.41
		(10.83)		(0.36)		(0.33)		(0.43)
1=Feedback positive		-4.12		1.13***		1.53***		1.27***
		(9.05)		(0.30)		(0.28)		(0.35)
Observations	1331	1331	1326	1326	1326	1326	1325	1325
Control Mean	65.61	65.61	6.96	6.96	7.07	7.07	5.84	5.84
Control SD	72.97	72.97	2.56	2.56	2.40	2.40	3.00	3.00
Adj R-Square	0.03	0.03	0.05	0.08	0.03	0.09	0.05	0.08
Demographic Controls	Y	Y	Y	Y	Y	Y	Y	Y

Notes: The dependent variables are the time (in seconds) volunteers spent writing campaign slogans, winsorized at the 95th percentile (Cols 1-2), and process perceptions of fairness (Cols 3-4), effectiveness (Cols 5-6), and transparency (Cols 7-8) measured from 0-10 in the feedback survey round. The independent variables are indicators for the treatment groups, with the Human/No Pay group serving as the omitted control category. They are interacted with a dummy indicating whether the volunteer received positive rather than negative feedback. Participant Demographic Controls are included in all specifications. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Turning to process evaluations, we again find that AI evaluations are perceived to be less effective and fair, regardless of whether people are compensated (Col. 3, 5).[16] Most strikingly, while receiving positive feedback improved absolute perceptions of the evaluation process across all groups, it failed to close the gap between human and algorithmic oversight. Perceptions of fairness and effectiveness within the AI groups remained significantly lower than those in the human groups even after positive feedback (Col. 4, 6). This suggests that the concerns about AI's capability and fairness documented earlier represent fundamental

[16]There is no difference in transparency (Col. 7), possibly because we explain the exact criteria used for the feedback.

rather than easily malleable attitudes. These findings complement (Margalit and Raviv, 2024), who document that AI managers fail to elicit the same motivational boost from positive feedback as humans.

While some of these downstream effects are estimated imprecisely, the results consistently suggest that AI evaluation generates persistent negative perceptions that extend beyond immediate task performance. For organizations, these findings suggest that the reputational costs of AI adoption may be difficult to mitigate and must be weighed against efficiency gains, especially in settings that rely on intrinsic motivation and voluntary effort.

6 Discussion

This paper studies how workers respond to AI-based performance evaluation in a prosocial task. Using an experiment with 1,491 volunteers writing fundraising messages, we show that AI evaluation reduces effort by 11-14% among workers with below-median mission alignment, while more aligned motivated workers are unaffected. This behavioral response is mediated primarily by beliefs that AI is less effective at identifying quality work. Explicitly programming AI to use "human-like" criteria does not mitigate these effects, and negative perceptions persist even after positive feedback.

One important concern is the external validity of our findings. Since our online sample likely represents relatively tech-savvy workers, our findings could understate resistance to AI integration in the broader labor force. However, perhaps surprisingly, we find no evidence that frequent AI users or those with positive attitudes toward technology respond more favorably to AI evaluation, suggesting that familiarity alone may be insufficient to overcome resistance. A related concern is that our study context, which recruits volunteers for a one-time prosocial task, may not generalize to conventional employment settings with more stable worker-firm relationships. While the fact that performance pay does not eliminate AI's negative effects suggests that the mechanisms we identify extend beyond volunteer settings, future research should study AI integration in traditional employment contexts.

Our findings carry several implications for organizations considering AI adoption in monitoring and performance evaluation. First, the heterogeneity by baseline mission alignment suggests that screening for workers with a strong commitment to the organizational mission

may become more important, as negative perceptions of AI oversight do not translate into reduced effort for this group in our setting. Second, interventions focused on algorithm design, such as programming AI to explicitly apply human evaluation criteria, appear insufficient to overcome negative responses to AI. Third, decisions about whether to integrate AI should account not only for cost savings but also for potential reductions in worker motivation and reputational effects among clients.

Overall, these results contribute to the understanding of how workers respond when algorithms replace human judgment in evaluation. While much attention in the emerging literature on AI has focused on fairness and bias concerns, we show that beliefs about evaluator effectiveness are an additional important factor shaping workers' responses and may also affect whether clients support the organization.

References

Abel, M. (2024): "Do workers discriminate against female bosses?" *Journal of Human Resources*, 59, 470–501.

Abel, M. and R. Johnson (2025): "AI Bias for Creative Writing: Subjective Assessment Versus Willingness to Pay," .

Acemoglu, D., D. Autor, J. Hazell, and P. Restrepo (2022): "Artificial Intelligence and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics*, 40, S293–S340.

Bénabou, R. and J. Tirole (2003): "Intrinsic and extrinsic motivation," *The Review of Economic Studies*, 70, 489–520.

——— (2006): "Incentives and prosocial behavior," *American Economic Review*, 96, 1652–1678.

Berger, B., M. Adam, A. Rühr, and A. Benlian (2020): "Watch me improve—algorithm aversion and demonstrating the ability to learn," *Business & Information Systems Engineering*, 63, 55–68.

Blut, M., C. Wang, N. V. Wünderlich, and C. Brock (2021): "Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI," *Journal of the Academy of Marketing Science*, 49, 632–658.

Bowles, S. and S. Polania-Reyes (2012): "Economic incentives and social preferences: substitutes or complements?" *Journal of Economic Literature*, 50, 368–425.

Burton, J. W., M.-K. Stein, and T. B. Jensen (2020): "A systematic review of algorithm aversion in augmented decision making," *Journal of Behavioral Decision Making*, 33, 220–239.

Castelo, N., M. W. Bos, and D. R. Lehmann (2019): "Task-Dependent Algorithm Aversion," *Journal of Marketing Research*, 56, 809–825.

Celebi, C., C. Exley, S. Harrs, H. Kivimaki, M. Serra-Garcia, and J. Yusof (2025): "Mission Possible: The Collection of High-Quality Data," .

Celli, V. (2022): "Causal mediation analysis in economics: Objectives, assumptions, models," *Journal of Economic Surveys*, 36, 214–234.

Cowgill, B. and C. E. Tucker (2019): "Economics, fairness and algorithmic bias," *Preparation for: Journal of Economic Perspectives*.

Dargnies, M.-P., R. Hakimov, and D. Kübler (2026): "Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence," *Management Science*, 72, 285–301.

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

Frey, B. S. and R. Jegen (2001): "Motivation crowding theory," *Journal of Economic Surveys*, 15, 589–611.

Germann, M. and C. Merkle (2023): "Algorithm aversion in delegated investing," *Journal of Business Economics*, 93, 1691–1727.

Gneezy, U., S. Meier, and P. Rey-Biel (2011): "When and why incentives (don't) work to modify behavior," *Journal of Economic Perspectives*, 25, 191–210.

Gneezy, U. and A. Rustichini (2000): "Pay enough or don't pay at all," *The Quarterly Journal of Economics*, 115, 791–810.

Granulo, A., S. Caprioli, C. Fuchs, and S. Puntoni (2024): "Deployment of algorithms in management tasks reduces prosocial motivation," *Computers in Human Behavior*, 152, 108094.

Hoffman, M., L. B. Kahn, and D. Li (2018): "Discretion in hiring," *The Quarterly Journal of Economics*, 133, 765–800.

Kellogg, K. C., M. A. Valentine, and A. Christin (2020): "Algorithms at Work: The New Contested Terrain of Control," *Academy of Management Annals*, 14, 366–410.

Kizilcec, R. F. (2016): "How much information? Effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395.

Kluger, A. N. and A. DeNisi (1996): "The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory," *Psychological Bulletin*, 119, 254.

Lavanchy, M., P. Reichert, J. Narayanan, and K. Savani (2023): "Applicants' Fairness Perceptions of Algorithm-Driven Hiring Procedures," *Journal of Business Ethics*, 188, 125–150.

Margalit, Y. and S. Raviv (2024): "When Your Boss is an Algorithm: The Effect of Algorithmic Management on Worker Performance," *Available at SSRN 4776355*.

Mellström, C. and M. Johannesson (2008): "Crowding out in blood donation: was Titmuss right?" *Journal of the European Economic Association*, 6, 845–863.

Newman, D. T., N. J. Fast, and D. J. Harmon (2020): "When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions," *Organizational Behavior and Human Decision Processes*, 160, 149–167.

Prahl, A. and L. M. Van Swol (2017): "Understanding algorithm aversion: When is advice from automation discounted?" *Journal of Forecasting*, 36, 691–702.

Yin, M., J. Wortman Vaughan, and H. Wallach (2019): "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

A Appendix

Table A1: Selection into Volunteering

	1=Agrees to Volunteer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.089***						0.091***
	(0.020)						(0.020)
Respondent Age		-0.001					-0.001
		(0.001)					(0.001)
Black or African American			0.042				0.033
			(0.027)				(0.027)
Hispanic or Latino			0.031				0.029
			(0.031)				(0.032)
Asian			-0.050				-0.059
			(0.037)				(0.037)
College				0.071**			0.066**
				(0.030)			(0.030)
Conservative					-0.096***		-0.086***
					(0.023)		(0.023)
Moderate					-0.122***		-0.115***
					(0.026)		(0.026)
1=Volunteered for charity						0.191***	
						(0.020)	
Observations	2278	2280	2280	2280	2280	2279	2278
Sample Mean	0.656	0.656	0.656	0.656	0.656	0.656	0.656
Sample SD	0.475	0.475	0.475	0.475	0.475	0.475	0.475
R-Square	0.009	0.001	0.002	0.003	0.013	0.039	0.027

Notes: The dependent variable is in indicator variable measuring if people agree to volunteer for the task. Standard errors in parentheses are robusts. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A2: Balance Table

	(1) Human/No Pay		(2) AI/No Pay		(3) Human/Pay		(4) AI/Pay		(5) AI/Human		T-test Difference			
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)
Female	293	0.549	303	0.531	297	0.545	301	0.512	296	0.564	0.018	0.004	0.038	-0.015
		(0.029)		(0.029)		(0.029)		(0.029)		(0.029)				
Age	294	44.310	303	45.347	297	45.657	301	46.296	296	45.720	-1.037	-1.347	-1.986	-1.410
		(0.914)		(0.890)		(0.921)		(0.895)		(0.922)				
Asian	294	0.071	303	0.069	297	0.077	301	0.100	296	0.074	0.002	-0.006	-0.028	-0.003
		(0.015)		(0.015)		(0.016)		(0.017)		(0.015)				
White	294	0.711	303	0.716	297	0.734	301	0.688	296	0.730	-0.005	-0.023	0.023	-0.019
		(0.026)		(0.026)		(0.026)		(0.027)		(0.026)				
Black	294	0.177	303	0.142	297	0.172	301	0.176	296	0.149	0.035	0.005	0.001	0.028
		(0.022)		(0.020)		(0.022)		(0.022)		(0.021)				
Hispanic	294	0.122	303	0.149	297	0.128	301	0.103	296	0.088	-0.026	-0.005	0.019	0.035
		(0.019)		(0.020)		(0.019)		(0.018)		(0.016)				
Educ yrs	294	15.190	303	15.340	297	15.333	301	15.266	296	15.436	-0.149	-0.143	-0.075	-0.245
		(0.146)		(0.146)		(0.144)		(0.139)		(0.147)				
Conservative	294	0.320	303	0.287	297	0.263	301	0.302	296	0.267	0.033	0.057	0.017	0.053
		(0.027)		(0.026)		(0.026)		(0.027)		(0.026)				
Technology Att.	286	0.280	301	0.445	293	0.304	298	0.329	294	0.367	-0.165***	-0.024	-0.049	-0.088
		(0.040)		(0.038)		(0.040)		(0.038)		(0.040)				
Intrinsic Motiv.	294	74.956	303	76.865	297	75.165	301	74.153	296	74.615	-1.909	-0.209	0.803	0.341
		(1.061)		(0.921)		(1.166)		(1.161)		(1.096)				
AI Usage	294	1.000	302	1.175	297	0.936	301	1.123	296	1.159	-0.175	0.064	-0.123	-0.159
		(0.078)		(0.077)		(0.079)		(0.080)		(0.077)				
F-test of joint significance (p-value)											0.152	0.840	0.601	0.277

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A4: Mediation Analysis: Effort Provision, Low Mission Alignment

	Writing Time (sec)				
	(1)	(2)	(3)	(4)	(5)
Human/Pay	24.36	23.31	24.18	23.14	24.48
	(18.81)	(18.69)	(18.66)	(18.85)	(18.68)
AI/Pay	-13.81	-5.01	-9.40	-12.82	-4.68
	(20.54)	(20.69)	(20.55)	(20.51)	(20.80)
AI/No Pay	-41.12**	-30.80*	-38.07**	-38.74**	-30.00*
	(16.41)	(16.64)	(16.21)	(16.40)	(16.77)
Process Effective		9.71***			9.95**
		(2.96)			(4.10)
Process Fair			6.82**		2.93
			(2.71)		(3.76)
Process Transparent				2.83	-3.42
				(2.35)	(3.18)
Observations	704	699	701	700	694
Control Mean	287.6	287.6	287.6	287.6	287.6
Control SD	164.2	164.2	164.2	164.2	164.2
Adj R-Square	0.05	0.06	0.05	0.05	0.06
Demographic Controls	N	N	N	N	N
Low: AI/Pay vs. Human/Pay	0.055	0.165	0.095	0.074	0.155
Low: AI/No Pay vs. AI/Pay	0.115	0.138	0.100	0.135	0.148

Notes: The sample is restricted to people with below median baseline level of support for the government supporting food shortage efforts. The dependent variable is the time (in seconds) people spent writing messages (winsorized at the 95th percentile). The independent variables are indicators for the four treatment groups with the Human/No Pay control group left out. Process perception variables are measured on a 0 to 10 scale. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Subgroup Analysis: Mission Alignment

	Writing Time		Quality	
	(1)	(2)	(3)	(4)
Human/Pay	28.64	29.73	-1.03	-1.18
	(18.25)	(18.19)	(2.01)	(1.97)
AI/Pay	-7.52	-10.21	-0.63	-0.36
	(18.29)	(18.28)	(1.96)	(1.91)
AI/No Pay	-36.16**	-35.85**	-3.49**	-3.58**
	(15.73)	(15.71)	(1.71)	(1.65)
High Align. x Human/Pay	25.92	25.58	3.41	3.26
	(25.77)	(25.68)	(2.82)	(2.75)
High Align. x AI/Pay	51.68**	55.74**	1.09	0.62
	(25.70)	(25.75)	(2.81)	(2.74)
High Motiv x AI/No Pay	57.21**	58.05***	3.28	2.94
	(22.31)	(22.21)	(2.35)	(2.24)
High Alignment	-11.05	-14.48	-2.07	-1.64
	(18.29)	(18.28)	(1.97)	(1.88)
Observations	1491	1491	9688	9688
Control Mean	294.3	294.3	56.7	56.7
Control SD	162.1	162.1	26.8	26.8
Adj R-Square	0.02	0.04	0.43	0.43
Demographic Controls	N	Y	N	Y
Low: AI/Pay vs. Human/Pay	0.050	0.030	0.844	0.690
Low: AI/No Pay vs. AI/Pay	0.072	0.106	0.108	0.073
High: AI/Pay vs. Human/Pay	0.559	0.581	0.312	0.342
High: AI/No Pay vs. AI/Pay	0.133	0.128	0.701	0.604

Notes: The dependent variable in Cols 1-2 is the time (in seconds) people spent writing messages (winsorized at the 95th percentile) and in Cols 3-4 is human-rated messaged quality. The independent variables are indicators for the four treatment groups with the Human/No Pay control group left out. High Align. is a dummy variable measuring whether people have above-median baseline level of support for the government supporting food shortage efforts. Participant Demographic Controls are included in even columns. Standard errors in parentheses are robust and in Cols 3-4 are clustered by participant and rater. Cols 3-4 include rater fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Effect of AI with Human Characteristics

	Quality		Writing Time		Effective		Fair	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human/Pay	0.49	0.49	42.49***	42.53***	0.07	0.07	0.08	0.08
	(1.32)	(1.32)	(12.80)	(12.81)	(0.14)	(0.14)	(0.17)	(0.17)
AI/Pay	-0.09	-0.10	18.48	18.50	-1.00***	-1.00***	-0.55***	-0.55***
	(1.35)	(1.35)	(13.08)	(13.09)	(0.16)	(0.16)	(0.17)	(0.17)
AI/No Pay	-2.06*	-2.45*	-6.67	1.32	-1.11***	-1.15***	-0.73***	-0.80***
	(1.17)	(1.37)	(10.98)	(12.93)	(0.14)	(0.16)	(0.15)	(0.18)
AI/No Pay x Human		0.77		-16.03		0.08		0.14
		(1.39)		(12.58)		(0.17)		(0.18)
Observations	9688	9688	1491	1491	1483	1483	1488	1488
Control Mean	56.7	56.7	294.3	294.3	8.1	8.1	7.7	7.7
Control SD	26.8	26.8	162.1	162.1	1.8	1.8	2.2	2.2
Adj R-Square	0.43	0.43	0.03	0.03	0.16	0.16	0.11	0.11
Demographic Controls	Y	Y	Y	Y	Y	Y	Y	Y
AI/Pay vs. Human/Pay	0.678	0.674	0.070	0.070	0.000	0.000	0.000	0.000
AI/Pay vs. AI/No Pay	0.130	0.110	0.027	0.197	0.423	0.362	0.210	0.141
AI/Pay vs. AI+Human		0.296		0.009		0.647		0.509

Notes: The dependent variable in Cols 1-2 is human-rated messaged quality and in Cols 3-4 is the time (in seconds) people spent writing messages, winsorized at the 95th percentile. Col 5-6 and 7-8 measure whether people perceive the process to be effective and fair, respectively. The independent variables are indicators for the treatment groups with the Human/No Pay control group left out. Participant Demographic Controls are included in even columns. Standard errors in parentheses are robust and in Cols 1-2 are clustered by participant and rater. Cols 1-2 include rater fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Expected Evaluation Criteria

	1=Anecdotes		1=Anec, -1=Stat	
	(1)	(2)	(3)	(4)
Human/Pay	0.030	0.030	0.009	0.009
	(0.045)	(0.045)	(0.041)	(0.041)
AI/Pay	-0.139***	-0.140***	-0.092**	-0.092**
	(0.048)	(0.048)	(0.039)	(0.039)
AI/No Pay	0.015	-0.095**	0.034	-0.068*
	(0.041)	(0.047)	(0.036)	(0.040)
AI/No Pay x Human		0.220***		0.205***
		(0.049)		(0.040)
Observations	1490	1490	1490	1490
Control Mean	0.34	0.34	0.38	0.38
Control SD	0.56	0.56	0.49	0.49
Adj R-Square	0.02	0.03	0.01	0.03
Demographic Controls	Y	Y	Y	Y
AI/Pay vs. Human/Pay	0.000	0.000	0.009	0.010
AI/Pay vs. AI/No Pay	0.000	0.351	0.000	0.533
Human/No Pay vs. AI+Human		0.010		0.001

Notes: The dependent variable in Cols 1-2 is the response to the expected criteria the evaluation is using, coded as -1 = statistics, 1= anecdotes, and 0 = both equally. The dependent variable in Cols 3-4 is a binary variable equal to 1 if the criteria is anecdotes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A7: Subgroup Analysis: Attitudes Toward Technology

	Quality		Writing Time		Effective		Fair	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human/Pay	0.98	0.82	46.95***	47.84***	-0.06	-0.02	-0.03	-0.00
	(1.64)	(1.64)	(16.48)	(16.25)	(0.21)	(0.21)	(0.25)	(0.25)
AI/Pay	1.16	1.23	27.12	32.17*	-1.17***	-1.15***	-0.52**	-0.50**
	(1.80)	(1.77)	(17.21)	(17.08)	(0.23)	(0.23)	(0.24)	(0.24)
AI/No Pay	-1.79	-1.98	10.93	15.81	-1.44***	-1.40***	-0.82***	-0.78***
	(1.64)	(1.61)	(14.58)	(14.66)	(0.20)	(0.20)	(0.21)	(0.21)
Positive x Human/Pay	-1.12	-1.12	-15.03	-13.00	0.22	0.24	0.16	0.17
	(2.72)	(2.62)	(26.91)	(26.77)	(0.29)	(0.29)	(0.35)	(0.35)
Positive x AI/Pay	-3.42	-3.38	-20.12	-29.80	0.40	0.37	-0.06	-0.08
	(2.81)	(2.67)	(27.07)	(27.11)	(0.32)	(0.32)	(0.34)	(0.34)
Positive x AI/No Pay	-0.52	-0.63	-44.25*	-48.10**	0.58**	0.60**	0.08	0.10
	(2.48)	(2.35)	(22.56)	(22.76)	(0.27)	(0.27)	(0.31)	(0.31)
Positive Tech Attitude	0.79	1.02	37.62**	42.89**	0.61***	0.61***	0.95***	0.95***
	(1.96)	(1.95)	(18.79)	(18.99)	(0.21)	(0.21)	(0.25)	(0.25)
Observations	9582	9582	1472	1472	1464	1464	1469	1469
Control Mean	56.7	56.7	294.3	294.3	8.1	8.1	7.7	7.7
Control SD	26.8	26.8	162.1	162.1	1.8	1.8	2.2	2.2
Adj R-Square	0.43	0.43	0.01	0.03	0.11	0.14	0.07	0.09
Demographic Controls	N	Y	N	Y	N	Y	N	Y

Notes: The dependent variable in Cols 1-2 is human-rated messaged quality and in Cols 3-4 is the time (in seconds) people spent writing messages, winsorized at the 95th percentile. Col 5-6 and 7-8 measure whether people perceive the process to be effective and fair, respectively. is an indicator variable equal to one for participants who are more positive about the impact of technology on society. The independent variables are indicators for the treatment groups with the Human/No Pay control group left out. Participant Demographic Controls are included in even columns. Standard errors in parentheses are robust and in Cols 1-2 are clustered by participant and rater. Cols 1-2 include rater fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A8: Subgroup Analysis: AI Usage

	Quality		Writing Time		Effective		Fair	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human/Pay	-0.15	-0.37	34.49**	36.70**	0.12	0.12	0.19	0.18
	(1.54)	(1.53)	(15.81)	(15.56)	(0.19)	(0.18)	(0.23)	(0.21)
AI/Pay	-1.87	-1.93	6.74	6.41	-1.16***	-1.25***	-0.61***	-0.70***
	(1.86)	(1.81)	(16.06)	(15.59)	(0.21)	(0.20)	(0.23)	(0.22)
AI/No Pay	-1.27	-1.58	-4.71	-1.77	-1.34***	-1.41***	-0.84***	-0.92***
	(1.36)	(1.34)	(14.00)	(13.92)	(0.18)	(0.17)	(0.20)	(0.19)
High AI x Human/Pay	2.37	2.48	19.24	15.60	-0.17	-0.14	-0.27	-0.26
	(2.90)	(2.78)	(27.59)	(27.23)	(0.31)	(0.31)	(0.38)	(0.37)
High AI x AI/Pay	4.38	4.67*	29.08	28.97	0.46	0.62*	0.24	0.38
	(2.90)	(2.78)	(27.63)	(27.37)	(0.33)	(0.33)	(0.36)	(0.35)
High AI x AI/No Pay	-1.18	-1.06	-9.05	-13.77	0.67**	0.72***	0.44	0.46
	(2.54)	(2.38)	(22.70)	(22.63)	(0.29)	(0.28)	(0.32)	(0.31)
High AI Use	-1.70	-1.61	16.38	14.68	0.42*	0.14	0.43*	0.15
	(2.05)	(1.90)	(18.97)	(18.80)	(0.22)	(0.23)	(0.26)	(0.26)
Observations	9685	9685	1490	1490	1482	1482	1487	1487
Control Mean	56.7	56.7	294.3	294.3	8.1	8.1	7.7	7.7
Control SD	26.8	26.8	162.1	162.1	1.8	1.8	2.2	2.2
Adj R-Square	0.43	0.43	0.02	0.03	0.09	0.18	0.04	0.12
Demographic Controls	N	Y	N	Y	N	Y	N	Y

Notes: The dependent variable in Cols 1-2 is human-rated messaged quality and in Cols 3-4 is the time (in seconds) people spent writing messages, winsorized at the 95th percentile. Col 5-6 and 7-8 measure whether people perceive the process to be effective and fair, respectively. High AI is an indicator variable equal to one for participants who use AI more frequently. The independent variables are indicators for the treatment groups with the Human/No Pay control group left out. Participant Demographic Controls are included in even columns. Standard errors in parentheses are robust and in Cols 1-2 are clustered by participant and rater. Cols 1-2 include rater fixed effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A9: Effect of Performance Feedback

	Writing Time				Effective			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human/Pay	9.12	1.15	5.83	-3.60	-0.04	0.03	-0.02	0.09
	(6.53)	(9.33)	(12.12)	(14.29)	(0.21)	(0.32)	(0.42)	(0.47)
AI/Pay	3.28	-5.01	-15.34	0.91	-0.64***	-0.49	-0.53	-0.39
	(6.47)	(9.17)	(11.56)	(13.53)	(0.23)	(0.34)	(0.46)	(0.50)
AI/No Pay	-1.63	-12.57	-14.24	-10.66	-0.64***	-0.58**	-0.39	-0.78*
	(5.45)	(7.65)	(10.06)	(11.49)	(0.19)	(0.29)	(0.38)	(0.44)
Pos Feedb x Human/Pay		16.28	5.15	26.54		-0.11	-0.15	-0.10
		(12.88)	(16.87)	(19.63)		(0.40)	(0.56)	(0.58)
Pos Feedb x AI/Pay		16.19	19.57	13.94		-0.33	-0.48	-0.30
		(12.83)	(17.08)	(18.66)		(0.43)	(0.60)	(0.62)
Pos Feedb x AI/No Pay		22.09**	10.43	31.25*		-0.18	-0.12	-0.26
		(10.83)	(14.38)	(16.33)		(0.36)	(0.49)	(0.55)
1=Feedback positive		-4.12	2.24	-7.95		1.13***	0.90**	1.36***
		(9.05)	(11.91)	(13.69)		(0.30)	(0.42)	(0.45)
Observations	1331	1331	659	672	1326	1326	658	668
Control Mean	65.61	65.61	62.62	68.75	6.96	6.96	6.86	7.06
Control SD	72.97	72.97	67.37	78.55	2.56	2.56	2.44	2.68
Adj R-Square	0.03	0.03	0.04	0.04	0.01	0.08	0.07	0.09
Demographic Controls	Y	Y	Y	Y	N	Y	Y	Y
Mission Alignment	Pooled	Pooled	Low	High	Pooled	Pooled	Low	High

Notes: The dependent variable in Cols 1-2 is the time (in seconds) people spent writing campaign slogans in the follow-up session, winsorized at the 95th percentile. Col 2-3, 4-5, and 6-7 measure whether people perceive the process to be effective, fair, and transparent respectively. The independent variables are indicators for the treatment groups with the Human/No Pay control group left out. Low/High Mission Alignment corresponds to above/below-median baseline support for government food shortage effort. Participant Demographic Controls are included. Standard errors in parentheses are robust. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A10: Effect on Donations and Effort in Optional Task

	Donation		Optional: Time	
	(1)	(2)	(3)	(4)
Human/Pay	0.17	-0.57	3.10	-0.49
	(1.62)	(2.24)	(3.99)	(5.37)
AI/Pay	-1.67	-2.54	0.45	-8.26
	(1.62)	(2.22)	(3.84)	(5.38)
AI/No Pay	-1.20	0.26	-5.76*	-9.68**
	(1.39)	(1.89)	(3.35)	(4.58)
High Alignm. x Human/Pay		1.27		7.01
		(3.21)		(7.95)
High Alignm. x AI/Pay		1.36		16.61**
		(3.21)		(7.70)
High Alignm. x AI/No Pay		-3.00		7.72
		(2.75)		(6.72)
High Alignment		5.94***		1.62
		(2.25)		(5.62)
Observations	1487	1487	1491	1491
Control Mean	26.4	26.4	59.3	59.3
Control SD	19.4	19.4	48.1	48.1
Adj R-Square	0.00	0.02	0.02	0.03
Demographic Controls	Y	Y	Y	Y
Low: AI/Pay vs. Human/Pay	0.26	0.39	0.50	0.15
Low: AI/No Pay vs. AI/Pay	0.74	0.15	0.06	0.76
High: AI/Pay vs. Human/Pay		0.41		0.74
High: AI/No Pay vs. AI/Pay		0.43		0.02

Notes: The dependent variable in Cols 1–2 is the amount donated to the non-profit organization (in cents), and in Cols 3–4 is the time (in seconds) spent on the optional hashtag-writing task (winsorized at the 95th percentile). High Alignm. is an indicator variable for individuals with above-median baseline support for government food shortage efforts. Participant Demographic Controls are included in all specifications. Robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A11: Stated Reasons for Reduced Willingness to Donate to AI-Using Organizations

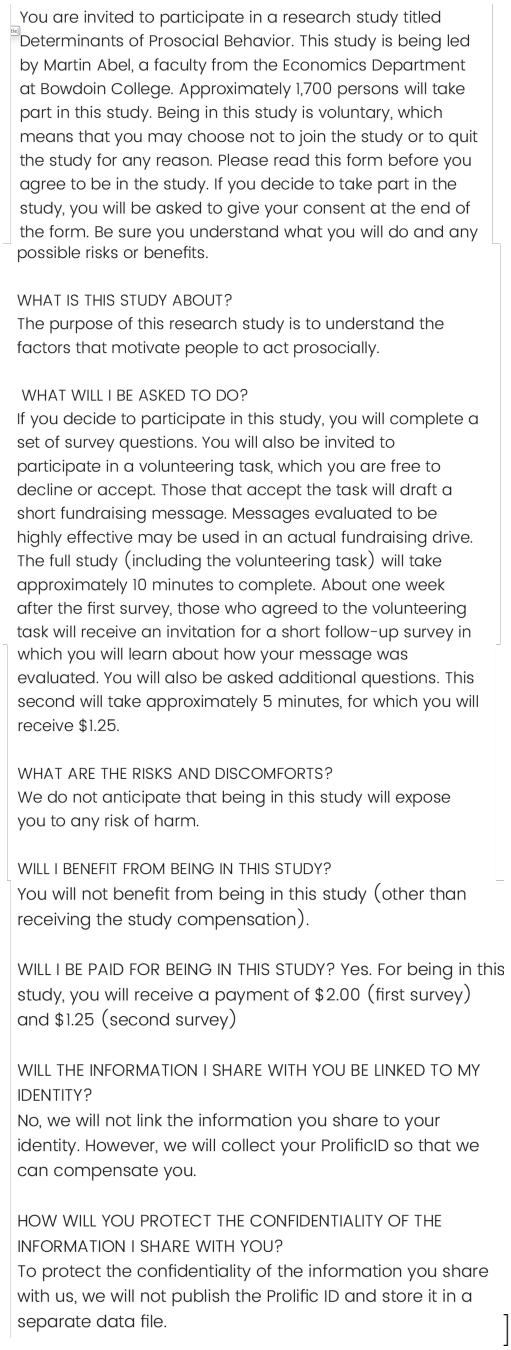
	N	Proportion
Inauthenticity/Lack of Genuineness	118	0.404
Low Effort/Laziness	55	0.188
General AI Opposition/Ideology	44	0.151
Trust/Manipulation Concerns	45	0.154
Lack of Emotional Understanding	34	0.116
Effectiveness/Quality Concerns	30	0.103
Resource Misallocation/Hypocrisy	29	0.099
Displacement of Human Workers	21	0.072
Other/Multiple Reasons	3	0.010

Notes: Sample includes 292 respondents who indicated they would be less likely to donate to organizations using AI. Respondents provided open-ended explanations which were independently coded into categories. Multiple categories could be assigned per response. Proportions therefore sum to more than 1.0.

B Online Appendix

B.1 Experimental Protocol

Figure B1: Consent Form (part 1)



You are invited to participate in a research study titled Determinants of Prosocial Behavior. This study is being led by Martin Abel, a faculty from the Economics Department at Bowdoin College. Approximately 1,700 persons will take part in this study. Being in this study is voluntary, which means that you may choose not to join the study or to quit the study for any reason. Please read this form before you agree to be in the study. If you decide to take part in the study, you will be asked to give your consent at the end of the form. Be sure you understand what you will do and any possible risks or benefits.

WHAT IS THIS STUDY ABOUT?
The purpose of this research study is to understand the factors that motivate people to act prosocially.

WHAT WILL I BE ASKED TO DO?
If you decide to participate in this study, you will complete a set of survey questions. You will also be invited to participate in a volunteering task, which you are free to decline or accept. Those that accept the task will draft a short fundraising message. Messages evaluated to be highly effective may be used in an actual fundraising drive. The full study (including the volunteering task) will take approximately 10 minutes to complete. About one week after the first survey, those who agreed to the volunteering task will receive an invitation for a short follow-up survey in which you will learn about how your message was evaluated. You will also be asked additional questions. This second will take approximately 5 minutes, for which you will receive $1.25.

WHAT ARE THE RISKS AND DISCOMFORTS?
We do not anticipate that being in this study will expose you to any risk of harm.

WILL I BENEFIT FROM BEING IN THIS STUDY?
You will not benefit from being in this study (other than receiving the study compensation).

WILL I BE PAID FOR BEING IN THIS STUDY? Yes. For being in this study, you will receive a payment of $2.00 (first survey) and $1.25 (second survey)

WILL THE INFORMATION I SHARE WITH YOU BE LINKED TO MY IDENTITY?
No, we will not link the information you share to your identity. However, we will collect your ProlificID so that we can compensate you.

HOW WILL YOU PROTECT THE CONFIDENTIALITY OF THE INFORMATION I SHARE WITH YOU?
To protect the confidentiality of the information you share with us, we will not publish the Prolific ID and store it in a separate data file.

Figure B2: Consent Form (part 2)

WHERE WILL THE INFORMATION I SHARE WITH YOU BE STORED?
The information that you share with us will be stored online on a password-protected OneDrive folder. We will make every effort to ensure the security of the data that you share with us. As you are likely aware, it is impossible to completely guarantee the security of data transmitted or stored electronically. In addition to the places where we will store your information (described above), your data may be saved on backups and activity logs (i.e., records of internet activity).

WHO WILL HAVE ACCESS TO THE INFORMATION I SHARE WITH YOU?
Only the researcher team will have access to the information you share with us. An anonymized version of the data set may later be publicly available.

WHAT WILL HAPPEN TO THE INFORMATION I SHARE AFTER THE STUDY IS OVER?
We will retain the information you share with us for at least five years. Data from this study may be used for future research studies or shared with other researchers or the research community at large to advance our understanding of this topic, without additional informed consent from you. We will remove or code any information that could identify you before data are shared to ensure that, by current standards and known methods, no one will be able to identify you from the information we share. Despite these efforts, we cannot guarantee the anonymity of your personal data.

DO I HAVE TO BE IN THIS STUDY?
No. You may choose not to take part in this study for any reason. If you join this study, you may change your mind and stop participating in the study at any time and for any reason. In either case, you will not lose any benefits to which you are otherwise entitled. You may also skip any questions or procedures that you do not wish to take part in.

WHO SHOULD I CONTACT WITH QUESTIONS?
If you have questions or problems related to this study, you should contact Martin Abel at m.abel@bowdoin.edu or at [redacted] If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Bowdoin College Institutional Review Board (IRB) at irb@bowdoin.edu. A copy of this document can also be downloaded here. CONSENT If you have read the above information, have received answers to any questions that you have, and are 18 years of age or older and you wish to consent to take part in the study, please check the box ("I agree to participate.") below.

- ○ I agree
- ○ I don't agree (will exit survey)

Figure B3: Volunteering Decision

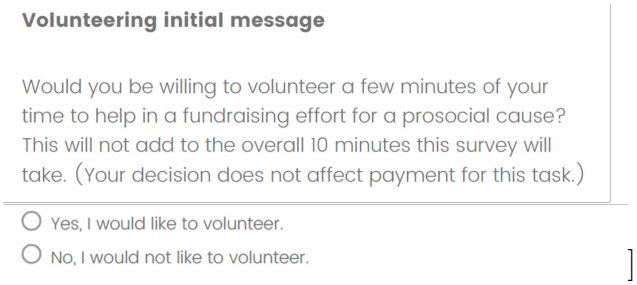


Figure B4: Video Transcript

They line up long before opening. Car, after car after car. Each person here for the most basic of human necessities. Food. But ever since the pandemic demand has been spiking. Experts say inflation is largely to blame. Higher food, housing, and overall living expenses ravaging the budget of lower income households, an unwinding of pandemic financial assistance, and the resumption in student loan payments are also driving more people to seek help.

Food banks across the United States have been left scrambling, forced to do more with less as food insecurity rises in America. According to a recent USDA report, one in eight households struggled to put food on the table last year. Significantly higher than the year before. If we're not able to feed our children, we're not able to secure a healthy nation. Nation building one car and one meal at a time.

Figure B5: Human Evaluation

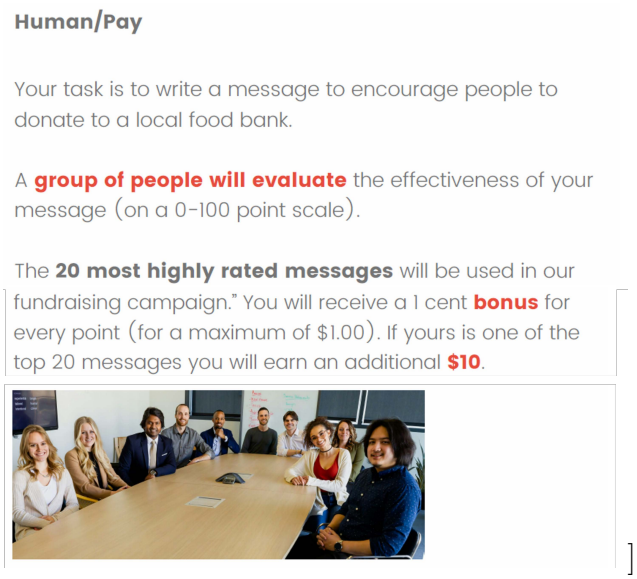


Figure B6: AI Evaluation

Treatment AI/No pay

Your task is to write a message to encourage people to donate to a local food bank.

An **Artificial Intelligence (AI) algorithm will evaluate** the effectiveness of your message (on a 0-100 point scale).

The **20 most highly rated messages** will be used in our fundraising campaign.

Figure B7: AI Evaluation with Human Criteria

AI w/ Human Characteristics

Your task is to write a message to encourage people to donate to a local food bank.

An **Artificial Intelligence (AI) algorithm will evaluate** the effectiveness of your message (on a 0-100 point scale). The AI's algorithm is **instructed to be "human-like"** and make assessments based on how sympathetic, emotionally impactful, and relatable your message is.

The **20 most highly rated messages** will be used in our fundraising campaign.

Figure B8: Optional Task

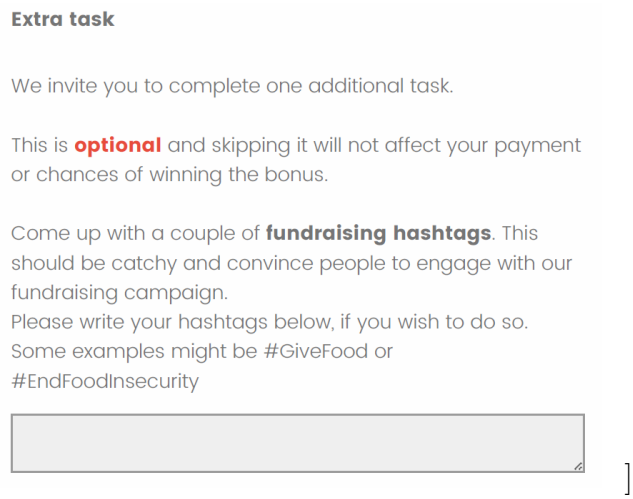


Figure B9: Process Perception

Figure B10: Donation Decision

As mentioned, an Artificial Intelligence algorithm will evaluate the message you drafted to identify effective fundraising strategies we evaluate fundraising messages for a **non-for-profit organization** that fights food shortage.

In appreciation of your time, we want to give you a **bonus of 50 cents**.

You have the option to **donate** part of your bonus to this organization.

How much of your bonus would you like to donate?

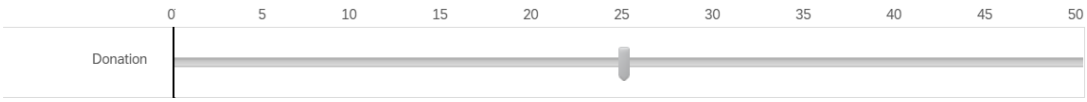


Figure B11: Feedback (Example: Human Evaluators)

We now want to share with you some feedback. As a reminder, you were asked to write a fundraising message to collect donations to fight food shortage in the U.S.

Your message was reviewed by **a group of people**.
The rating of the message was **above average**; however, it was **not selected** to be used in the fundraising drive (and you did not receive the $10 bonus).

We still want to thank you for your effort. We also want to acknowledge that the evaluation process is noisy and may not capture the true effectiveness of the message.

We now want to share with you some feedback. As a reminder, you were asked to write a fundraising message to collect donations to fight food shortage in the U.S.

Your message was reviewed by **a group of people**.
Unfortunately, the rating of the message was **below average** and was **not selected** to be used in the fundraising drive (and you did not receive the $10 bonus).

We still want to thank you for your effort. We also want to acknowledge that the evaluation process is noisy and may not capture the true effectiveness of the message.

Figure B12: Rating Instructions