OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

The Effect of Artificial Intelligence Feedback and Social Norms on Performance

Katherine Wu
katherinewu653@gmail.com

ABSTRACT

As artificial intelligence (AI) systems become common sources of feedback in workplaces and classrooms, understanding how people respond to AI-generated evaluations is increasingly important, particularly whether recipients distrust feedback simply because it comes from a machine.

Using randomized controlled experiments, this study examines how two factors simultaneously shape human performance: the attributed source of feedback (from AI versus from a human) and social norm information, namely information about peers' performance.

The findings suggest that social norm information raises effort across all settings, whether feedback is attributed to an AI or to a human, whereas the source of information alone has little effect. Social comparison, however, reveals how recipients respond to negative feedback. While negative feedback incentivizes greater effort if recipients are also provided with a social comparison benchmark, it does not have a statistically detectable effect in its absence. This pattern does not depend on the messenger though, when AI delivered feedback and human delivered feedback is paired with social norms, positive and negative feedback have similar effects.

Together, these results challenge the assumption that people inherently distrust AI-generated evaluations and suggest that information about relative performance affects recipients' behavior far more than the messenger's identity, demonstrating how powerful the influence of comparative performance information is in shaping motivation and productivity.

INTRODUCTION

Artificial intelligence (AI) has become common in people's everyday lives, whether during assessments used by teachers and workplaces or automated teaching platforms and workplace performance trackers (Schmidt et al., 2025).

One increasingly common role for these systems is the delivery of feedback, or evaluative information about a person's performance. Feedback can be motivational, but it also serves as social information people use to understand norms and comparisons among others (Brudner et al., 2022). A particularly important form of such information is the social norm: information about how one's peers perform or what strategies they use, which provides a benchmark against which individuals judge their own behavior.

As AI takes on more of this evaluative role, investigating how people react to feedback from artificial intelligence is vital to understanding the social and behavioral effects of technological integration into daily life.

Research suggests that the identity of the messenger matters. Some individuals may exhibit algorithm aversion, or the tendency to prefer human decision-makers to algorithms, reacting negatively to AI feedback even when it is accurate and consistent (Burton et al., 2020; Tong et al., 2021). A separate and growing literature shows that the content of feedback matters as well, particularly the social information it carries. Studies of social norms find that providing individuals with information about average peer performance increases motivation, individual performance, and adherence to norms, and that such norms influence cooperation and effort in organizations and educational contexts (Fehr & Fischbacher, 2004; Festré, 2009; Gächter et al., 2017; Stewart et al., 2012). More recently, researchers have turned to the behavioral effects of AI-mediated and algorithmic feedback itself, though much of this work centers on algorithmic decision-making or high-stakes evaluations (Tong et al., 2021; Bai et al., 2022; Lee et al., 2023).

These two literatures have largely developed in parallel. Work on algorithm aversion emphasizes who delivers feedback, while work on social norms emphasizes what the feedback contains. However, there is little research on how the two interact. Specifically, it remains unclear how the content of feedback, particularly social norm information, combines with attribution to an AI versus a human source to shape behavior, and whether the influence of the messenger persists once peer-comparison information is introduced.

This paper addresses that gap by investigating how attributions to artificial intelligence versus human supervisors, together with the presence or absence of social norm information, jointly affect motivation and individual performance. Participants complete an activity for which they are randomly assigned to receive performance feedback that either is or is not attributed to an AI source, and that either contains or does not contain information about the average performance of their peers. Feedback in this study takes two valences: negative feedback signals that performance fell short and offers strategies for improvement, while positive feedback affirms that performance was strong or improving. I measure both behavioral and subjective responses to this feedback including effort and performance on the task as well as perceptions of fairness and legitimacy. This design allows for the isolation of the causal effects of feedback source and norm information on motivation, enabling a direct test of whether the messenger or the message more strongly shapes behavior.

METHODS

The research examines how the source of feedback and the presence of social norm information affect individual performance in an experimental setting. The study was executed by using an online survey platform, Qualtrics, where participants performed a word formation task that assessed their cognitive abilities and feedback reaction time. Each participant received a random group of six letters and had to generate as many valid English words as possible within a two-minute time limit. This task helps to evaluate effort and linguistic problem-solving skills through an objective assessment that produces measurable results during multiple sessions.

The experiment consisted of three rounds. Round 1 provided a baseline measure of performance, while Rounds 2 and 3 introduced different types of feedback depending on the treatment group. Participants were randomly assigned to one of five treatment conditions. The control group received no feedback between rounds. The second group received feedback that was said to be created from an artificial intelligence system ("AI Feedback"), while the third group was told the feedback came from a human evaluator ("Human Feedback"). The fourth group received AI feedback that included positive and negative details about peers' strategies and performance ("AI Feedback and Social Norms"). The fifth group received human feedback with positive and negative information about peers' strategies and performance ("Human Feedback and Social Norms"). The latter treatments introduced an element of normative comparison which enabled the study to explore how people act when performance data gets presented within social groups.

Feedback could be positive (participants performed above-average) or negative (participants performed below-average). Negative feedback indicated that the participant had struggled to form valid words, noted the point penalty for incorrect answers, and offered strategies for improvement (for example, beginning with short three- to four-letter words and scanning for common prefixes and suffixes). Positive feedback affirmed that the participant had performed well or improved relative to the previous round (for example, noting that they had generated more valid words than before). In the two social-norm conditions, this feedback was supplemented with descriptive information about peers, for instance, a negative message might note that most players use short words to gain quick points and that not doing so puts the participant behind, while a positive message might share the strategies most participants found effective.

Of the participants, 64.4% were female and 35.6% were male. Participants ranged in age from 15 to 60 years, with a mean age of 25.81 years. Age was included as a control variable; its estimated association with score improvement was negative but not statistically significant. With respect to education, 40.7% had a high school education or below and 59.3% had education above the high school level.

The analysis includes three demographic control variables collected from participants. Sex is coded as a binary variable equal to one for female participants and zero for male participants; no participants in the sample identified outside these two categories. Age is measured in years. Education is coded as a binary variable equal to one for participants with education above the high school level and zero for those with a high school education or below ("Above High School"). Binary variables of this kind, often termed indicator or dummy variables, take a value of one when a condition is met and zero otherwise, allowing categorical characteristics to be included in the regression.

All responses were verified with a standardized English dictionary, ensuring objective scoring and eliminating the possibility of human grading bias. The resulting dataset is mainly numerical, recording each participant of the number of valid words produced in each round, the assigned treatment condition, and limited demographic characteristics such as age and gender. The main outcome variable is the total number of valid words produced per round, which serves as a measure of individual performance.

A key strength of this design is the use of random assignment. Random assignment minimizes concerns around selection bias or omitted variable confounding. By comparing individual changes in performance across rounds, the experiment also controls for unobserved heterogeneity in baseline ability. These features help to strengthen the validity of the findings.

RESULTS

Empirical Strategy

This section examines how different feedback sources and social information affect individual performance. The goal is to identify whether feedback labeled as coming from AI or a human, with or without social norms, influences effort and improvement across rounds. To investigate this, I estimate the following equation:

$$\triangle \mathit{Score}_{t,i} = \alpha_0 + \alpha_1\, \mathit{Treatment}_{t,i} + \alpha_2 X_{t,i} + \varepsilon_{t,i}$$

Where $\triangle \mathit{Score}_{t,i}$ is the difference in test scores between Round 1 and a later round (e.g., where *t* is either rounds 2 or 3, or an average of the two rounds) for participant *i*. $\mathit{Treatment}_{t,i}$ represents the treatment received by participants *i* (e.g., AI disclosure, Human disclosure, with/without social norms). The treatment indicator equals one for participants assigned to the condition of interest and zero for participants in the no-feedback control group; α_1 therefore measures the effect of that condition relative to receiving no feedback. Treatment in Round 2 is negative feedback and treatment in Round 3 is positive feedback. $X_{t,i}$ is a vector of controls (e.g., gender, age, education). $\varepsilon_{t,i}$ is the error term for individual *i*. α_1 is the coefficient of interest, plotted in the tables of results that follow. Specifically, α_1 gives the estimated average difference in score improvement between participants in the treatment condition and those in the control group, holding sex, age, and education constant.

The Direct Effect of AI and Human Disclosure

Table 1 reports the results for AI Disclosure. Across all specifications, the estimated coefficients are negative relative to the control group, with a decrease of 2.3 points for any feedback, 3.6 points for negative feedback, and 1.1 points for positive feedback. None of these effects are statistically significant. The low explanatory power of the model ($R^2 = 0.07$ for the pooled specification) confirms that disclosing AI as the feedback source neither improves nor alters performance meaningfully. If anything, the direction of the estimates suggests a small reduction in effort when participants are aware that feedback originates from an algorithm.

Table 1: The Effect of AI Disclosure

Dep. Variable	Feedback		
	Any	Negative	Positive
	(1)	(2)	(3)
AI Disclosure	-2.348	-3.587	-1.108
	(3.947)	(5.035)	(4.169)
R-Squared	0.07	0.813	0.036
Observations	22	22	22

Notes: Demographic controls include sex (an indicator equal to one for female participants, zero for male), age (in years), and education (an indicator equal to one for participants with education above high school, zero otherwise). Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Data source: own experiment.

These findings indicate that AI feedback, on its own, lacks social salience. The estimates are consistent with AI feedback that, when not socially framed, does not reliably trigger social comparison or reputational motives, although the absence of a statistically significant effect means the data cannot confirm such a reduction. In behavioral economics terms, this points to an absence of "social utility" in AI interactions: individuals do not internalize feedback from AI in the same way as they do from other people (Boyd & Markowitz, 2026).

Table 2 presents the results for Human Disclosure. The coefficients are again negative and statistically indistinguishable from zero. Relative to the control group, performance decreases by 1.5 points on average, by 2.2 points under negative feedback, and by 0.9 points under positive feedback. R^2 values remain low across all models, showing that human labeling alone does not explain variation in scores. These results mirror the AI effect, suggesting that simple disclosure, without a social frame, fails to alter behavior.

Table 2: The Effect of Human Disclosure

Dep. Variable	Feedback		
	Any	Negative	Positive
	(1)	(2)	(3)
Human Disclosure	-1.535	-2.208	-0.861
	(3.905)	(4.453)	(4.169)
R-Squared	0.165	0.099	0.210
Observations	23	23	23

Notes: Demographic controls include sex (an indicator equal to one for female participants, zero for male), age (in years), and education (an indicator equal to one for participants with education above high school, zero otherwise). Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Data source: own experiment.

The lack of significant response to Human Disclosure implies that recognition of a human feedback source is not sufficient to influence effort. Behavioral theories emphasize that feedback affects outcomes through social pressure or perceived expectations (Weiß et al., 2024). In this setting, human feedback unaccompanied by social information does not activate these mechanisms.

The Role of Social Norms

Table 3 introduces the interaction between AI Disclosure and Social Norms. Once social comparison is added, the treatment effect turns positive and statistically significant. The combined condition increases performance by 5.2 points overall, with a 5.7-point rise under negative feedback and a 4.7-point rise under positive feedback. The strongest effect appears in the negative feedback condition, suggesting that social norms mitigate the discouraging effect of AI feedback and transform it into a motivator for improvement.

Table 3: The Effect of AI Disclosure and Social Norms

Dep. Variable	Feedback		
	Any	Negative	Positive
	(1)	(2)	(3)
AI Disclosure & Social Norms	5.206*	5.713*	4.699
	(2.723)	(3.113)	(2.889)
R-Squared	0.176	0.192	0.114
Observations	30	30	30

Notes: Demographic controls include sex (an indicator equal to one for female participants, zero for male), age (in years), and education (an indicator equal to one for participants with education above high school, zero otherwise). Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Data source: own experiment.

This result highlights that social framing, not feedback tone, drives behavioral change. Social norms create a reference point, activating competition or conformity motives even when feedback is delivered by AI. Individuals appear motivated by information about peer performance, regardless of whether their own feedback is positive or negative. The interaction is asymmetric, however: social norms paired with AI disclosure work better when feedback is negative, suggesting that underperformance relative to peers prompts greater adjustment when mediated by AI.

Table 4 examines Human Disclosure with Social Norms. The combined treatment increases performance by 6.6 points, representing a 146% improvement relative to the control mean of 4.5 points. The effect is consistent across negative and positive feedback, with 6.5- and 6.6-point increases respectively. These effects are statistically significant and substantially larger than those observed for the AI-plus-norms condition.

Table 4: The Effect of Human Disclosure and Social Norms

Dep. Variable	Feedback		
	Any	Negative	Positive
	(1)	(2)	(3)
Human Disclosure & Social Norms	6.572**	6.508**	6.637**
	(2.663)	(3.560)	(2.803)
R-Squared	0.403	0.364	0.291
Observations	23	23	23

Notes: Demographic controls include sex (an indicator equal to one for female participants, zero for male), age (in years), and education (an indicator equal to one for participants with education above high school, zero otherwise). Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Data source: own experiment.

The results suggest that social norms are the central driver of performance gains, not the valence of the feedback. The near-identical increases across positive and negative conditions indicate that individuals respond primarily to the presence of social information. The effect is not uniform across disclosure types: social norms have a stronger and more consistent influence when paired with human feedback. While AI with norms boosts performance mainly after negative feedback, human feedback with norms raises performance across all contexts. The largest coefficient, 6.6 points, appears for the human-plus-norms treatment, confirming that behavioral responses intensify when social and interpersonal cues are combined.

Robustness Test

To assess the stability of the findings, robustness checks were conducted. The four treatment groups (treatment_1– treatment_4) were aggregated into a single dummy indicator (treatment_all = 1 if any of the treatment group is included and treatment_all = 0 if control group is included) and were compared against the control group. The aggregated dummy indicator allows the estimation of the average effect of receiving any treatment relative to the control condition.

The results in Table 5 indicate that participants exposed to any treatment scored, on average, approximately 35 points higher than those in the control group and the coefficient for treatment_all is statistically significant ($p < 0.05$). This provides strong evidence that the treatment effect is not spurious when considered collectively.

Table 5: Robustness Test of All Four Treatment Groups

Source	SS	df	MS
Model	3408.06136	1	3408.06136
Residual	18143.4083	31	585.271237
Total	21551.4697	32	673.483428

Number of obs	=	33
F(1,31)	=	5.82
Prob > F	=	0.0219
R-squared	=	0.1581
Adj R-squared	=	0.131
Root MSE	=	24.192

Points	Coefficient	Std.err.	t	P> |t|	[95% Conf. interval]	
Treatment_all_cons	35.35	14.64921	2.41	0.022	5.472732	65.22727
	23.333333	13.96748	1.67	0.105	-5.153524	5182019

DISCUSSION

The results show that neither AI nor human disclosure alone affects performance. Both yield coefficients close to zero, with large standard errors and high p-values, meaning their effects are statistically indistinguishable from zero. The null result provides information that social frameworks should exist for people to work based on informational precision. People process feedback as social-based information when they create their decision-making process based on social-based information. The participants show no motivation to change their performance or work ethic simply because they receive feedback from an AI system or human evaluator. The study shows that accuracy and perceived fairness do not drive behavior change or produce an effect on work output.

The introduction of social norms results in different outcomes. The disclosure of information by both AI systems and humans shows substantial growth when peer performance information is exposed. The results demonstrate that social elements drive human behavior whereas technological elements do not. Feedback only influences performance when it provides information about an individual's current position in comparison to others or about expectations that other people hold. In the absence of social norms, feedback provides information about past rounds but does not alter performance results or perceived motivation that would justify exerting additional effort to perform better. Positive and negative feedback results in similar outcomes for the human feedback and social norms treatment, supporting this study's interpretation. The presence of social norms creates a motivational effect which persists through all feedback methods, proving that social context has stronger influence than message positivity or negativity.

The differences between AI and human conditions help explain why these results make sense even if they depart from the theoretical expectation. AI feedback with social norms raises performance mainly under negative feedback. Participants may view AI criticism as impersonal and less threatening, using it as a neutral signal to catch up with peers. By contrast, human feedback combined with social norms produces strong effects across both positive and negative feedback. This pattern suggests that participants attribute greater social meaning to human evaluations, interpreting them as signals of encouragement. Human feedback carries implicit social expectations even when anonymized, while AI feedback is interpreted as purely mechanical. In short, when a human and a social frame coexist, motivation increases across the board; when AI delivers feedback, social comparison only helps when individuals seek to correct underperformance.

From a behavioral economics perspective, these findings reveal that performance in feedback settings depends on social motivation, not informational precision. People respond to relative standing and perceived expectations, not to the technical quality of the feedback. Systems designed to enhance effort or learning should embed feedback within a social context, regardless of the source.

CONCLUSION

The results show that feedback labeled as AI or human produces no meaningful change in performance. The coefficients are small and statistically indistinguishable from zero, so disclosure alone does not influence effort. Social norms shift the outcome. Once peer information appears, performance rises across treatments. The pattern indicates that social comparison drives behavior, not the identity of the feedback source.

Social norms also interact differently with AI and human disclosure. When paired with AI, norms are most effective under negative feedback. Participants respond strongly when they learn they performed below the group average. When paired with human disclosure, norms raise performance under both positive and negative feedback, which produces larger and more stable gains. The strongest effects occur when norms are linked to human feedback, which suggests that peer information becomes more influential when tied to a human evaluator.

These findings point to clear policy implications. Human or AI feedback alone does not shift performance. Social norms are powerful and should be included when the goal is to raise effort. The pairing of norms with human feedback produces the largest improvements. The pairing of norms with AI feedback is also effective, especially under negative comparisons. Feedback systems should emphasize relative performance, since peer effects drive the shifts in behavior. People don't necessarily change when they are told how they are doing; they change when they see how they measure up.

REFERENCES

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., Mckinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., & Kerr, J. (2022). *Constitutional AI: Harmlessness from AI Feedback*. https://arxiv.org/pdf/2212.08073

Boyd, R. L., & Markowitz, D. M. (2026). Artificial Intelligence and the Psychology of Human Connection. Perspectives on Psychological Science. https://doi.org/10.1177/17456916251404394

Brudner, E. G., Fareri, D. S., Shehata, S. G., & Delgado, M. R. (2022). Social feedback promotes positive social sharing, trust, and closeness. *Emotion*, *23*(6). https://doi.org/10.1037/emo0001182

Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2). https://doi.org/10.1002/bdm.2155

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Festré, A. (2009). INCENTIVES AND SOCIAL NORMS: A MOTIVATION-BASED ECONOMIC ANALYSIS OF SOCIAL NORMS. *Journal of Economic Surveys*, *24*(3), 511–538. https://doi.org/10.1111/j.1467-6419.2009.00594.x

Gächter, S., Gerhards, L., & Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, *97*, 72–86. https://doi.org/10.1016/j.euroecorev.2017.06.001

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., & Rastogi, A. (2023, September 1). *RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback*. ArXiv.org. https://doi.org/10.48550/arXiv.2309.00267

Schmidt, D. A., Alboloushi, B., Thomas, A., & Magalhaes, R. (2025). Integrating artificial intelligence in higher education: perceptions, challenges, and strategies for academic innovation. *Computers and Education Open*, *9*, 100274. https://doi.org/10.1016/j.caeo.2025.100274

Stewart, G. L., Courtright, S. H., & Barrick, M. R. (2012). Peer-based control in self-managing teams: Linking rational and normative influence with individual and group performance. *Journal of Applied Psychology*, *97*(2), 435–447. https://doi.org/10.1037/a0025303

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*(9), 1600–1631. https://doi.org/10.1002/smj.3322

Weiß, M., Gollwitzer, M., & Johannes Hewig. (2024). Social influence and external feedback control in humans. F1000Research, 12, 438–438. https://doi.org/10.12688/f1000research.133295.3