Journal of Economic Behavior and Organization 217 (2024) 692–704

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research Paper

Additive vs. subtractive earning in shared human-robot work environments ☆

Bnaya Dreyfuss [a], Ori Heffetz [b,c,d,*], Guy Hoffman [e], Guy Ishai [b], Alap Kshirsagar [e]

[a] Department of Economics, Harvard University, United States of America
[b] Bogen Family Department of Economics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel
[c] S.C. Johnson Graduate School of Management, Cornell University, United States of America
[d] National Bureau of Economic Research, United States of America
[e] Sibley School of Mechanical and Aerospace Engineering, Cornell University, United States of America

A B S T R A C T

The performance of robots working alongside humans might positively or negatively affect humans' earnings, depending on the economic setting. In a new real-effort lab experiment, we study the impact of economic conditions in hybrid human-robot workplaces on workers' effort provision and attitudes. In a previous *subtractive*-earnings experiment (Kshirsagar et al., 2019), subjects' expected earnings negatively depend on a robot's performance, while in our new *additive*-earnings experiment, they depend on the robot's performance positively. Both experiments are guided by a past human-human experiment and by a model of expectations-based reference-dependent preferences. As the theory predicts and as previously found, increasing robot performance discourages effort under subtractive earnings—but, as the theory also predicts and as we find here, this effect disappears and perhaps reverses under additive earnings. Additionally, increasing robot performance negatively affects subjects' perceptions of themselves and of their robotic coworker under subtractive earnings, but we find that these effects weaken or reverse under additive earnings. These findings suggest a relationship between workers' earning structures and robots' performance that should be considered when designing hybrid workplaces.

1. Introduction

Robots, algorithms and artificial intelligence (AI) are being used alongside human workers in an increasing variety of work environments. As new forms of automation, they can have various direct effects on the production process. For example, they can be viewed as making both capital and labor more productive (e.g., Bessen, 2020; Nordhaus, 2021) or, alternatively, as causing worker displacement, by replacing tasks previously performed by humans (e.g., Acemoglu and Restrepo, 2018).

However, beyond their direct effects on production, the mere presence of robots who work alongside human workers can also indirectly affect human labor. Specifically, robots can affect workers' behavior (e.g., effort and productivity) and perceptions (e.g., self-competence and well-being). These effects are less well understood, and may depend on how a robot's performance affects expected human earnings.

We analyze two real-effort lab experiments (data from Kshirsagar et al., 2019, $N = 60$; and a new experiment, $N = 60$), both adapting Gill and Prowse's (2012) human-human design to a human-robot design, to study how the combination of a robot's perfor-

☆ We thank Gabriela Cohen-Hadid, Julia Katz, Ofer Rubinstein, and Song Ye for excellent research assistance. This project was supported in part by the Planning and Budgeting Committee and the Israel Science Foundation (grants no. 1821/12 and 2968/21).

* Corresponding author at: S.C. Johnson Graduate School of Management, Cornell University, United States of America.
E-mail addresses: bdreyfuss@g.harvard.edu (B. Dreyfuss), oh33@cornell.edu (O. Heffetz), hoffman@cornell.edu (G. Hoffman), guy.ishai@mail.huji.ac.il (G. Ishai), ak2458@cornell.edu (A. Kshirsagar).

https://doi.org/10.1016/j.jebo.2023.11.024
Received 17 April 2023; Received in revised form 21 November 2023; Accepted 22 November 2023

Available online 12 December 2023
0167-2681/© 2023 Elsevier B.V. All rights reserved.

mance and the sign of its impact on a human worker's expected earnings affects the human worker's outcomes. Holding constant the human's expected return to marginal effort, a high-performing robot working next to the human negatively affects both the human's effort and her perceptions when she expects her earnings to decrease with the robot's output (Kshirsagar et al., 2019). Adding a new variation to the original experimental design, we find that these negative effects could be mitigated and perhaps reversed when the human expects her earnings to increase with the robot's output.

We make two main contributions. First, we study the *economics of Human-Robot Interaction* (HRI), i.e., how the economic incentives which govern human-machine interaction in the workplace shape human workers' response to the performance of robot coworkers—both in terms of their real-effort decisions and their subjective perceptions. Second, our effort-provision results are in line with recent models of expectations-based reference-dependent (EBRD) preferences. We use a novel empirical-design variation that merely flips the sign of the robot performance's effect on human expected earnings across two otherwise identical settings.

This variation allows for a clean test of the EBRD model, which has opposite predictions on how agents react to an increase in the robot's performance across the two scenarios.

To illustrate, consider a human and a robot working side-by-side in an automobile factory. The human assembles car body parts and hands them over to a robot that welds the parts together. The factory has a bonus incentive program, which pays out an end-of-year bonus based on a number of considerations, such as total factory output and individual worker evaluation.

We can compare two opposite scenarios. In one, the bonus depends only on the total factory output. In this case, both increased worker assembly effort e_h and increased robot welding performance e_r ("robot effort") increase the chance p of getting the bonus. Call that the *additive*-earning scenario.

In the opposite scenario, workers' contributions are evaluated individually for the bonus. In this case, a worker might be seen as slowing down the team if they perform their assembly task slower than the robot's welding speed. If this is true, increased robot effort e_r now decreases the worker's chance p of getting the bonus. Call that the *subtractive*-earning scenario. Notably, in both scenarios, additional human effort has the same positive effect on their chance for getting the bonus.

The factory example highlights additiveness vs. subtractiveness as a result of an incentive scheme design. However, additiveness and subtractiveness could also result from an exogenous state resulting from given economic conditions. For example, a freelance graphic design marketplace could have both human designers and AI systems offer design work for customers. Better AI models could draw more customers to the marketplace, increasing the chance for human designers to earn, thus creating an additive scenario. Conversely, AI design algorithms could perform so much better than a human designer that they take away potential customers from the human, creating a subtractive scenario. Note that, in this scenario, too, the marketplace could be made additive by design, by charging a fixed subscription fee rather than a pay-for-design scheme and paying human designers a fixed fraction of overall fees.

In this paper, we ask: What are the effects on the human's effort and perceptions of robot performance under these two opposite earning scenarios?

In line with the examples above, we combine data from two experiments to investigate how the compensation scheme shapes the reaction of participants to their robotic co-workers' performance. We are interested in two types of outcomes: chosen effort and subjective perceptions. As we show below, the EBRD model has precise predictions on how workers react to changes in the robot's performance, depending on the compensation scheme. Moving beyond this main outcome, we also collect subjective perceptions to investigate how workers' attitudes—which we believe are important in the context of technology adoption, but are not part of the EBRD model—are shaped by the economic setting.

Results from the first experiment are presented in Kshirsagar et al. (2019), a conference proceeding geared towards the HRI community that presented new insights into an unexplored area of HRI research—incentivized real-effort experiments with physical robots. In the next section, we review the literature and discuss the contribution of this paper (and in particular, the addition of the new experiment) relative to our previous work.

We present the experimental design in Section 2. In the two (subtractive and additive) experiments, each human participant works alongside a robotic arm for the chance of winning a monetary prize, for ten rounds. Each round lasts two minutes, in which the human and robot's separate tasks are to count the number of "G" letters in as many randomly generated strings as possible, by placing a block in one of three possible bins that correspond to three possible numeric answers for each string. In both experiments, in each round there is a monetary prize (uniformly drawn from $\{\$0.1, \$0.2, \ldots \$3.9\}$) that the human wins on the next day with probability p, where p depends on the human's score, e_h, and on the robot's score e_r (the number of strings they counted correctly). In both experiments, increasing the human's score by one increases the probability of winning the prize by one percent. In the subtractive experiment the robot's score negatively affects this probability: p is given by $p(e_h, e_r) = (e_h - e_r + 50)/100$. In the additive experiment, the robot's score affects it positively: $p(e_h, e_r) = (e_h + e_r)/100$.

Our design builds on and extends Gill and Prowse's (2012) experimental paradigm, and it is closely guided by a general-purpose EBRD model (Kőszegi and Rabin 2006, 2007, 2009; for a review of the theory and applications, see O'Donoghue and Sprenger, 2018). Designing our experiments using EBRD theory is meant to induce robot effects on human workers that operate solely through workers' expectations (regarding material payoffs), and not through peer effects. Modeling peer effects often relies on feelings towards others, e.g., inequity aversion, altruism, envy and social pressure. These motives do not easily translate to human-robot contexts without imposing strong assumptions on anthropomorphism (i.e., the attribution of human characteristics, feelings or behavior to objects), and there is evidence that such feelings are diminished when working with robots (Chugunova and Sele, 2022). Robotic-coworkers' effects on expectations, on the other hand, do not depend on how they are perceived as peers.

Notice that our primary interest is the way human-robot interaction is shaped by incentives, and not contrasting the responses of humans to robots vs. other humans. As a result, our empirical design does not compare our human-robot experiments to some

human-human benchmark. Instead, we introduce variation in the compensation scheme and examine its impact on participants' attitudes and behavior towards the robot.

In Section 3 we present the theoretical predictions of the EBRD model in our experimental design, which can be interpreted as manifestations of disappointment aversion. To illustrate, consider the subtractive experiment: a human worker works alongside a robot for a chance to win a monetary prize. A better-performing robot decreases the human's chance of winning the prize: for every level of effort exerted by the human, their expectations of winning are lower the better the robot performs. Hence, holding the human's effort fixed, the magnitude of disappointment that the human faces in case of losing is also smaller the better the robot performs. If the human is disappointment averse, i.e., their risk of getting disappointed looms larger than their potential of being positively surprised, their incentive to exert effort in order to avoid the potential disappointment decreases. In total, a *discouragement* effect of the robot on the human is predicted. If, instead, a better-performing robot increases the human's chance of winning the prize, it is predicted to have the opposite, *encouragement* effect, where better-performing robots increase the human's performance. In the EBRD model, loss aversion (Kahneman and Tversky, 1979) is the underlying cause of disappointment aversion: disappointment is modeled as a "loss" with respect to an expectations-based reference point.

In Section 4 we present our results. Combining both experiments, we have data on 120 participants, 60 in each of the two experiments (and each working alongside the robot in 10 rounds). Starting with chosen effort, in the subtractive experiment, we find a small and precisely estimated negative effect of robot score on human score consistent with a discouragement effect. A one-point increase in robot score reduces human score by 0.043 points on average (SE = 0.014; standardized effect = 0.109). In the additive experiment, we find a positive but imprecisely estimated effect of 0.023 (SE = 0.018; standardized effect = 0.041). Importantly, these two effects are statistically different from each other in the direction predicted by the theory (diff. = 0.066, SE = 0.023). Therefore, changing the compensation scenario from subtractive to additive, while keeping everything else constant, makes the effect of robot performance on human performance significantly more positive. The effect of monetary incentives is small and imprecisely estimated. A one-dollar increase in the value of the monetary prize increases human performance by 0.311 and 0.373 in the subtractive- and additive-compensation experiments, respectively (SE = 0.172 and 0.207; standardized effect = 0.061 and 0.059, respectively). Last, we construct an adjusted measure of robot performance by flipping the robot score's sign in the additive-compensation experiment, and adding a constant, such that this adjusted variable has the same predicted effect in the two experiments. Pooling the data, we first find no statistically significant difference between the adjusted robot's effect across the two experiments, supporting the view that differences in the experiment setting other than the incentive formula do not drive our results, as intended by our design. Second, we find an overall effect of 0.034 (SE = 0.011; standardized effect = 0.072) of the adjusted robot score variable on human score, consistent with the theory, and a small and more precisely estimated prize effect of 0.340 (SE = 0.135; standardized effect = 0.058) in the pooled data.

To interpret the magnitudes of the coefficients, we emphasize two related points. First, the elasticity of effort with respect to monetary incentives is low, as is often the case in experiments like ours, where most of the variation in workers' effort provision comes from the intensive margin.[1] Second, and more importantly, while the absolute magnitude of the robot-performance effect is small in both experiments, its magnitude relative to monetary incentives is sizable (70–179 percent of the standardized prize effect). On the one hand, these findings imply that our directional results should be treated with caution, as their economic magnitude is small. On the other hand, our findings might also suggest that in contexts where performance is more responsive to monetary incentives, the robot's performance effect could also be large.[2]

Moving on to subjective attitudes, we find the additive scenario to be better than the subtractive one in maintaining positive human attitudes towards themselves and towards their robotic coworkers. This may suggest a form of attribution bias (Ross, 1977; Gilbert and Malone, 1995) where participants attribute favorable characteristics to the robot and themselves when it benefits them financially, and unfavorable (or less favorable) traits when it harms them. In both experiments, the robot's score positively affects a subjective rating of robot competence (a standardized effect of 0.48 in each experiment, SE = 0.03). In addition, a higher robot score makes the robot less likable to participants when the compensation scheme is subtractive, and more likable when it is additive (standardized effects of −0.23 and 0.34, respectively, SEs = 0.03). The robot's score also negatively affects a subjective rating of participants' own competence in both experiments. However, this negative effect is much stronger in the subtractive scenario than in the additive one (standardized effects of −0.27 and −0.08, respectively, SEs = 0.02–0.03). These findings, combined with our real-effort results, suggest that the way workers react to the performance of robots strongly depends on how it affects their earnings.

We close Section 4 by analyzing open-ended responses, and find that they are overall consistent with the EBRD mechanism. We find roughly twice as many participants who reported being motivated to work as predicted by theory—i.e., work harder when the robot is slower (faster) in subtractive (additive) compensation—as those who reported the opposite motivation.

We conclude in Section 5.

[1] For example, DellaVigna and Pope (2017) find that an order-of-magnitude increase in the piece rate (from 1 to 10 cents) increases effort by less than 10 percent. We can only speculate why our prize-size effect is lower and more noisily estimated than in Gill and Prowse (2012; see Appendix D.1). Perhaps our interface makes the variation in the probability of winning more salient than the variation in the size of the prize. Alternatively, waiting a day until the resolution of the lottery may have caused participants to allocate less attention to the size of the prize in each round.

[2] Indeed, as we discuss in Section 4, the EBRD model also predicts the encouragement/discouragement effect to increase with the prize effect; see Gill and Prowse (2012), Proposition 3.

1.1. Related literature and contributions

Our paper is most closely related to an interdisciplinary literature studying human-machine interaction, and to the economic literature on EBRD preferences.

Human-machine interaction Our paper adds to a literature on human behavior due to human-machine interaction (see Chugunova and Sele, 2022 for a recent interdisciplinary review). A great deal of this literature has focused on understanding human behavior and perception when collaborating with or using machines, e.g., algorithm aversion (a tendency to ignore algorithm guidance or advice in decision making even at a visible cost of productivity or at other costs; see Burton et al., 2020, April for a review) or automation bias (over-reliance on automation; see Parasuraman and Manzey, 2010 for a review).

Our first major contribution is showing how human-robot interactions can have different results depending on the economic setting (i.e., compensation scheme), thus proposing a new perspective on these interactions. Existing work mostly documents and studies adverse effects of robots on human behavior and proposes methods to mitigate them, which are unrelated to the compensation scheme. For example, recent studies have investigated ways to reduce the extent of algorithm aversion, e.g., by allowing humans to incorporate their prior opinions into the algorithm (Kawaguchi, 2021), or by re-designing the algorithm to learn from commonly observed human deviations and take them into account (Sun et al., 2022). We propose that the economic setup itself could also play a major role in purposely mitigating such adverse effects.

Relative to our previous work (Kshirsagar et al., 2019), which used a fixed economic setting and only varied the prize and robot performance, the contribution of this paper is to introduce variation into the compensation scheme—allowing the comparison of subtractive vs. additive compensation. In other words, while our previous work establishes that robot performance can affect human effort, this paper investigates how these effects can change as a function of the economic setting. Studying this question requires the combination of both experiments.

Expectations-based reference-dependent (EBRD) preferences Our second major contribution is providing and testing a new prediction of the EBRD model, with useful implications to the analysis and design of workplace incentives. EBRD models such as Kőszegi and Rabin's are increasingly used in economics (e.g., Thakral and Tô, 2021; Dreyfuss et al., 2022b) but to date, experimental evidence supporting them is mixed, and has been shown to depend on implementation details (e.g., Marzilli Ericson and Fuster, 2011; Abeler et al., 2011; Gill and Prowse, 2012; Heffetz and List, 2014; Gneezy et al., 2017; Heffetz, 2021).

Our work contributes to this body of research by proposing what we view as a particularly sharp test of the theory. Since the two experiments we analyze are essentially identical except for the sign of the effect of the robot's performance on expected earnings, the comparison between them provides a new clean test of the EBRD model. In particular, a discouragement effect in the subtractive experiment is consistent with other explanations that do not depend on the effect of the robot's performance on expectations about future earnings. For example, it may be consistent with theories of goal-setting and framing (Heath et al., 1999; Pierce et al., 2020), where the robot's projected score serves as a reference, a goal, or a "score to beat" and with a discouragement effect driven by participants "giving up" when the goal is too high. However, the fact that the effect disappears or even reverses under additive compensation is consistent with EBRD, but not with such theories.

In addition, recent work shows how incorporating EBRD preferences in designing mechanisms and policies can be quite useful (see Dreyfuss et al., 2022a for a recent example in the context of designing assignment mechanisms). By combining data from the two experiments, we show that it can also be useful for understanding behavior and designing incentives in hybrid human-robot workplaces.

2. Experimental design

The full text of the experiment instruments is in Online Appendix A. For further details on the experimental design, see Online Appendix B.

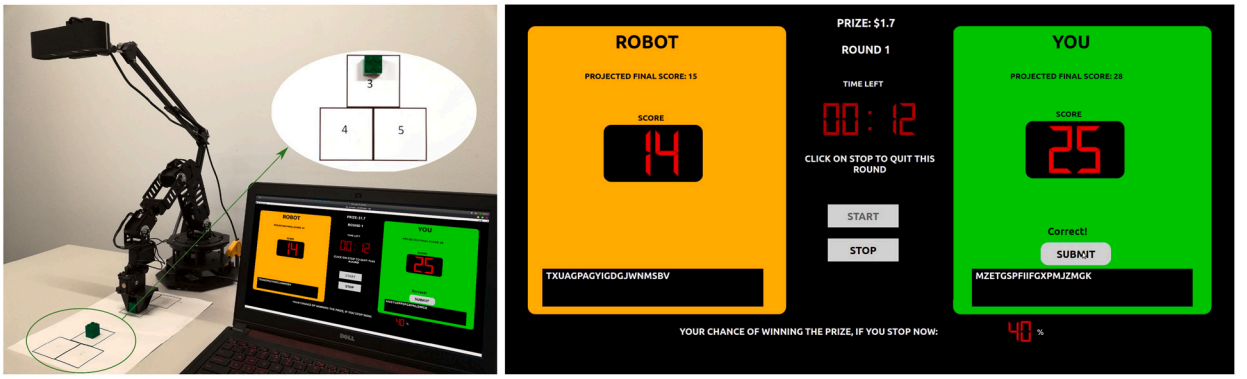
The experimental paradigm (Gill and Prowse, 2012) We start by describing the experimental paradigm in general. Below we review details that are relevant specifically to our implementation.

There is a pair of players: a First Mover and a Second Mover, who complete a real-effort task. As suggested by their names, the First Mover first completes the task and receives some score, denoted e_1. We are primarily interested in the behavior of the Second Mover. The Second Mover observes e_1, and then completes the same task and receives a score, denoted e_2. There is a prize $\$v$. In the basic version of the experiment, the Second Mover wins the prize with probability

$$p(e_1, e_2) = \frac{e_2 - e_1 + 50}{100}. \tag{1}$$

Notice that this is a *subtractive* compensation scheme: the probability linearly increases in e_2, and linearly *decreases* in e_1. Also notice that the marginal expected monetary benefit from an increase in e_2 is independent of e_1. Therefore, if agents care only about absolute amount of money earned and their cost of effort, the Second Mover's work effort should not depend on the effort of the First Mover. However, (as shown in the next section), the EBRD model predicts a *discouragement* effect in this scenario.

The implementation in Gill and Prowse (2012) has two human participants completing the task sequentially: the First Mover chooses a level of effort first, and the Second Mover attempts the task after observing e_1. Subjects play the same role (First/Second



Notes: Our setup consists of a robotic arm, a vision sensor, bins for placing the block in the letter-counting task and a user interface.

Fig. 1. Experiment setup.

Mover) for 10 rounds. To make interactions one-shot, two players can be paired with each other only once. For further details, see Gill and Prowse (2012).

In our setting, the First Mover is a robotic arm programmed to work at some fixed speed that varies randomly between rounds, and the Second Mover is a human participant. As we explain below, the robot and human work simultaneously. However, before the beginning of each round, the human participant learns the robot's projected score (given the round's randomly-chosen speed), i.e., what the robot's final score in that round is projected to be.

Letter-counting task The experiment setup is shown in Fig. 1. In both experiments, each participant worked next to a robotic arm for 10 rounds, for the chance of winning a monetary prize. In each round, the participant and the robot each received a randomly generated string of 20 characters. Their task was to count the number of "G" letters in their texts, place a block in one of three possible bins that correspond to three possible numeric answers, and click on the "Submit" button in the user interface to verify their block placement. Following a correct placement, participants received one point and were shown the next string of characters. For an incorrect placement, they did not receive a point and the "Submit" button in the user interface was disabled for 10 seconds, after which they could submit a new block placement for verification. Each round lasted two minutes, but participants could choose to stop a round at any earlier time.[3] During the round, participants were continuously shown their current score, their projected final round score (based on their average speed so far), the robot's score and projected final round score, their winning probability if they stop now, and the prize that they were working for.

Compensation scheme In each round a participant worked for a chance to win a monetary prize, randomly and uniformly drawn from all the \$0.1 multiples between \$0.1 and \$3.8 in the subtractive-compensation experiment, and between \$0.1 and \$3.9 in the additive-compensation experiment.[4] In both experiments the participants' score positively affected their probability of winning the prize, however the robot's score affected this probability differently in the two experiments. In the subtractive-compensation experiment (Kshirsagar et al., 2019), the robot's score negatively affected the human participant's winning probability. The probability for the human winning the prize was determined using the formula used in Gill and Prowse (2012):

$$p^s(e_h, e_r) = \frac{e_h - e_r + 50}{100}, \tag{2}$$

where p is the probability of winning the current round's prize, and e_h and e_r are the human's score and the robot's score in that round, respectively. In this paper we added a new additive-compensation experiment, where the robot's score *positively* affected the human participant's winning probability. The probability for the human winning the prize was determined using the formula

$$p^a(e_h, e_r) = \frac{e_h + e_r}{100}. \tag{3}$$

In case a participant chose to stop the round early, e_h would be the participant's score at the time of stopping and e_r would be equal to the robot's projected score—its score assuming that it would have continued uninterrupted until the end of the two-minute round, given by $e_{r,\text{final}} = e_{r,\text{now}}/\text{time}_{\text{now}} \times 120$.[5]

The resolution of each round's lottery given participants' final winning probabilities, as well as actual payment of prizes, were conducted in a different lab session on the next day (see details below).

[3] This option was not available to participants in the original implementation of Gill and Prowse (2012).

[4] This small difference is due to a coding error in the subtractive-compensation experiment.

[5] We wanted to allow for behavior to change on the extensive margin by allowing participants to choose not to compete at all in certain rounds. In practice, participants rarely used this option. For example, a round lasts 20 seconds or less in only 2.7 percent of participant-round observations.

Robot performance The robot was programmed to always place its block in the correct bin, earning one point. However, the robot's movement speed was randomly chosen before each round, such that its initially projected final score in the round is distributed uniformly in $\{5, 6, ..., 45\}$. Participants learned the robot's projected score before the beginning of each round and were able to act based on this information.

Due to inaccuracies in the robot's motion planner, there were small differences between the initially projected robot score and the final robot score (concentrated in the lower and higher ends of the distribution). These differences are small enough and shrink quickly enough during the round, that we model each participant's belief distribution as a point mass distribution determined by the round's final robot score.[6] In Section 4 and Online Appendix D.4 we verify that our results are robust to using the initially projected score as the robot's performance measure instead of its final score.

Subjective-attitudes elicitation To investigate the design effects on subjective attitudes, participants filled out a short questionnaire after each round:

1. Robot Competence: "Please rate how much you consider the robot to be competent on the following scale." (1–5)
2. Robot Likability: "Please rate how much you like the robot on the following scale." (1–5)
3. Self Competence: "Please rate how true the following sentence is for you with respect to this task: I feel confident in my ability to do this word-counting task well." (1–7)

These measures (and their different scales) are based on previously validated instruments (Bartneck et al., 2009; Williams and Deci, 1996).

Expectations elicitation To investigate whether the combination of compensation scheme and robot performance in our design was successful at manipulating participants' winning expectations, the additive-compensation experiment included a subsample of randomly chosen half of the participants, who were asked to estimate their final winning probabilities before each round. Before starting to work and just after observing the robot's projected final score, they were asked:

"With what probability of winning the prize do you expect to finish this round?"

Online Appendix E analyzes these elicited expectations, and finds that our robot-score manipulation strongly affects participants' expectations, as intended by our design.[7]

We only elicited expectations from half the participants to verify that the elicitation itself is not driving our main results (the subtractive-compensation experiment did not include expectations elicitation). Reproducing our analysis without the 29 participants whose expectations were elicited essentially reproduces our main results, albeit with larger standard errors.

Procedure As described in Kshirsagar et al. (2019), the subtractive-compensation experiment was conducted between July 25, 2018 and August 26, 2018. The additive-compensation experiment was conducted between April 25, 2019 and May 3, 2019.

We recruited participants from the Cornell Business Simulation Lab online pool of participants.[8] In practice, all participants in the subtractive-compensation experiment were students except for one staff member and one person whose student/staff status is unknown, and we limited recruitment to students only in the additive-compensation experiment. In addition to the monetary prizes won, participants earned a show-up fee of $10 in the subtractive-compensation experiment, and 1 lab credit (valuable to students) in the additive-compensation experiment.[9] During recruitment, participants were informed only about the show-up fee or lab credit, but not about the round prizes.[10]

Each participant signed up online for their own, individual session. Upon arrival in the experiment room, the participant signed a consent form and read printed instructions. The experimenter answered questions, if any, related to the experimental task and

[6] The average difference between projected robot score and final robot score at the beginning of a round was slightly below 1, and the average absolute difference was 3. That difference quickly shrank once the round actually started, and the projected score started getting updated (after each increase in the robot's actual score). By the time the robot's score reached 2, average difference was effectively zero (and remained so until the end of the round), and average absolute difference was around 1.5, and kept decreasing. Due to these inaccuracies, we also ended up with a small fraction of final robot scores outside the $\{5, 6, ..., 45\}$ set: 2.5 percent of rounds in the subtractive-compensation experiment and 1.3 percent in the additive-compensation experiment had robot scores lower than 5, and another 0.5 percent of rounds in the subtractive-compensation experiment had a robot score of 46.

[7] Beyond this paper's main focus, Online Appendix E further investigates whether elicited expectations conform to the rational-expectations benchmark of a one-to-one response to both robot and human score. In a regression of elicited expectations on both robot score and human score, we find that both coefficients are smaller than the rational-expectations benchmark of 1, with the coefficient on robot score (0.64, SE = 0.05) closer to 1 than the coefficient on human score (0.24, SE = 0.09). Whether these results reflect an actual deviation from rational expectations, measurement error, or both, they are in line with the direction in which our exogenous expectations manipulation should theoretically affect participants' effort choice.

[8] All Cornell students and staff are eligible to sign up on this portal.

[9] Professors of some sections of some classes throughout the university (but mostly in the business school, for example, in economics classes offered to MBA students) may choose to offer their students extra credit toward their final grade for participating in studies at the lab. Earned credit is valid only for the semester in which it is earned and generally cannot be carried forward to the next semester.

[10] Motivated by the theory, this was meant to increase the effectiveness of our expectations-shifting manipulation, by preventing formation of prior expectations about the chance of winning round prizes.

compensation scheme. The participant then filled out a comprehension quiz, designed to make sure that they understood the compensation scheme. This was followed by a demonstration of the lottery resolution procedure (see below) and a practice round lasting two minutes to familiarize them with the task. The experimenter then left the room and the participant completed ten paying rounds lasting two minutes each.

At the end of the ten rounds, participants were asked to optionally answer two open-ended questions about their experience working with the robot and about how they decided how hard to work.

Participants had to return to the lab on the next day to resolve the lotteries and get their payment. During this second lab visit, we used a fair 100-sided die on a public website. If the die roll result was less than or equal to the participant's chance of winning the prize, the participant was paid the prize for that round.[11]

The first-day sessions lasted around 50 minutes in the subtractive-compensation experiment and 50–55 minutes in the additive-compensation one (due to expectations elicitation). The second-day sessions lasted around 5 minutes in both experiments.

Sample Excluding the pilot runs as well as 4 subjects whose data were not appropriately logged due to errors, we have data on 60 participants in each experiment. Each participant played 10 paying rounds (Total $N = 1{,}200$).[12] In the subtractive-compensation experiment, one participant did not fill out the subjective-attitudes questionnaire after one of the rounds. We thus have 599 observations for the data analysis involving these questionnaires.

3. Theoretical predictions

In this section we reproduce the main theoretical predictions from Gill and Prowse (2012). We solve for the optimal effort choice of a participant working alongside the robot, as a function of the compensation scheme, robot performance and monetary prize.

Participants choose a level of effort (score), e_h, to maximize their utility. Traditional models of labor supply highlight two terms in the utility function: $-C(e_h)$, an increasing cost function of effort, and pv, the expected payoff, with p the probability of winning and v the size of the prize. Humans maximizing the sum $-C(e_h)+pv$ are predicted to increase their effort e_h the larger v is, independently of the robot's score e_r and independently of the experimental condition—subtractive vs. additive—expressed through the linear dependency of p on e_h and e_r (see eq. (2), (3)).

Kőszegi and Rabin's EBRD model includes two additional terms that represent expected gains and losses relative to expectations, i.e., positive surprises vs. disappointments in our setting. In particular, the choice of e_h determines the probability of winning, which becomes a stochastic reference point. This makes the reference point *expectations-based*: rational expectations are the reference point, and gains and losses are evaluated with respect to this stochastic reference point. With probability p, the human gets the prize v and experiences a gain relative to the potential outcome of not winning the prize; the gain equals $(1-p)v$ (because the probability of not winning is $1-p$, and the prize is v). Thus, $p(1-p)v$ is the expected size of a positive surprise. With probability $1-p$ the human gets nothing and experiences a loss relative to the potential outcome of winning the prize; the loss equals $-\lambda pv$. Thus, the expected size of disappointment is $-(1-p)\lambda pv$. The weighting of a loss relative to a gain by $\lambda > 1$ formalizes loss-aversion (i.e., the notion that the pain from a loss is greater than the elation from an equally sized gain), and in this setting translates to disappointment aversion. The resulting expected utility as a function of e_h and e_r is therefore:

$$\begin{aligned} U(e_h,e_r) &= -C(e_h)+p(e_h,e_r)v+p(e_h,e_r)(1-p(e_h,e_r))v-(1-p(e_h,e_r))\lambda p(e_h,e_r)v \\ &= -C(e_h)+(2-\lambda)p(e_h,e_r)v+(\lambda-1)p(e_h,e_r)^2v. \end{aligned} \tag{4}$$

As the expression highlights, with these additional EBRD terms the optimal effort level now also depends on e_r.[13] To see why, notice that the two EBRD terms (on the right) constitute a quadratic function of p. The quadratic function is U-shaped since losses loom larger than gains ($\lambda > 1$). This means that the marginal benefit of a one-percent increase in the winning probability increases with the baseline chance p, which depends on e_r. In other words, the human has a smaller incentive to increase their winning probability when it is low to begin with. See Gill and Prowse (2012), p. 479, for more discussion of the intuition behind the quadratic term and how it affects behavior.

For a given baseline probability, a marginal increase in the robot's projected score in the additive experiment has the same effect as a marginal decrease in the robot's projected score in the subtractive treatment. Therefore, in the subtractive-compensation experiment (where high-performing robot decreases the chances of winning) participants are predicted to achieve a lower score when the robot performs better. However, in the additive-compensation experiment (where high-performing robot increases the chances of winning) they are predicted to achieve a higher score when the robot performs better. We refer to these effects as a *discouragement* and an *encouragement* effect, respectively (recall that under classical preferences, the choice of effort is independent of e_r). Last, under both compensation structures, participants are predicted to increase their effort with the prize size v.

[11] In Gill and Prowse's (2012) original experiment, winners and losers were determined at the end of each round; however, motivated by the theory, we wanted to extend the period between expectations formation and lottery resolution. In particular, the analysis below solves the problem using Kőszegi and Rabin's (2007) "Choice-acclimating Personal Equilibrium" (CPE). However, as shown in Kőszegi and Rabin (2009), using CPE is only appropriate when the resolution of the lottery occurs at least one period after the lottery choice. We therefore delayed the resolution of the lottery to make sure that this assumption is valid in our context.

[12] For each experiment we recruited the same number of participants (playing the same number of rounds) as Gill and Prowse (2012).

[13] The original versions of the EBRD models (Kőszegi and Rabin, 2006, 2007, 2009) have an additional term, η, that determines the relative weight on gain-loss utility. However, it has since been shown that η is typically not separately identified from λ. We therefore use the standard practice of normalizing $\eta = 1$ and expressing utility in terms of λ alone.

Table 1
Descriptive statistics.

	Subtractive Compensation $N = 600$[a]				Additive Compensation $N = 600$			
	mean	SD	min	max	mean	SD	min	max
Prize ($)	2.0	1.2	0.1	3.8	2.0	1.2	0.1	3.9
Robot Score	24.4	14.9	1.0	46.0	23.2	13.4	4.0	43.0
Human Score	20.4	5.9	0.0	35.0	22.2	7.4	0.0	42.0
Incorrect Attempts	2.3	1.4	0.0	7.0	1.8	1.3	0.0	6.0
Time Worked (Seconds)	115.1	19.7	2.0	120.0	113.8	23.2	4.0	120.0
Robot Competence	3.8	1.2	1.0	5.0	3.7	1.3	1.0	5.0
Robot Likability	3.0	1.1	1.0	5.0	3.6	1.1	1.0	5.0
Self Competence	4.4	1.3	1.0	7.0	5.4	1.3	1.0	7.0

Notes: Descriptive statistics for the two experiments.
[a] Subjective measures were not collected for one subject in one round in the subtractive-compensation experiment, and hence $N = 599$ for the bottom three rows in this experiment.

Finally, note that this analysis (like the analysis in Gill and Prowse, 2012) effectively assumes "narrow-bracketing" (Benartzi and Thaler, 1999; Rabin and Weizsäcker, 2009), i.e., that participants view each round in isolation, and do not aggregate the lotteries across the ten rounds into a single portfolio. To encourage subjects construing the experiment this way, we did not tell subjects in advance the number of paying rounds they will play, but rather only mentioned "several paying rounds." In Online Appendix C we theoretically prove that the encouragement/discouragement effect is still predicted if participants expectations' aggregate all rounds, but its magnitude declines with the number of aggregated rounds.

4. Results

4.1. Descriptive statistics

Table 1 provides summary statistics of the robot's and the human's performance and attitudes in the two experiments.

4.2. Effects on human score

Guided by the model, we test for the existence and direction of two effects: that of the size of the monetary prize (prize effect), and that of robot score (discouragement/encouragement effect), on human score. We first estimate the following linear specification for the two experiments separately using OLS:

$$\text{Human}_{it} = \beta_0 + \beta_1 \times \text{Prize}_{it} + \beta_2 \times \text{Robot}_{it} + c_i + d_t + \varepsilon_{it} \tag{5}$$

where Human_{it}, Prize_{it} and Robot_{it} are participant i's score, the prize they work for, and the robot score in round t, respectively, c_i and d_t are participant and round fixed effects and ε_{it} is an error term. The coefficients of interest are β_1 and β_2, which capture the prize effect and the discouragement/encouragement effect, respectively.

Table 2 summarizes our main results. The results from the subtractive-compensation experiment (Panel A; column 1; first presented in Kshirsagar et al., 2019, Table 3) show a small and imprecisely estimated prize effect ($\beta_1 = 0.311$, SE $= 0.172$), and a small but precisely estimated discouragement effect, in line with the predictions of the theory.[14] A one point increase in robot score reduces human score by 0.043 points (SE $= 0.014$) on average. In standardized terms, a one-standard-deviation increase in robot score reduces human score by 0.109 standard deviations (SE $= 0.035$). The results from the additive-compensation experiment (Panel A.; column 2) show a small and imprecisely estimated prize effect ($\beta_1 = 0.373$, SE $= 0.207$), and an imprecisely estimated robot score coefficient consistent with a small encouragement effect ($\beta_2 = 0.023$, SE $= 0.018$).

To test how the effects of prize and robot score change between the two experiments, we estimate a regression of human score on prize, robot score, and fixed effects, similar to eq. (5), but add interaction terms of each of these variables with a dummy variable that takes 1 for observations from the additive-compensation experiment, and 0 otherwise. These interaction terms estimate the difference between the effects in the two experiments. The estimated differences are reported in Panel A, 3rd column of Table 2 ("Raw Diff."). First, there is no significant difference in the prize effect between the experiments (diff. $= 0.062$, SE $= 0.269$). Second, although we could not reject a non-positive effect of robot score on human score under additive compensation (2nd column), we can reject the equality of the robot score effect between the two experiments (diff. $= 0.066$, SE $= 0.023$, $p = 0.004$). Therefore, reversing the sign of robot score in the probability formula significantly reverses the direction of its effect on human score in the direction predicted by the theory (recall that traditional, reference-independent models predict a zero effect of the robot's score under both compensation schemes).

Next, we note that according to the probability formulas, the incentives of a human working alongside a robot whose score is e_r in the subtractive-compensation experiment are identical to the incentives of a human working alongside a robot whose score is $50 - e_r$

[14] Our p-value threshold statistical significance/precision is of 0.005; see Benjamin et al. (2018).

Table 2
Effect of robot score and prize on human score and subjective measures.

	Subtractive	Additive	Pooled		
			Raw Diff.	Abs. Diff.	Abs. Avg.
A. Human Score					
Prize ($)	0.311	0.373	0.311	0.311	0.340
	(0.172)	(0.207)	(0.191)	(0.191)	(0.135)
Robot Score	−0.043	0.023	−0.043	−0.043	−0.034
	(0.014)	(0.018)	(0.015)	(0.015)	(0.011)
Prize × Additive			0.062	0.062	
			(0.269)	(0.269)	
Robot Score × Additive			0.066	0.021	
			(0.023)	(0.023)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200
B. Robot Competence					
Prize ($)	0.035	−0.030	0.035		
	(0.027)	(0.031)	(0.029)		
Robot Score	0.038	0.047	0.038		
	(0.002)	(0.003)	(0.002)		
Prize × Additive			−0.065		
			(0.041)		
Robot Score × Additive			0.008		
			(0.003)		
N	599	600	1199		
C. Robot Likability					
Prize ($)	−0.023	−0.010	−0.023		
	(0.026)	(0.029)	(0.028)		
Robot Score	−0.016	0.029	−0.016		
	(0.002)	(0.003)	(0.002)		
Prize × Additive			0.013		
			(0.040)		
Robot Score × Additive			0.045		
			(0.003)		
N	599	600	1199		
D. Self Competence					
Prize ($)	−0.037	−0.001	−0.037		
	(0.028)	(0.028)	(0.028)		
Robot Score	−0.023	−0.008	−0.023		
	(0.002)	(0.002)	(0.002)		
Prize × Additive			0.036		
			(0.040)		
Robot Score × Additive			0.015		
			(0.003)		
N	599	600	1199		

Notes: OLS regression results based on eq. (5) and (6). Each panel reports results from five (panel A) or three (panels B–D) separate regressions, where the dependent variable is indicated in the panel's title, and the (condition-based) sample is indicated in the column's title. Standard errors in parenthesis. Pooled 4th and 5th columns: "Robot Score" equals Robot Score in the subtractive-compensation experiment and 50 − Robot Score in the additive-compensation experiment (as indicated by "Robot sign flipped in Additive").

in the additive-compensation experiment. We exploit this symmetry between the experiments to construct an adjusted variable of robot's score, which takes the true e_r value in the subtractive-compensation experiment and the transformed value $50 - e_r$ in the additive-compensation experiment.[15] The theory predicts the same negative effect of this adjusted robot's score variable on human's score in both experiments.

We first estimate a similar regression to the one in Panel A's 3rd column, using the pooled data of both experiments and interaction terms with an additive-compensation-experiment dummy, but replacing robot score with the adjusted robot score variable (indicated by "X" in the row titled "Robot sign flipped in Additive"). Results are reported in Panel A, 4th column in Table 2 ("Abs. Diff."). As implied by the "Abs. Diff." column, there is no statistically significant difference in the absolute value of robot-score effect across the compensation schemes (diff. = 0.021, SE = 0.023). This is consistent with the design's intention to render the robot score's sign in the compensation formula the only important difference between the two experiments.

[15] To see why, notice that $p^s(e_h, e_r) = \frac{e_h + (50 - e_r)}{100} = p^a(e_h, 50 - e_r)$. Furthermore, since e_r is distributed uniformly between 5 and 45, $50 - e_r$ is also distributed uniformly between 5 and 45, so the adjusted variable has the same distribution across the two experiments.

Last, to see the average absolute effect across the two experiments, we estimate a similar specification using the adjusted robot score, but in order not to differentiate the two experiments, we drop the interaction terms. Results are reported in Panel A, 5th column in Table 2 ("Abs. Avg."). The average prize effect across the two experiments becomes more precisely estimated and it becomes clearer that it is a small effect in magnitude (0.058 in standardized terms, with SE = 0.023). The average effect of the adjusted robot score (−0.034, SE = 0.011) is small, negative and precisely estimated, suggesting an overall significant behavioral effect with the sign predicted by the theory.

Comparison to Gill and Prowse's (2012) results, and the interaction between prize and robot score The intentional similarity of our design to Gill and Prowse's (2012) allows us to compare effect magnitudes across the experiments. We show in Online Appendix D.1 that our discouragement effect (Table 2's 2nd column), as well as our two-experiment-average effect (5th column), are similar in magnitude to Gill and Prowse's average discouragement effect (which equals −0.045, SE = 0.026).

In addition, as mentioned above (see footnote 2), the EBRD model predicts a second-order prize effect: the robot-score effect should increase (in absolute terms) with the size of the prize. This prediction can be tested by interacting the prize-size and the robot-score variables, as done by Gill and Prowse (2012). This specification is not included in our main analysis above since, from the point of view of the model, a necessary condition for finding such a second-order prize effect is the presence of a first-order prize effect, which in our setting is small and imprecisely estimated.[16] Online Appendix D.1 uses a specification with the interaction term and finds that its coefficient is indeed statistically insignificantly different from zero (it is 0.003, SE = 0.012 and 0.009, SE = 0.015 in the subtractive- and additive-compensation schemes, respectively, and its average absolute in a pooled analysis similar to the one in Table 2, Panel A, 5th column is −0.003, SE = 0.010). This result is different from Gill and Prowse's, who found an interacted discouragement effect (as well as a statistically significant and positive prize effect).

Quantity vs. quality and intensive margin vs. extensive margin tradeoffs We investigate whether our results depend on quantity-quality and intensive-margin-extensive-margin tradeoffs specific to our task and setting.[17] Human score, our main measure of effort, encapsulates both quantity of effort—the number of attempted submissions—and quality of effort—the rate of correct submissions. It also encapsulates both intensive-margin effort decisions—the choice of quality and speed conditional on working—and extensive-margin effort decisions—the choice of how long to work during a round before terminating it (or a choice not to terminate it).

In Online Appendix D.2 we decompose human score into quantity—the number of attempted submissions—and quality—the percentage of correct submissions—and use them as two alternative outcome variables. We do not find a quality-quantity tradeoff. Instead, results seem to be driven by both quantity and quality that move in (weakly) the same direction. In terms of relative importance, the overall effect on human score is driven mostly by quantity effects (pooled two-experiment-average coefficient on robot score in a regression of the number of attempts = −0.032, SE = 0.011; similar to the effect on total human score), with quality playing a much smaller role.

In Online Appendix D.3 we replace human score with a dummy for completing the full two minutes in a round or with the round duration (until quitting or completing). We find that results remain qualitatively the same, suggesting that encouragement/discouragement are in both the intensive- and extensive-margin decisions.[18]

Expectations dynamics While our setting is designed to induce a single expectations update at the beginning of each round, it is possible that expectations change during a round or "leak" between rounds.

In Online Appendix D.4 we replace the final-robot-score variable with the initially projected robot score shown to participants—which was not perfectly accurate during the robot's first movements in each round (see footnote 6). We find that these two very similar measures of participants' expectations yield very similar results, both qualitatively and quantitatively.

In Online Appendix D.5 we add the previous round's robot score as another independent variable in our main regression. We do not find that any possible cross-round "expectations leakage" changes the original (intra-round) effects of robot score and prize more than trivially.

4.3. Effects on human attitudes

In addition to studying the effects on behavior, predicted by the EBRD model, we also investigate participants' attitudes toward the robot and toward themselves as measured after each round: their liking of the robot, perceived robot competence and perceived self competence.

We expected people to consider a better-performing robot more competent. We also explored a possible effect of robot score on the participant liking of the robot, without a strong prior on the effect's sign. Finally, we hypothesized that the robot's performance

[16] The lack of strong and precisely estimated first-order prize effect also prevents us from structurally estimating the model as in Gill and Prowse (2012), since the identification of the model parameters relies on moments of the data capturing participants' reaction to prize variation (see their Table 4 in Appendix B).

[17] For example, under the hood of an overall discouragement effect, people may hypothetically be encouraged to attempt to solve *more* strings following a better-performing robot (in contrast with the theory's prediction), but also tend to make more mistakes or to quit the round earlier under these conditions.

[18] In the subtractive-compensation experiment, 63 out of 600 rounds lasted less than the full two minutes (out of which, 44 were stopped before 110 seconds). In the additive-earnings experiment, 52 out of 600 rounds lasted less than two minutes (out of which, 47 were stopped before 110 seconds). Participants are somewhat more likely to stop a round in the last five rounds (39/300 in the subtractive experiment and 37/300 in the additive experiment) than in the first five rounds (24/300 and 15/300, respectively).

would negatively affect people's self-competence. We tested for these effects by regressing each of the dependent variables on robot score and prize, while controlling for participant and round fixed effects, using OLS:

$$\text{Attitude}_{it} = \beta_0 + \beta_1 \times \text{Prize}_{it} + \beta_2 \times \text{Robot}_{it} + c_i + d_t + \epsilon_{it}, \tag{6}$$

where Attitude_{it} is participant i's attitude (robot competence, robot likability or self competence), and the rest of the variables are the same as in equation (5). The coefficient of interest is β_2, capturing the effect of the robot's score on attitudes.

Our results are summarized in Panels B–D in Table 2. Robot score positively predicts perceived robot competence in both experiments. In standardized terms, a one-standard-deviation increase in robot score increases perceived robot competence by 0.478 and 0.485 standard deviations (SE = 0.027 and 0.028), respectively, in the subtractive- and additive-compensation experiments. The robot's score negatively predicts its likability in the subtractive-compensation experiment (standardized effect = −0.227, SE = 0.029) whereas it positively predicts it in the additive-compensation experiment (standardized effect = 0.336, SE = 0.031). Robot score negatively predicts human-perceived self-competence in both experiments, however the effect is greater under subtractive compensation (standardized effect = −0.266, SE = 0.026) relative to additive compensation (standardized effect = −0.078, SE = 0.025).[19] Using the same method as above to evaluate the effect raw differences, we find that the experiments significantly differ in their effect on robot likability and human perceived self competence. As mentioned above, the effect on robot liking has opposite signs in the two experiments, and the effect is roughly twice as strong in the additive-compensation experiment. The effect on self-competence is negative in both experiments, but about three times as strong in the subtractive-compensation one. The effect of robot's score on its perceived competence remains roughly the same in both experiments.

Overall, these findings are consistent with participants exhibiting a form of attribution bias by attributing favorable characteristics to the robot and themselves when the robot's performance benefits them financially, and unfavorable (or less favorable) characteristics when it harms them financially.

4.4. Open-ended responses

Online Appendix F investigates participants' open-ended responses at the end of the experiment. Specifically, in one of the questions they are asked to explain their considerations when choosing how hard to work in each round. Two coders classified responses into common themes. In both experiments, more participants (17, 20 participants in subtractive compensation by Coder 1 and 2 respectively; 13, 17 in additive compensation) reported being motivated to work as predicted by theory—i.e., work harder when the robot is slower (faster) in subtractive (additive) compensation—than those who reported the opposite motivation (10, 6 in subtractive compensation; 5, 5 in additive compensation). In both experiments many participants also reported giving their best effort regardless of the robot, and there were more such reports under additive compensation than under subtractive compensation (8, 12 in subtractive compensation; 21, 25 in additive compensation). For more details including quotes from the responses, see the Online Appendix.

5. Conclusion

Our results suggest that when studying the economic implications of increased automation, one should not overlook how the effects of robot coworkers on workers' effort and perceptions may depend on the economic setting.

Our past work shows that in economic settings that induce—by nature or by design—earning scenarios where the performance of robots negatively affects the likelihood of a good outcome for the workers, workers may respond by decreasing the amount of effort they provide. In the present paper we find that this effect dissipates and perhaps reverses when we reverse the effect of robot performance on the likelihood of the good outcome. The return to effort for the worker was held constant across all earning scenarios and robot-performance levels. This fact strongly suggests that traditional economic models of behavior may not tell the whole story, and more psychologically realistic models of preferences, such as the EBRD model we use, may improve the analysis and the design of these new economic contexts.

In particular, depending on the compensation scheme and economic conditions, the highest-performing robot does not necessarily maximize production, and depending on the robots' performance, evaluating workers' production relative to robotic workers may not be the best way to motivate them to work.

The performance of the robot also affected participants' ratings of the robot and, perhaps more interestingly, of themselves. In the subtractive earning scenario, participants liked a high-performing robot less than a low-performing one, even though they considered the former to be more competent. We find an opposite effect, twice as large, on the robot's likability in the additive scenario. Participants' perception of their own competence decreased with the performance of the robot in both earning scenarios, but the coefficient is three times smaller when the compensation scheme is additive. These findings may suggest additional benefits, beyond effects on economic behavior, to using an additive compensation scheme, of the type we use in this experiment, in human-robot workplaces.

[19] This effect is present even after controlling for human score. In the case of subtractive compensation, it changes from −0.023 to −0.021; SE remains 0.002 (new standardized effect = −0.239, SE = 0.025). In the case of additive compensation, it changes from −0.008 to −0.009; SE remains 0.002 (new standardized effect = −0.090, SE = 0.023).

Finally, since our experiments are designed to isolate and investigate a specific kind of human-robot interaction, their implications for actual hybrid human-robot workplace design are still limited. In our experiments the human and robot worked on two separate tasks; the tasks were simplistic; and the EBRD-guided compensation schemes we applied are only two examples among many that firms can choose and that economic contexts can induce. Following our discussion of the literature in Section 1.1, a particularly interesting future direction would be to study how variation in the compensation scheme interacts with productivity-reducing behavior such as algorithm aversion or automation bias. Can purposefully designed economic settings help mitigate such adverse effects of automation? We hope that our empirical results will motivate further studies exploring how the interaction of humans and robots is affected by the economic setup while also involving more realistic tasks in more natural settings.

Declaration of competing interest

I declare that I have no relevant or material financial interests that relate to the research described in this paper.

Bnaya Dreyfuss
Ori Heffetz
Guy Hoffman
Guy Ishai
Alap Kshirsagar

Data availability

Data will be publicly shared once the paper is published.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jebo.2023.11.024.

References

Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. Am. Econ. Rev. 101 (2), 470–492.
Acemoglu, D., Restrepo, P., 2018. Modeling automation. Am. Econ. Assoc. Pap. Proc. 108, 48–53. https://doi.org/10.1257/pandp.20181020.
Bartneck, C., Kulić, D., Croft, E., Zoghbi, S., 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. 1 (1), 71–81.
Benartzi, S., Thaler, R.H., 1999. Risk aversion or myopia? Choices in repeated gambles and retirement investments. Manag. Sci. 45 (3), 364–381.
Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., et al., 2018. Redefine statistical significance. Nat. Hum. Behav. 2 (1), 6–10. https://doi.org/10.1038/s41562-017-0189-z.
Bessen, J., 2020. Automation and jobs: when technology boosts employment. Econ. Policy 34 (100), 589–626. https://doi.org/10.1093/epolic/eiaa001.
Burton, J.W., Stein, M.-K., Jensen, T.B., 2020, April. A systematic review of algorithm aversion in augmented decision making. J. Behav. Decis. Mak. 33 (2), 220–239. https://doi.org/10.1002/bdm.2155.
Chugunova, M., Sele, D., 2022. We and it: an interdisciplinary review of the experimental evidence on how humans interact with machines. J. Behav. Exp. Econ. 99, 101897. https://doi.org/10.1016/j.socec.2022.101897.
DellaVigna, S., Pope, D., 2017. What motivates effort? Evidence and expert forecasts. Rev. Econ. Stud. 85 (2), 1029–1069. https://doi.org/10.1093/restud/rdx033.
Dreyfuss, B., Glicksohn, O., Heffetz, O., Romm, A., 2022a. Deferred acceptance with news utility. Discuss. Pap. - Maurice Falk Inst. Econ. Res. Isr. 2, 1–36. https://www.proquest.com/docview/2695514376/abstract/BA75085E07074AF6PQ/1.
Dreyfuss, B., Heffetz, O., Rabin, M., 2022b. Expectations-based loss aversion may help explain seemingly dominated choices in strategy-proof mechanisms. Am. Econ. J. Microecon. https://doi.org/10.1257/mic.20200259.
Gilbert, D.T., Malone, P.S., 1995. The correspondence bias. Psychol. Bull. 117 (1), 21.
Gill, D., Prowse, V., 2012. A structural analysis of disappointment aversion in a real effort competition. Am. Econ. Rev. 102 (1), 469–503. https://doi.org/10.1257/aer.102.1.469.
Gneezy, U., Goette, L., Sprenger, C., Zimmermann, F., 2017. The limits of expectations-based reference dependence. J. Eur. Econ. Assoc. 15 (4), 861–876. https://doi.org/10.1093/jeea/jvw020.
Heath, C., Larrick, R.P., Wu, G., 1999. Goals as reference points. Cogn. Psychol. 38 (1), 79–109. https://doi.org/10.1006/cogp.1998.0708.
Heffetz, O., 2021. Are reference points merely lagged beliefs over probabilities? J. Econ. Behav. Organ. 181, 252–269. https://doi.org/10.1016/j.jebo.2020.11.010.
Heffetz, O., List, J.A., 2014. Is the endowment effect an expectations effect? J. Eur. Econ. Assoc. 12 (5), 1396–1422. https://doi.org/10.1111/jeea.12084.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291. https://doi.org/10.2307/1914185.
Kawaguchi, K., 2021. When will workers follow an algorithm? A field experiment with a retail business. Manag. Sci. 67 (3), 1670–1695. https://doi.org/10.1287/mnsc.2020.3599.
Kőszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. Q. J. Econ. 121 (4), 1133–1165. https://doi.org/10.1093/qje/121.4.1133.
Kőszegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. Am. Econ. Rev. 97 (4), 1047–1073.
Kőszegi, B., Rabin, M., 2009. Reference-dependent consumption plans. Am. Econ. Rev. 99 (3), 909–936. https://doi.org/10.1257/aer.99.3.909.
Kshirsagar, A., Dreyfuss, B., Ishai, G., Heffetz, O., Hoffman, G., 2019. Monetary-incentive competition between humans and robots: experimental results. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 95–103.
Marzilli Ericson, K.M., Fuster, A., 2011. Expectations as endowments: evidence on reference-dependent preferences from exchange and valuation experiments. Q. J. Econ. 126 (4), 1879–1907. https://doi.org/10.1093/qje/qjr034.
Nordhaus, W.D., 2021. Are we approaching an economic singularity? Information technology and the future of economic growth. Am. Econ. J. Macroecon. 13 (1), 299–332. https://doi.org/10.1257/mac.20170105.
O'Donoghue, T., Sprenger, C., 2018. Reference-dependent preferences. In: Handbook of Behavioral Economics: Applications and Foundations, vol. 1. Elsevier, pp. 1–77. https://linkinghub.elsevier.com/retrieve/pii/S2352239918300034.

Parasuraman, R., Manzey, D.H., 2010. Complacency and bias in human use of automation: an attentional integration. Hum. Factors 52 (3), 381–410. https://doi.org/10.1177/0018720810376055.

Pierce, L., Rees-Jones, A., Blank, C., 2020. The negative consequences of loss-framed performance incentives (Tech. Rep. No. NBER Working Paper No. 26619), National Bureau of Economic Research.

Rabin, M., Weizsäcker, G., 2009. Narrow bracketing and dominated choices. Am. Econ. Rev. 99 (4), 1508–1543.

Ross, L., 1977. The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process. Advances in Experimental Social Psychology, vol. 10. Elsevier, pp. 173–220.

Sun, J., Zhang, D.J., Hu, H., Van Mieghem, J.A., 2022. Predicting human discretion to adjust algorithmic prescription: a large-scale field experiment in warehouse operations. Manag. Sci. 68 (2), 846–865. https://doi.org/10.1287/mnsc.2021.3990.

Thakral, N., Tô, L.T., 2021. Daily labor supply and adaptive reference points. Am. Econ. Rev. 111 (8), 2417–2443. https://doi.org/10.1257/aer.20170768.

Williams, G.C., Deci, E.L., 1996. Internalization of biopsychosocial values by medical students: a test of self-determination theory. J. Pers. Soc. Psychol. 70 (4), 767.