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]

November 21, 2023

Abstract

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.

*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.
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).
[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

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 performance 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 sizeable (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 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, Heffetz, et al., 2022) but to date, experimental evidence supporting them is mixed, and have 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, Glicksohn, et al., 2022 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 & 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 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).

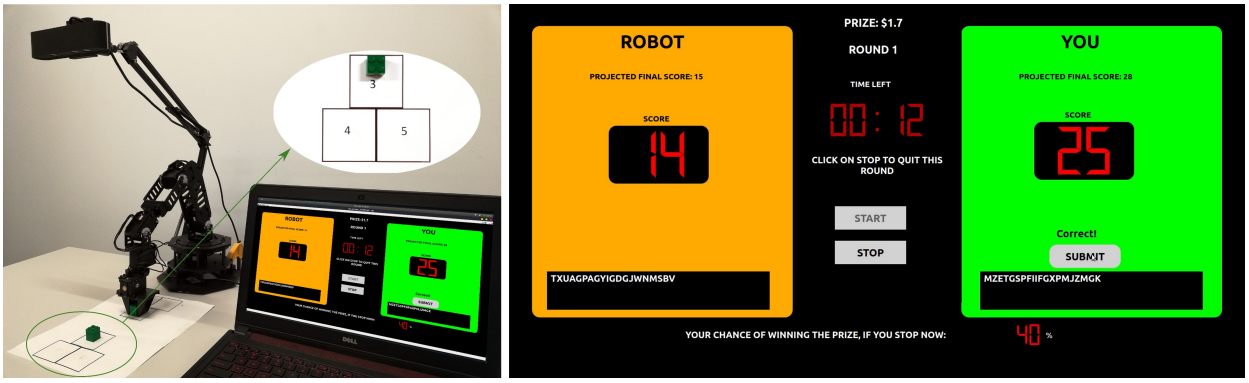
In our setting, the First Mover is a robotic arm programmed to work at some fixed speed that varies randomly between rounds, and the Second Mover is a human participant. As we explain below, the robot and human work simultaneously. However, before the beginning of each round, the human participant learns 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 Figure 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

[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.

Figure 1: Experiment setup



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.

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).

[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 & 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

[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.

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 compensation

[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.

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

[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).

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:

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

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

[13]The original versions of the EBRD models (Kőszegi & 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.

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.

Finally, note that this analysis (like the analysis in Gill and Prowse, 2012) effectively assumes "narrow-bracketing" (Benartzi & Thaler, 1999; Rabin & 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 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 Likeability	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.

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 $= 0.023$, $p = 0.004$).

[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 Likeability					
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").

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 $50 - e_r$ 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.

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

[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.

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),

[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.

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 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:

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

where Attitude_{it} is participant i's attitude (robot competence, robot likability or self compe-

[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).

tence), 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,

[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).

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.

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.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, *101*(2), 470–492.

Acemoglu, D., & Restrepo, P. (2018). Modeling automation. *AEA Papers and Proceedings*, *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. *International Journal of Social Robotics*, *1*(1), 71–81.

Benartzi, S., & Thaler, R. H. (1999). Risk aversion or myopia? choices in repeated gambles and retirement investments. *Management Science*, *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., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bessen, J. (2020). Automation and jobs: when technology boosts employment*. *Economic Policy*, *34*(100), 589–626. https://doi.org/10.1093/epolic/eiaa001

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *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. *Journal of Behavioral and Experimental Economics*, *99*, 101897. https://doi.org/10.1016/j.socec.2022.101897

DellaVigna, S., & Pope, D. (2017). What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies*, *85*(2), 1029–1069. https://doi.org/10.1093/restud/rdx033

Dreyfuss, B., Glicksohn, O., Heffetz, O., & Romm, A. (2022). Deferred acceptance with news utility. *Maurice Falk Institute for Economic Research in Israel. Discussion paper series*, (2), 1–36. https://www.proquest.com/docview/2695514376/abstract/BA75085E07074AF6PQ/1

Dreyfuss, B., Heffetz, O., & Rabin, M. (2022). Expectations-based loss aversion may help explain seemingly dominated choices in strategy-proof mechanisms. *American Economic Journal: Microeconomics*. https://doi.org/10.1257/mic.20200259

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*(1), 21.

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *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. *Journal of the European Economic Association*, *15*(4), 861–876. https://doi.org/10.1093/jeea/jvw020

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, *38*(1), 79–109. https://doi.org/https://doi.org/10.1006/cogp.1998.0708

Heffetz, O. (2021). Are reference points merely lagged beliefs over probabilities? *Journal of Economic Behavior & Organization*, *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? *Journal of the European Economic Association*, *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. *Management Science*, *67*(3), 1670–1695. https://doi.org/10.1287/mnsc.2020.3599

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, *121*(4), 1133–1165. https://doi.org/10.1093/qje/121.4.1133

Kőszegi, B., & Rabin, M. (2007). Reference-dependent risk attitudes. *The American Economic Review*, *97*(4), 1047–1073.

Kőszegi, B., & Rabin, M. (2009). Reference-dependent consumption plans. *American Economic Review*, *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. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 95–103. https://doi.org/10.1109/HRI.2019.8673201

Marzilli Ericson, K. M., & Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *The Quarterly Journal of Economics*, *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. *American Economic Journal: Macroeconomics*, *13*(1), 299–332. https://doi.org/10.1257/mac.20170105

O'Donoghue, T., & Sprenger, C. (2018). Reference-dependent preferences. *Handbook of behavioral economics: Applications and foundations 1* (pp. 1–77). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/S2352239918300034

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human 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. NBER Working Paper No. 26619). National Bureau of Economic Research. https://doi.org/10.3386/w26619

Rabin, M., & Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *American Economic Review*, *99*(4), 1508–1543.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in experimental social psychology* (pp. 173–220). Elsevier.

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. *Management Science*, *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. *American Economic Review*, *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. *Journal of Personality and Social Psychology*, *70*(4), 767.

Online Appendix for

Additive vs. Subtractive Earning in Shared Human-Robot Work Environments

November 21, 2023

A Experiment Instruments

Note: Black text is common to both the experiments. **Red text corresponds to the competition experiment**, while **blue text corresponds to the collaboration experiment.** (Those texts originally appeared black and not bold in both experiments.)

A1: Study Information (Used for recruiting participants)

Study Name: Decision Making with a Robot

Duration: **50** **55** minutes

Pay: **$10** **1 credit**

Abstract: We are seeking participants for an experimental study on decision making.

Description: Participants will make decisions in the presence of a robotic arm. The task will be quite strenuous, so if you choose to participate, please select a sign-up slot in which you are well-rested and alert. The interactions will be logged and recorded. The study will last approximately **50** **55** minutes. Participants will receive **$10** **1 credit** for their time. **The payment will be made on the day following the study. To earn the credit, participants will also have to visit the lab on the next day for approximately 5 minutes.**

A2: Consent Form

We are asking you to participate in a research study on people's decision making. We will describe this study to you and answer any of your questions.

This study is being conducted by Alap Kshirsagar, Sibley School of Mechanical and Aerospace Engineering, Cornell University. The Faculty Advisor for this study is Guy Hoffman. Ori Heffetz, another faculty member, is also a member of the research team.

The purpose of this research is to investigate how people make decisions. We will ask you to **compete collaborate** with a robotic arm by performing a task that involves counting letters and arranging blocks. The experiment will last approximately **50 55** minutes.

We do not anticipate any risks from participating in this research. You can choose to quit at any time.

Compensation for participation

You will receive **$10 1 credit** for your time. In addition to this, based on your performance you have a chance to win more money, as explained in the experiment.

Audio/Video Recording

During the experiment, we will log your task-progress. The data will not have any personally identifiable information.

Privacy/Confidentiality/Data Security

You will be assigned subject numbers for purposes of recording and analysis of data. All information which could link the subject number to a participant will be in a locked file under the Investigator's control only. No identifiers will be published, subjects will be referred only using the subject number in subsequent publications and presentations.

Data Sharing

De-identified data from this study may be shared with the research community at large to advance science and health. We will remove or code any personal information that could identify you before files are shared with other researchers to ensure that, by current scientific standards and known methods, no one will be

able to identify you from the information we share. Despite these measures, we cannot guarantee the anonymity of your personal data.

Taking part is voluntary

Your involvement is voluntary, you may refuse to participate before the study begins, discontinue at any time, or skip any questions/procedures that may make you feel uncomfortable, with no penalty to you, and no effect on the compensation earned before withdrawing, or your academic standing, record, or relationship with the university or other organization or service that may be involved with the research.

Follow up studies

May we contact you again to request your participation in a follow up study? (Yes/No)

If you have questions

Please ask any questions you have now. If you have questions later, you may contact Alap Kshirsagar at ak2458@cornell.edu. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) for Human Participants at 607-255-5138 or access their website at http://www.irb.cornell.edu. You may also report your concerns or complaints anonymously through Ethicspoint online at www.hotline.cornell.edu or by calling toll free at 1-866-293-3077. Ethicspoint is an independent organization that serves as a liaison between the University and the person bringing the complaint so that anonymity can be ensured.

Statement of Consent:

I have read the above information, and have received answers to any questions I asked. I consent to take part in the study.

Your Signature_____ Date_____

Your Name (printed)_____

Signature of person obtaining consent_____ Date_____

Printed name of person obtaining consent_____

This consent form will be kept by the researcher for five years beyond the end of the study.

A3: Instructions

Thank you for participating in this session. Please read these instructions carefully.

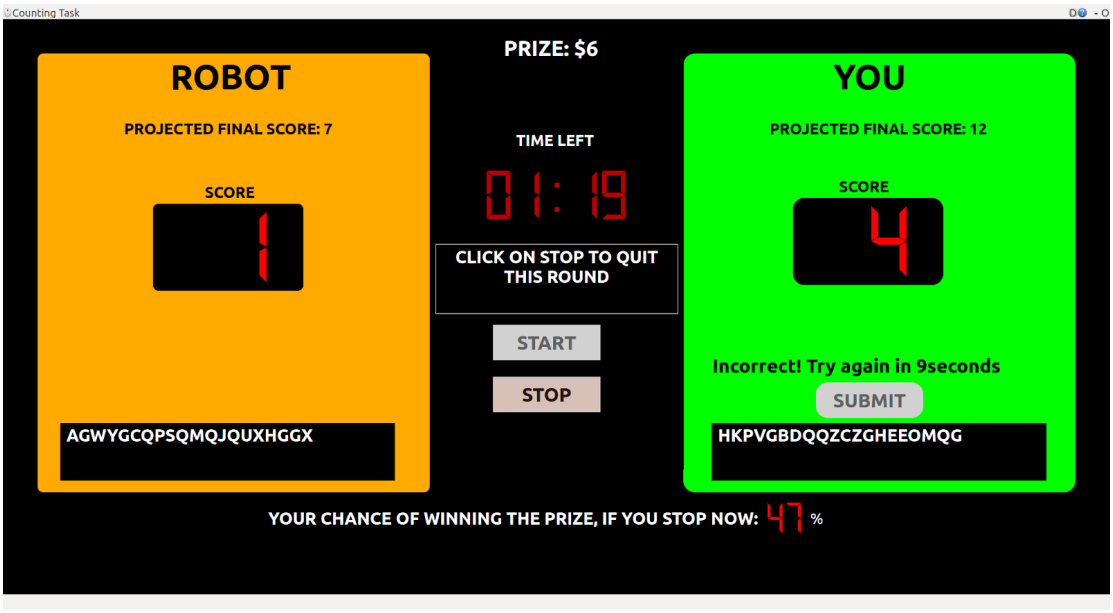
Please turn off your mobile phones or keep them on silent mode (not vibrate).

In this session, you will **compete** **collaborate**, for real monetary prizes, with the robotic arm in front of you. You will **compete** **collaborate** in several rounds. Each round will result in a chance of winning a monetary prize. This chance will determine a lottery that will happen tomorrow, when you return to the lab. Each round has its own prize and its own separate lottery draw.

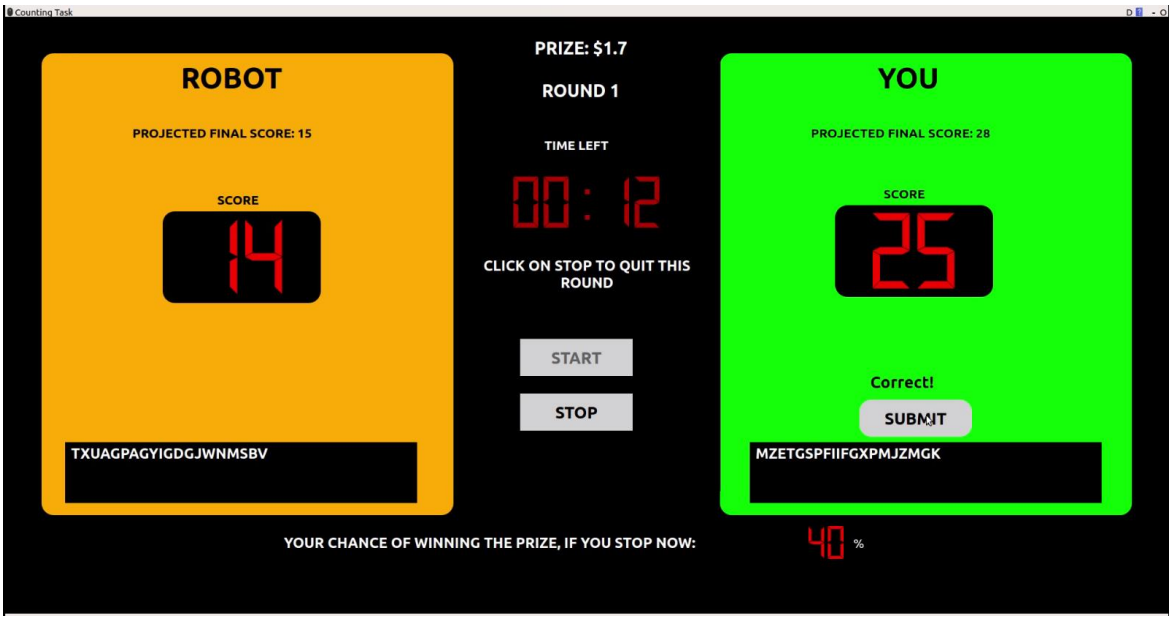
Everything written in these instructions is true. For example, when we mention a monetary prize, the prize is real. When you return to the lab for payment tomorrow, if you win the lottery related to a certain prize, you will actually receive that prize, in cash.

If you have any questions while reading these instructions, please make sure to ask the experimenter.

On the screen you will see information related to the task you and the robot are doing. Below is a screenshot of what you see on the screen. We will explain soon what you see here.

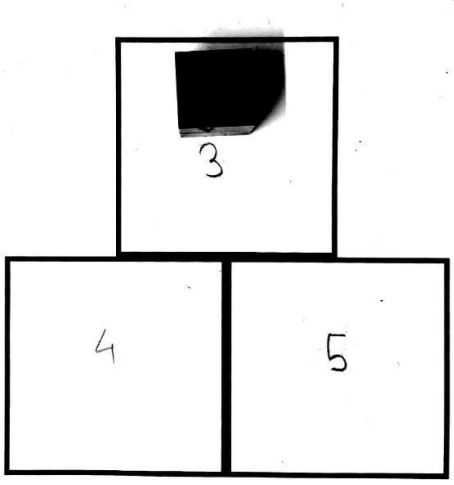


(Competition Experiment)



(Collaboration Experiment)

Here is an image of what the bins used for the task look like (we will explain this soon too):



You and the robot will work on identical tasks as described below:

1. You and the robot each receive a randomly generated text. Your text is displayed on the bottom of the "YOU" side of the screen in a black box, and the robot's text is displayed at the bottom of the "ROBOT" side of the screen. Please locate the text now.

2. You (and the robot) have to count the number of G letters in your texts. There will be either 3,4, or 5 G letters in the text. In the example above, your text contains the letter G three (3) times, and the robot's text contains four (4) occurrences.
3. You (and the robot) have to place the block in the corresponding bin. In the example above, you have to place the block in the bin labeled "3". (The robot will place the block in the bin labeled "4")
4. Once you've arranged your block, click the 'Submit' button on the screen above the text, to validate your block arrangement. The mounted camera will then analyze your block arrangement.

- If your block is placed in the **correct location**, your 'Score' will increase by 1 point, and your next text will appear on the screen.
- If your block is placed in the **incorrect location**, you will get no points, and the 'Submit' button will become disabled for 10 seconds. After 10 seconds, the 'Submit' button will be enabled again, and you can try again. (Of course, you can recount the letters and rearrange the block while the 'Submit' button is disabled; you just cannot click 'Submit' until the button is enabled again.)

The session will have a practice round, followed by several paying rounds. Each round will last two (2) minutes. Different rounds are for different amounts of money. The robot will follow a pre-programed algorithm in each round. The robot may have different speeds in different rounds, but its speed is not affected in any way by your performance and speed. The 'Projected Final Score' of the Robot on a new round's screen reflects the robot's speed in that round.

You do not have to **compete** **collaborate** for the full two minutes and can choose to move to the next round at any time during these two minutes. In other words, you can analyze as many texts as you want to up to two minutes, or click the 'Stop' button if you do not want to continue. This will start the next round.

Points to Note: You and the robot are doing the same task, but texts will be different. If you click 'Stop' before the two minutes are up, the assumption is that the robot would have continued until the end of the two-minute round, with the same average performance it had until the moment when you clicked 'Stop', and would therefore have achieved the robot's projected final score.
Please let the experimenter know at this point if there any questions about the task description.

We now explain what you see on the screen above the texts.

[On each side - bottom to top]

1. SCORE (ROBOT / YOU) : Points accumulated so far
2. PROJECTED FINAL SCORE (ROBOT / YOU) : Total points expected at the end of the 2-minute round. This value is calculated based on the average speed so far: the computer calculates how many points per second you've made so far, and multiplies this number by the total time of round. Of course, this number assumes a constant average speed for you, and may change if you speed up or slow down. The calculation is as follows:

PROJECTED FINAL SCORE = SCORE/TIME ELAPSED (in seconds)*120

[In the center - top to bottom]

3. PRIZE: The monetary prize for this round. Note that this prize will be paid out based on a lottery happening tomorrow, when you return to the lab. The lottery is related to the points you collect, as described in the next section.
4. TIME LEFT: Time left in the round
5. START: Click here to begin the round
6. STOP: Click here to end the round
7. YOUR CHANCE OF WINNING THE PRIZE, IF YOU STOP NOW: If you choose to stop the round, the robot's final score will be made equal to its projected final score and your final score will be made equal to your score (not your projected final score), at that instant. The calculation of this number is as follows:

Your chance of winning, if you stop now (%) = 50 + (Your Score - Robot's Projected Final Score)

Your chance of winning, if you stop now (%) = (Your Score + Robot's Projected Final Score)

Every 5 seconds, you will hear this number **and the prize** in a robotic voice.

Prize Scheme:

Your chance of winning the prize for each round depends on the **difference sum** of the robot's final score and your final score. **If the scores are the same, you and the robot each have 50% chance of winning the prize. If the scores are not the same, the chance of winning for whoever has the higher points score increases by 1 percentage point for every increase of 1 in the difference between the scores, while the chance of winning for whoever has the lower score correspondingly decreases by 1 percentage point. That means that your chance of winning the price increases by 1% for every increase of 1 point in your final score, and decreases by 1% for every increase of 1 point in the robot's final score. For example, if the sum of your and the robot's scores is 50, you have 50% chance of**

winning the prize. Your chance of winning increases by 1 percentage point for every increase of 1 in either of the robot's score or your score.

Your chance of winning (%) = 50 + (Your Final Score - Robot's Final Score)

Your chance of winning (%) = (Your Final Score + Robot's Final Score)

For example:

Your Final Score	Robot's Final Score	Your Final Score minus Robot's Final Score	Your Chance of Winning the Prize	Robot's Chance of Winning the Prize
10	10	0	50%	50%
5	2	3	53%	47%
25	10	15	65%	35%
10	25	-15	35%	65%
0	40	-40	10%	90%

Your Final Score	Robot's Final Score	Your Final Score plus Robot's Final Score	Your Chance of Winning the Prize
10	10	20	20%
5	45	50	50%
25	10	35	35%
30	30	60	60%
0	40	40	40%

Points to note: Your chance of winning the prize depends on the **difference between sum of** the robot's final score and your final score. Assuming the robot's final score will equal its projected final score, your chance of winning the prize increases by 1% for every additional point that you score. (Remember that if you choose to quit a round early, then the robot's final score will be made equal to its projected final score for that round.)

The lottery for each paying round will happen tomorrow when you visit the lab again. We will use a public website (http://www.roll-dice-online.com/) for this purpose. For each paying round, we will roll a 100-sided die on that website, meaning the website randomly chooses a number between 1 and 100. If the result of the die roll is less than or equal to your chance of winning the prize, then you will win the prize for that round. For example, if your chance of winning the prize is 80%, then you will win the prize for any number between 1 and 80, and not win the prize for any number between 81 and 100. Please be assured that you will be paid in a fair way.

Here is a screenshot of the website we will use tomorrow:

Please let the experimenter know at this point if there are any questions about the prize scheme.

Comprehension Quiz

To make sure you understand the prize scheme correctly, please fill out the following brief comprehension quiz. [see **A4**]

Please let the experimenter know when you are done filling out the quiz.

Lottery Resolution Demonstration

To make sure you are familiar with the lottery-resolution procedure that will be followed tomorrow when you return to the lab, please complete the following demonstrations. [see **A5**]

Please let the experimenter know when you are finished with the demonstrations.

Questionnaires

[See **A6**]
After each round, you will be asked to fill out a short questionnaire.

[TO A RANDOMLY SELECTED HALF OF THE SAMPLE] Before the start of each round, you will be asked to fill out another short questionnaire.

Also, at the end of all paying rounds, you will be asked to fill out a questionnaire.

When you visit the lab again tomorrow, you will be asked to fill out another questionnaire.

Practice Round

To familiarize yourself with the task and reward scheme, you will now participate in a practice round. The task description for the practice round is same as that for the actual **competition task.** The practice round will last for 2 minutes. There is no monetary prize for the practice round.

Are there any final questions?

A4: Comprehension Quiz

Please answer these questions based on the information provided to you. These questions are designed to test whether you understand the reward scheme for this **competition task**.

Your Score	Robot's Projected Final Score	**Difference between Sum of** Your Score and Robot's Projected Final Score	Your Chance of Winning the Prize, if You Stop Now.
5	5		
5	25		
5	45		
25	5		
25	25		
25	45		
45	5		
45	25		
45	45		

Q1: If the Robot's 'Projected Final Score' is 10 and your 'Score' increases from 20 to 21, what is the increase in your chance of winning the prize?

Q2: If the Robot's 'Projected Final Score' is 20 and your 'Score' increases from 20 to 21, what is the increase in your chance of winning the prize?

Q3: Is it true that your chance of winning the prize increases by 1% for every 1 point increase in your 'Score', for any given robot's 'Projected Final Score'?

A5: Die-Rolling Demonstration

Please go to the website http://www.roll-dice-online.com/ for a demonstration of the die-rolling procedure that will be followed when you return to the lab tomorrow.

Set the parameters as follows:

- Number of sides = 100
- Number of dice to roll = 1
- Number of rolls = 1

Hypothetical Round 1: Prize = $1, Your score = 15, Robot's score = 10

- Imagine that in one of today's rounds, the prize were $1, your score were 15, and the robot's score were 10.
- In that case, your chance of winning that round's prize tomorrow would be ___%.
- Imagine that you came back to the lab tomorrow to roll the die for that round. Please click 'Roll dice.' What number came up? ___
- With that die-roll outcome, would you win the $1, **or would the robot win**? **I \ ROBOT**. **Yes\No**
- If the answer is **"I" "Yes"** please write **"I" "Y"** in the leftmost square in the table below. If the answer is **"ROBOT," "No"** please write **"R" "N"** in the square.
- Now repeat this process 9 more times (overall 10 die rolls) to get a feel for the uncertainty of winning the prize. After each roll, determine in your head **who if you** would win the prize, and fill the leftmost empty square with **an "I" or "R" a "Y" or "N"** accordingly.

Hypothetical Round 2: Prize = $1, Your score = 9, Robot's score = 38

- In that case, your chance of winning that round's prize tomorrow would be ___%.
- Now repeat the process above (with another 10 die rolls) and fill the table below with **"I"s and "R"s "Y"'s or "N"'s** accordingly.

Hypothetical Round 3: Prize = $1, Your score = 26, Robot's score = 5

- In that case, your chance of winning that round's prize tomorrow would be ___%.
- Now repeat the process above (with another 10 die rolls) and fill the table below with **"I"s and "R"s "Y""s or "N""s** accordingly.

A6: Questionnaires

[TO A RANDOMLY SELECTED HALF OF THE SAMPLE] (To be filled before the start of each round)

Round Number ____

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

(To be filled **once before the start of the study** and then at the end of each round)

Round Number ____

Please rate how much you like the robot on the following scale (circle the number):

Dislike 1 2 3 4 5 Like

Please rate how much you consider the robot to be competent on the following scale:

Incompetent 1 2 3 4 5 Competent

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	2	3	4	5	6	7
Not at all true			Somewhat true			Very true

(To be filled **at the end of the study when they visit again the next day)**

PLEASE MARK THE APPROPRIATE COLUMN WITH AN "X"

*

	Not At All True	Hardly True	Moderately True	Exactly True
1. I can always manage to solve difficult problems if I try hard enough				
2. If someone opposes me, I can find the means and ways to get what I want.				
3. It is easy for me to stick to my aims and accomplish my goals.				
4. I am confident that I could deal efficiently with unexpected events.				
5. Thanks to my resourcefulness, I know how to handle unforeseen situations.				
6. I can solve most problems if I invest the necessary effort.				
7. I can remain calm when facing difficulties because I can rely on my coping abilities.				
8. When I am confronted with a problem, I can usually find several solutions.				
9. If I am in trouble, I can usually think of a solution				
10. I can usually handle whatever comes my way				

(To be filled **at the end of the study** **when they visit again the next day**)

PLEASE MARK THE APPROPRIATE COLUMN WITH AN "X"

	Very untrue of me	Untrue of me	Somewhat untrue of me	Neutral	Somewhat true of me	True of me	Very true of me
	1	2	3	4	5	6	7
1. I put money ahead of pleasure							
2. I firmly believe that money can solve all of my problems							
3. I would do practically anything legal for money if it were enough							
4. I believe that a person's salary is very revealing in assessing their intelligence							
5. I worry about my finances much of the time							
6. I enjoy working in situations involving competition with others							
7. I feel that winning is important in both work and games							

8. It is important to me to perform better than others on a task							
9. It annoys me when other people perform better than I do							
10. I try harder when I'm in competition with other people							

(To be filled at the end of the study)

Please write a few sentences about your experience of **competing** **collaborating** with this robot.

Finally, if you have any comments or thoughts you would like to share with us, please write them here. We are especially curious to know: how did you decide in each round how strongly to **compete** **work**?

B Experimental Design: More Details

Letter counting task. In the subtractive-compensation experiment, the strings given to participants and to the robot were randomly generated using the 26 letters of the English alphabet. In the additive-compensation experiment, the alphabets similar to the letter "G", i.e., "C", "Q" and "O", were not drawn to reduce human error. In case of an incorrect placement, during the 10 seconds in which the submission button was disabled, participants could recount the letters and move their block in the workspace, but could not resubmit. After 10 seconds, they could submit a new answer.

Experiment setup technical details. The experiment setup consisted of an autonomously operating WidowX Mark II robot arm, an Orbbec Astra vision sensor and a laptop as shown in Figure 1. The robot arm was controlled via the MoveIt! motion planner. The vision sensor feedback was used to verify the human's block position and to pick-up the robot's block from its starting position. A screen displayed the accumulated points ("Score"), projected points at the end of the two-minute round ("Projected Final Score"), the cash prize for the round, the time left in the round and the participant's chance of winning the prize if they stopped the round ("Your Chance of Winning the Prize, if You Stop Now"). The user interface also had "Start", "Stop" and "Submit" buttons to start the round, stop the round and submit the block arrangement for verification, respectively. Participants interacted with this screen using a USB mouse. In the subtractive-compensation experiment, a robotic voice read out the participant's probability of winning the prize every five seconds. In the additive-compensation experiment, the prize was also read out every five seconds along with the probability of winning. We made this modification to give equal importance to the prize and the probability of winning (however, as our results indicate, this did not dramatically change the effect of prize size). We used the Robot Operating System (ROS) framework to develop the entire software including the robot's motion controller, computer vision and user interface.

Personality questionnaires. Participants answered a standard personality questionnaire at the end of the first-day session in the subtractive-compensation experiment and at the second-day session in the additive-compensation experiment. We do not use these data in our study.

C Theoretical Predictions: Robustness to Relaxing Narrow Bracketing

Our main theoretical analysis assumes that participants choose each round's effort while narrow bracketing (e.g., Benartzi and Thaler 1999; Rabin and Weizsäcker 2009), i.e., acting as if this round's lottery is the only lottery that will take place at the end of the experiment. This appendix shows that the prediction of a discouragement/encouragement effect is robust to replacing this assumption with an assumption that when choosing a round's effort, participants consider the aggregate lottery resulting from all rounds' outcomes (i.e., no narrow bracketing). While the magnitude of the discouragement/encouragement effect is predicted to decrease with the number of rounds considered, the predicted direction of the effect remains the same.

Using the KR model without narrow bracketing, the EBRD utility, i.e., the expectations-based surprises and disappointments (gains/losses) that enter a participant's utility function (as in Section 3), does not stem only from the expected resolution of the current round's (two-outcome) lottery. Instead, it stems from the aggregate resolution of all rounds' lotteries. However, we assume that the probability of winning this round's lottery is the only one affected by the current decision, while other rounds' probabilities are fixed (due to both rounds that already finished and expectations regarding future rounds).

It is therefore useful to model the aggregate lottery as $L+G$, where L is this round's lottery and G is the lottery stemming from on all past and future rounds. The following claim establishes that an EBRD utility given the lottery $L+G$ is still convex in the probability of winning L, similarly to the case of L alone, discussed in Section 3.

Claim. Let $G=[z_1,Q_1;...;z_n,Q_n]$ be a lottery with a finite support and let $L=[v,p;0,1-p]$ be the lottery of a current round. Both L and G are resolved together in the same future period (when participants return to the lab to receive payment) and L is determined in the current period by setting p. Then the EBRD (gain-loss) utility satisfies

$$\frac{d^2E\left(U_{\text{EBRD}}\right)}{dp^2}\geq 0.$$

Specifically, the marginal EBRD utility with respect to p is

$$\frac{dE\left(U_{\text{EBRD}}\right)}{dp}=\left(\lambda-1\right)\left(2p-1\right)\left[\sum_{l=1}^{n}Q_l^2v+2\sum_{l=1}^{n}\sum_{m=l+1}^{n}Q_lQ_m\max\left\{v-\left|z_l-z_m\right|,0\right\}\right],$$

where $\tilde{v}\in\{0,v\}$, $z\in\{z_1,...,z_n\}$ are possible resolutions of L, G respectively.

Proof. Denote the outcomes and probabilities of $L+G$ as

$$\begin{array}{cc} s_1 = z_1 + v & q_1 = pQ_1 \\ \vdots & \vdots \\ s_n = z_n + v & q_n = pQ_n \\ s_{n+1} = z_1 & q_{n+1} = (1-p)\,Q_1 \\ \vdots & \vdots \\ s_{2n} = z_n & q_{2n} = (1-p)\,Q_n \end{array}.$$

The expected EBRD utility, similarly to Section 3, includes a pairwise comparison of each possible realization to all other possible realizations. Combined together, we get:

$$U \equiv U_{\text{EBRD}} = -(\lambda-1)\sum_{i=1}^{2n}\sum_{j=i+1}^{2n} q_i q_j \left|s_i - s_j\right|.$$

Calculating the derivative with respect to p:

$$\begin{aligned} \frac{dU}{dp} =& \sum_{l=1}^{2n} \frac{\partial U}{\partial q_l}\frac{dq_l}{dp} \\ =& -(\lambda-1)\sum_{l=1}^{2n}\sum_{m=1}^{2n}\left(q_m + \frac{dq_m}{dq_l}q_l\right)\left|s_l - s_m\right|\frac{dq_l}{dp}. \end{aligned}$$

We use $2n$ independent q coordinates, therefore $\frac{dq_m}{dq_l} = 1$ only if $l = m$. For these indices $s_l = s_m$, so their contribution to the sum is zero. Therefore we can write

$$\begin{aligned} \frac{dU}{dp} =& -(\lambda-1)\sum_{l=1}^{2n}\sum_{m=1}^{2n}\frac{dq_l}{dp}q_m\left|s_l - s_m\right| \\ =& -(\lambda-1)\left[\sum_{l=1}^{n}\sum_{m=1}^{n}\frac{dq_l}{dp}q_m\left|s_l - s_m\right| + \sum_{l=1}^{n}\sum_{m=n+1}^{2n}\frac{dq_l}{dp}q_m\left|s_l - s_m\right|\right] \\ &-(\lambda-1)\left[\sum_{l=n+1}^{2n}\sum_{m=1}^{n}\frac{dq_l}{dp}q_m\left|s_l - s_m\right| + \sum_{l=n+1}^{2n}\sum_{m=n+1}^{2n}\frac{dq_l}{dp}q_m\left|s_l - s_m\right|\right] \\ =& -(\lambda-1)\left[\sum_{l=1}^{n}\sum_{m=1}^{n}Q_l p Q_m\left|z_l - z_m\right| + \sum_{l=1}^{n}\sum_{m=1}^{n}Q_l\,(1-p)\,Q_m\left|(z_l+v) - z_m\right|\right] \\ &-(\lambda-1)\left[\sum_{l=1}^{n}\sum_{m=1}^{n}-Q_l p Q_m\left|z_l - (z_m+v)\right| + \sum_{l=1}^{n}\sum_{m=1}^{n}-Q_l\,(1-p)\,Q_m\left|z_l - z_m\right|\right]. \end{aligned}$$

Replacing the the order of the sums of the third term and renaming the indices we can obtain

$$\begin{aligned}
&\sum_{l=1}^{n}\sum_{m=1}^{n} -Q_l p Q_m \left|z_l - (z_m + v)\right| \\
=&\sum_{m=1}^{n}\sum_{l=1}^{n} -Q_l p Q_m \left|z_l - (z_m + v)\right| \\
=&\sum_{l=1}^{n}\sum_{m=1}^{n} -Q_m p Q_l \left|z_m - (z_l + v)\right| \\
=&\sum_{l=1}^{n}\sum_{m=1}^{n} -Q_m p Q_l \left|(z_l + v) - z_m\right|,
\end{aligned}$$

and then

$$\begin{aligned}
\frac{dU}{dp} =& -(\lambda-1)\left[\sum_{l=1}^{n}\sum_{m=1}^{n} pQ_lQ_m\left|z_l - z_m\right| + \sum_{l=1}^{n}\sum_{m=1}^{n}(1-p)Q_lQ_m\left|(z_l+v) - z_m\right|\right] \\
&-(\lambda-1)\left[-\sum_{l=1}^{n}\sum_{m=1}^{n} pQ_lQ_m\left|(z_l+v) - z_m\right| - \sum_{l=1}^{n}\sum_{m=1}^{n}(1-p)Q_lQ_m\left|z_l - z_m\right|\right] \\
=& -(\lambda-1)\sum_{l=1}^{n}\sum_{m=1}^{n} Q_lQ_m\left[(2p-1)\left|z_l - z_m\right| - (2p-1)\left|(z_l+v) - z_m\right|\right] \\
=& -(\lambda-1)(2p-1)\sum_{l=1}^{n}\sum_{m=1}^{n} Q_lQ_m\left[\left|z_l - z_m\right| - \left|v + (z_l - z_m)\right|\right] \\
=& (\lambda-1)(2p-1)\sum_{l=1}^{n} Q_l^2 v \\
&+(\lambda-1)(2p-1)\sum_{l=1}^{n}\sum_{m=l+1}^{n} Q_lQ_m\left[\left|v + (z_l - z_m)\right| + \left|v + (z_m - z_l)\right| - 2\left|z_l - z_m\right|\right].
\end{aligned}$$

We only need to take care of the sum of absolute values in the brackets. Denote $x_{lm} = z_l - z_m$, then

$$\begin{aligned}
|v + x_{lm}| + |v - x_{lm}| - 2|x_{lm}| &= \begin{cases} v + x_{lm} - v + x_{lm} - 2x_{lm} & v < x_{lm} \\ v + x_{lm} + v - x_{lm} - 2x_{lm} & 0 \le x_{lm} < v \\ v + x_{lm} + v - x_{lm} + 2x_{lm} & -v \le x_{lm} < 0 \\ -v - x_{lm} + v - x_{lm} + 2x_{lm} & x_{lm} < -v \end{cases} \\
&= \begin{cases} 0 & v < x_{lm} \\ 2v - 2|x_{lm}| & 0 \le x_{lm} < v \\ 2v - 2|x_{lm}| & -v \le x_{lm} < 0 \\ 0 & x_{lm} < -v \end{cases} \\
&= 2\max\{v - |x_{lm}|, 0\}.
\end{aligned}$$

Combining everything together,

$$\frac{dU}{dp} = (\lambda - 1)(2p - 1)\left[\sum_{l=1}^{n} Q_l^2 v + 2\sum_{l=1}^{n}\sum_{m=l+1}^{n} Q_l Q_m \max\{v - |z_l - z_m|, 0\}\right].$$

This is an increasing function of p. QED.

A simple example for calibration. To get an approximation of the magnitude of the discouragement/encouragement effect without narrow bracketing, assume a total of $n + 1$ rounds considered by a participant. To simplify, assume that prizes of all rounds equal v, and the probability of winning each round alone is $\frac{1}{2}$. Without the last round, the possible outcomes of the total lottery are 0, v, $2v$,..., nv, and the probability to get each outcome is

$$\begin{aligned}
P(\tilde{v} = kv) &= \binom{n}{k} \cdot \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \\
&= \binom{n}{k} \cdot \frac{1}{2^n}.
\end{aligned}$$

The differences $|z_l - z_m|$ equal at least v when l, m refer to different outcomes, therefore the double-sum term in the Claim's equation vanishes. The marginal expected EBRD utility with respect to the last round's probability p is

$$\frac{dU_{\text{EBRD}}}{dp_{n+1}} = (\lambda - 1)(2p_{n+1} - 1)\, v \sum_{k=0}^{n} \left(\binom{n}{k} \cdot \frac{1}{2^n}\right)^2.$$

Using this sum's closed formula, we get

$$\frac{dU_{\text{EBRD}}}{dp_{n+1}} = (\lambda - 1)\,(2p_{n+1} - 1)\,v\frac{\left(n - \frac{1}{2}\right)!}{\sqrt{\pi}n!},$$

where $\left(n - \frac{1}{2}\right)! = \Gamma\left(n + \frac{1}{2}\right)$. We can compare it to the marginal expected EBRD utility of one round only,

$$\frac{dU_{\text{EBRD}}}{dp} = (\lambda - 1)\,(2p - 1)\,v,$$

and see that the n-dependent factor weighs the marginal expected EBRD utility of n rounds relative to the marginal utility of one round. Following are some examples for this weight, to illustrate its decline with the number of considered additional rounds:

n	0	1	2	3	5	9
$\frac{\left(n-\frac{1}{2}\right)!}{\sqrt{\pi}n!}$	1.000	0.500	0.375	0.312	0.246	0.185

D Additional Results

D.1 Comparison to Gill and Prowse's (2012) Results, and Interaction of Prize and Robot Score

Although our design is not identical to Gill and Prowse's (2012), they are similar enough to allow comparison of effect magnitudes, reported in Table D.1. Under both designs participants need to perform a repetitive task for two minutes; our task was calibrated such that our participants obtain scores similar on average to theirs; and our subtractive experiment's compensation formula is identical to theirs.

Panels A and B compare our main discouragement/encouragement effects to Gill and Prowse's discouragement effect. The specification we use for Gill and Prowse's data in panel B is identical to our panel A's main specification (which repeats the main results from Table 2, with Second Mover effort instead of human score and First Mover effort instead of robot score).[20] Our prize-size effect is somewhat weaker than Gill and Prowse's and is less precisely estimated, but we find robot score and First Mover effort effects that are remarkably close. In standardized terms, the discouragement effect in our subtractive-compensation experiment (-0.109, SE $= 0.035$) may be stronger than Gill and Prowse's average discouragement effect (-0.051, SE $= 0.030$), but the difference between them is not statistically significant.

[20] It is not the main specification in Gill and Prowse's (2012) Table 2; it does not include a Prize × First Mover effort interaction term and it includes fixed effects rather than random effects. We estimate it based on their publicly available data.

Table D.1: Effect of robot score, prize and their interaction on human score, and the parallel effects based on Gill and Prowse's (2012) data

	Subtractive	Additive	Pooled
			Abs. Avg.
A. Human Score			
Prize ($)	0.311	0.373	0.340
	(0.172)	(0.207)	(0.135)
Robot Score	−0.043	0.023	−0.034
	(0.014)	(0.018)	(0.011)
Robot sign flipped in Additive			X
N	600	600	1200
B. Gill and Prowse (2012): Second Mover Effort			
Prize ($)	0.445		
	(0.155)		
First Mover Effort	−0.045		
	(0.026)		
N	590		
C. Human Score			
Prize ($)	0.242	0.160	0.410
	(0.334)	(0.407)	(0.280)
Robot Score	−0.049	0.004	−0.028
	(0.028)	(0.036)	(0.022)
Prize × Robot Score	0.003	0.009	−0.003
	(0.012)	(0.015)	(0.010)
Robot sign flipped in Additive			X
N	600	600	1200
D. Gill and Prowse (2012): Second Mover Effort			
Prize ($)	1.716		
	(0.605)		
First Mover Effort	0.047		
	(0.050)		
Prize × First Mover Effort	−0.051		
	(0.024)		
N	590		

Notes: Panel A: Main results from Table 2. Panel B: An OLS regression using Gill and Prowse's (2012) data, and the same specification as in Panel A (note that it includes fixed effects, rather than random effects as in Gill and Prowse, 2012). Panel C: OLS regressions as those in Table 2, with an added Prize × Robot Score interaction term. Panel D: An OLS regression using Gill and Prowse's (2012) data, and the same specification as in Panel C. This specification is identical to the main one in Gill and Prowse's (2012) Table 2, except that it includes fixed effects rather than random effects as in their paper. Standard errors in parenthesis. Pooled column: see Table 2's notes.

Next, Gill and Prowse's main results concern the effect of the interaction between prize size and First Mover effort, as the theory predicts not only a first-order discouragement/encouragement effect, but also a second-order effect which should increase with prize size. In panel C we estimate a specification that includes an interaction term of prize and robot score. Panel D shows Gill and Prowse's main results.[21] Unlike their results, we do not find an interaction effect—possibly related to a weak first-order prize effect in our experiment.

D.2 Decomposing Human Score into Quantity and Quality

Our main outcome variable, the human's score in a round, can be thought of as a combination of the number of submission attempts made by the participant and the quality of these attempts—how many are correct (recall that incorrect attempts penalize participants by 10 seconds).

This appendix investigates our main findings when our main measure, human score, is replaced with either its quantity or quality component: total number of attempts and percent of correct attempts, respectively.

Table D.2 shows that our main findings generally hold when replacing human score with either of these two alternative measures.

D.3 Round Completion and Duration

This section investigates whether discouragement/encouragement do not only affect the intensive-margin decision how hard/fast to work, but also the extensive-margin decision how long to work for. Table D.3 shows that results remain qualitatively the same when replacing human score with either a 0/1 indicator for whether the participant completed the round, or a continuous measure of round duration (in seconds, up to the full round's 120 seconds or until the participant quit).[22]

[21] There is a slight difference between these results and the original results presented in their Table 2: we choose to use fixed effects rather than random effects. This barely changes Gill and Prowse's original results (which are a prize effect of 1.639, SE = 0.602, a First-Mover-effort effect of 0.044, SE = 0.049, and an interaction effect of −0.049, SE = 0.023).

[22] The round duration for 6 participants was recorded as exactly 119 seconds in 57 of their 60 rounds. This is likely a bug in the data logger, and we hence treat these participants as having completed the full 120 seconds in all their rounds. (There are no other records of 119-second-duration rounds. Dropping these participants has a small effect on our main results and does not change them qualitatively.)

Table D.2: Effect of robot score and prize on human score and its decomposition into quantity and quality measures

	Sub.	Add.	Pooled		
			Raw Diff.	Abs. Diff.	Abs. Avg.
A. Overall: Human Score					
Prize ($)	0.311	0.373	0.311	0.311	0.340
	(0.172)	(0.207)	(0.191)	(0.191)	(0.135)
Robot Score	−0.043	0.023	−0.043	−0.043	−0.034
	(0.014)	(0.018)	(0.015)	(0.015)	(0.011)
Prize × Additive			0.062	0.062	
			(0.269)	(0.269)	
Robot Score × Additive			0.066	0.021	
			(0.023)	(0.023)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200
B. Quantity: Number of attempts					
Prize ($)	0.350	0.422	0.350	0.350	0.385
	(0.168)	(0.205)	(0.188)	(0.188)	(0.132)
Robot Score	−0.037	0.025	−0.037	−0.037	−0.032
	(0.013)	(0.018)	(0.015)	(0.015)	(0.011)
Prize × Additive			0.072	0.072	
			(0.265)	(0.265)	
Robot Score × Additive			0.062	0.012	
			(0.022)	(0.022)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200
C. Quality: Percent of correct attempts					
Prize ($)	0.430	0.111	0.430	0.430	0.265
	(0.325)	(0.266)	(0.297)	(0.297)	0.211
Robot Score	−0.079	0.009	−0.079	−0.079	−0.048
	(0.026)	(0.023)	(0.024)	(0.024)	(0.017)
Prize × Additive			−0.319	−0.319	
			(0.421)	(0.421)	
Robot Score × Additive			0.089	0.070	
			(0.035)	(0.035)	
Robot sign flipped in Additive				X	X
N	595	589	1184	1184	1184

Notes: OLS regressions based on eq. 5, while replacing the dependent variable of (overall) human score (panel A) with total number of attempts (a measure of quantity; panel B) or with the percent of correct attempts (a measure of quality; panel C). Standard errors in parenthesis. Pooled 4th and 5th columns: see Table 2's notes. Some observations are missing in panel C due to rounds with a human score of zero.

Table D.3: Effect of robot score and prize on round completion and round duration

	Subtractive	**Additive**	**Pooled**		
			Raw Diff.	Abs. Diff.	Abs. Avg.
A. Human Score					
Prize ($)	0.311	0.373	0.311	0.311	0.340
	(0.172)	(0.207)	(0.191)	(0.191)	(0.135)
Robot Score	−0.043	0.023	−0.043	−0.043	−0.034
	(0.014)	(0.018)	(0.015)	(0.015)	(0.011)
Prize × Additive			0.062	0.062	
			(0.269)	(0.269)	
Robot Score × Additive			0.066	0.021	
			(0.023)	(0.023)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200
B. Completed round					
Prize ($)	0.0178	0.0265	0.0178	0.0178	0.0220
	(0.0103)	(0.0087)	(0.0095)	(0.0095)	(0.0067)
Robot Score	−0.0025	0.0008	−0.0025	−0.0025	−0.0017
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0006)
Prize × Additive			0.0087	0.0087	
			(0.0135)	(0.0135)	
Robot Score × Additive			0.0033	0.0017	
			(0.0011)	(0.0011)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200
C. Round duration (sec)					
Prize ($)	1.829	1.888	1.829	1.829	1.850
	(0.667)	(0.746)	(0.709)	(0.709)	(0.500)
Robot Score	−0.188	0.104	−0.188	−0.188	−0.150
	(0.053)	(0.066)	(0.057)	(0.057)	(0.042)
Prize × Additive			0.059	0.059	
			(1.001)	(1.001)	
Robot Score × Additive			0.292	0.085	
			(0.084)	(0.084)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200

Notes: OLS regressions based on eq. 5, while replacing the outcome variable of human score (panel A) with either completion of the full round's duration (panel B), or with the time elapsed since the beginning of the round once the round is terminated due to quitting or due to completing its full duration (panel C). Pooled 4th and 5th columns: see Table 2's notes.

D.4 Robustness to the Difference between Initially Projected Robot Score and Final Robot Score

As mentioned in footnote 6, there were some differences between projected and final robot score. Due to the linear approximation of the non-linear relation between the robot's movement speed's scaling factor and the robot's score, in some rounds the initially projected robot score shown to participants differ from the actual, final robot score, and the difference quickly shrinks once the robot gains a few score points.

Our main-text specifications in Table 2 effectively assume this difference away by using final robot score as main independent variable. Table D.4 shows that replacing final robot score with initially projected robot score does not change our results (qualitatively or quantitatively), suggesting that initial differences between the two are too small to matter for our results.

D.5 Effect of Prior Rounds

This appendix investigates whether winning expectations from previous rounds affect a current round's effort choice. To do so, we test the effect of a previous round's robot score on this round's effort choice.

Table D.5 shows that the including expectations manipulation from the previous round, i.e., the previous round's robot score, has little effect on the current round's robot score's effect on human score. Panel A of the Table shows our main results without including lagged robot score, using only rounds $2-10$ (i.e., using the relevant sample for the lagged analysis), and Panel B adds the lagged variable.

E Expectations Elicitation

This appendix analyzes winning expectations, which were elicited from a random half of the participants in the additive-compensation experiment. It tests whether our design effectively changed participants' winning expectations according to the probability formula (eq. 3). Consistent with our design rationale, we find that the exogenous variation in robot score strongly affects elicited expectations in the direction predicted by the theory, however with with a coefficient smaller than 1. We also find that elicited expectations' positively correspond to human score, with an even smaller coefficient. Using only the participants who did not undergo elicitation, we find that the elicitation itself is unlikely to drive the differences we find between the additive and subtractive compensation schemes.

Table D.4: Effect of either robot score or initially projected robot score and prize on human score

	Subtractive	Additive	Pooled		
			Raw Diff.	Abs. Diff.	Abs. Avg.
A. Human Score					
Prize ($)	0.311	0.373	0.311	0.311	0.340
	(0.172)	(0.207)	(0.191)	(0.191)	(0.135)
Robot Score	−0.043	0.023	−0.043	−0.043	−0.034
	(0.014)	(0.018)	(0.015)	(0.015)	(0.011)
Prize × Additive			0.062	0.062	
			(0.269)	(0.269)	
Robot Score × Additive			0.066	0.021	
			(0.023)	(0.023)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200
B. Human Score					
Prize ($)	0.313	0.366	0.313	0.313	0.338
	(0.172)	(0.207)	(0.191)	(0.191)	(0.135)
Projected Robot Score	−0.051	0.035	−0.051	−0.051	−0.043
	(0.017)	(0.020)	(0.019)	(0.019)	(0.013)
Prize × Additive			0.054	0.054	
			(0.270)	(0.270)	
Projected Robot Score × Additive			0.086	0.016	
			(0.027)	(0.027)	
Robot sign flipped in Additive				X	X
N	600	600	1200	1200	1200

Notes: OLS regressions based on eq. 5, while replacing the independent variable of (final) robot score (panel A) with the initially projected robot score (panel B). Standard errors in parenthesis. Pooled 4th and 5th columns: see Table 2's notes.

Table D.5: Effect of lagged robot score on the main results

	Subtractive	**Additive**	**Pooled**
			Abs. Avg.
A. Human Score			
Prize ($)	0.328	0.363	0.346
	(0.187)	(0.223)	(0.146)
Robot Score	−0.047	0.027	−0.038
	(0.015)	(0.020)	(0.012)
Robot sign flipped in Additive			X
N	540	540	1080
B. Human Score			
Prize ($)	0.342	0.356	0.331
	(0.189)	(0.221)	(0.146)
Robot Score	−0.049	0.020	−0.036
	(0.015)	(0.020)	(0.012)
Robot Score in previous round	−0.009	−0.050	0.018
	(0.015)	(0.020)	(0.012)
Robot sign flipped in Additive			X
N	540	540	1080

Notes: OLS regressions based on eq. 6, where panel A shows the main results as in Table 2 but for rounds 2–10 only, and panel B adds the effect of lagged robot score. Standard errors in parenthesis. Pooled column: see Table 2's notes.

Specifically, the additive-compensation experiment included a randomly selected subsample of 29/60 participants, who were asked in each round, after observing the robot's projected score and before starting to work: "With what probability of winning the prize do you expect to finish this round?"

Based on their responses, we estimate the following regression:

$$\text{ExpectedChance}_{it} = \beta_0 + \beta_1 \times \text{Robot}_{it} + \beta_2 \times \text{Human}_{it} + \varepsilon_{it}, \tag{E.1}$$

where Robot_{it} and Human_{it} are the robot and human scores in participant i's round t, respectively, and ε_{it} is an error term.

Given that the actual probability to win increases by 1 percent for each unit increase in human or robot score, rational expectations, that accurately respond to changes in scores, imply both $\beta_1 = 1$ and $\beta_2 = 1$. In contrast, if people's expectations do not respond to changes in score, we would predict $\beta_1 = 0$ and $\beta_2 = 0$. The results in Table E.1 indicate that while (a) both coefficients are strongly above 0, (b) people's expectations respond more to changes in the robot's score than they co-move with participants' own score, and (c) both coefficients are smaller than 1. (The β_2 estimator does not have a causal interpretation, because the expected winning chance and human's score—i.e., their chosen level of effort—may be jointly determined.) In the rest of this appendix we discuss these results in more detail.

Table E.1: Elicited winning expectations as a function of robot and human score

Variable	Estimate	Std. Error	p-value of coeff. $= 1$	p-value of coeff. $= 0$
Robot Score	0.64	0.05	< 0.001	< 0.001
Human Score	0.24	0.09	< 0.001	0.011

Notes: OLS regression results based on eq. E.1. Dependent variable: elicited winning expectations.

First, these results strengthen the interpretation that our design successfully manipulates humans' expectations, and that expectations are a main driver of our findings as suggested by the EBRD model. Specifically, manipulating robot score indeed strongly shifts winning expectations, suggesting that the combination of compensation scheme and robot performance operates through humans' expectations—as intended by our design.

Second, these results suggest that participants' expectations may only partially satisfy the rational-expectations assumption of the EBRD model. Specifically, they suggest that expectations may be closer to the rational benchmark in their response to an exogenous manipulation of robot's score, and farther from this benchmark in their (weaker) response

to human's score, i.e., in the participants' ability to accurately predict their own actions.

However, it is possible that the estimated coefficients are smaller than 1 due to non-standard noise in elicited expectations (e.g., truncation at 0 and 1), and due to standard noise in human score relative to planned human score. More detailed modeling of measurement error and of participants' error in effort choice, which is outside the scope of this work, may be required to study participants' expectations more thoroughly.

To conclude, while these results may suggest more complicated models of expectations than rational expectations, and/or measurement error, they are still in line with the general mechanism described by the EBRD model and are hence generally consistent with the theoretical predictions and interpretations of our main results.

E.1 Effects of the Elicitation Itself

Eliciting expectations can be an intrusive form of measurement, and may change participants' choices and outcomes in our experiment. Since it was only conducted as a part of the additive-compensation experiment and not as a part of the subtractive-compensation one, we verify that it is not a main driver of the differences we find between the two experiments.

Using only the subsample that did not undergo expectations elicitation in the additive-compensation experiment ($N = 310/600$), we find, albeit with less statistical power, that results remain similar to those in the full sample. The effect of robot score on human score changes from 0.023 (SE = 0.018) to 0.011 (SE = 0.026), such that the difference between robot-score effects in the two experiments (Table 2 in the main text) changes from 0.066 (SE = 0.023) to 0.055 (SE = 0.027). The difference in adjusted robot score (flipping the robot's sign in the additive experiment) across the experiments changes from 0.021 (SE = 0.023) to 0.032 (SE = 0.027), and the average effect of adjusted robot score changes remains -0.034 (SE changes from 0.011 to 0.012). The effects on human attitudes remain quantitatively very similar to those in the full sample, and remain highly statistically significant.

F Open Ended Responses

At the end of all 10 rounds, we asked participants to write a few sentences about their experience of working alongside the robot and also to share their comments or thoughts on how they decided the amount of effort to put in the task in each round:

- "Please write a few sentences about your experience of [competing/collaborating] with this robot."

- "Finally, if you have any comments or thoughts you would like to share with us, please write them here. We are especially curious to know: how did you decide in each round how strongly to [compete/work]?"

In both the experiments all participants wrote at least one comment. Two coders independently analyzed the comments to find common sentiments.

In the subtractive-compensation scenario (Sub.), some participants (Coder 1: 13, Coder 2: 15) had fun "competing" against the robot while some (Coder 1: 17, Coder 2: 15) were stressed or frustrated at the robot's superior performance. In the additive-compensation scenario (Add.), some participants (Coder 1: 10, Coder 2: 15) liked "collaborating" with the robot, while roughly half of this number of participants (Coder 1: 5, Coder 2: 7) thought there was no collaboration.

- P059 (Add.): "I enjoyed this experience. It was very fun. I thought the robot was competent from the very start because I am impressed that it can do this at all."

- P039 (Sub.): "[The experience was] frustrating, because it became obvious that I would never be as fast as the robot at reading and discerning, even if I could physically move my block faster—my speed was not an advantage when it came to the robot's accuracy."

- P033 (Add.): "I liked collaborating with the robot and it was nice to know that my probability of winning the prize was not determined by just my own effort (fall back net)."

- P018 (Add.): "[It] was a good experience but I am not sure how much collaborating was really happening. Mostly each person does their part independently and the scores are just summed. I am not sure if I would call it collaborating."

In the additive-compensation scenario, some participants (Coder 1: 11, Coder 2: 11) perceived the robot as a competitor, even though they were technically on the same team:

- P051 (Add.): "I never really collaborated with the robot. It seemed almost like a competition that I was winning even if I knew that we were trying to help me win."

- P058 (Add.): "The robot provided a form of competition even though we were working together."

- P012 (Add.): "As my goal was to beat the robot, it was super challenging and I always gave my maximum concentration and tried to do it as fast as possible."

- P039 (Add.): "Even when the robot was not performing good [sic], I wanted to do well so that my score would increase. And when the robot was performing well, I tried even harder so I would try to 'beat' it."

On the other hand, no participant perceived the robot as a collaborator in the subtractive-compensation scenario.

In accordance with the EBRD model, some participants in the subtractive-compensation scenario (Coder 1: 17, Coder 2: 20) were demotivated from working harder when the robot was expected to get a higher score, while in the additive-compensation scenario a higher robot score motivated some participants (Coder 1: 13, Coder 2: 17) to work harder.

- P012 (Sub.): "I felt myself less motivated to compete when the robot's projected score was very high."

- P029 (Add.): "I realized that the robot's effort influenced my effort. As the robot performed better, I was encouraged to perform better."

Fewer participants reported the opposite patterns of working harder when the robot was expected to get a higher score under subtractive compensation (Coder 1: 10, Coder 2: 6), or of working worked harder when the robot was expected to get a lower score in the additive-compensation scenario (Coder 1: 5, Coder 2: 5):

- P058 (Sub.) : "I felt as though I tried much harder when the robot was better which makes sense under the 'play to your opponents level' mantra"

- P048 (Add.): "When the robot was not going to be as fast (with a lower projected speed), I tried to go faster and be more careful in order to make up for their speed."

In both scenarios, several participants (Sub. – Coder 1: 8, Coder 2: 12; Add. – Coder 1: 21, Coder 2: 25) reported that they gave their best efforts in each round, regardless of both the robot's score and the prize:

- P062 (Sub.): "I competed as strongly as I could each round because I wanted to maximize my chances of winning."

- P064 (Sub.): "I competed with full effort in every round."

- P021 (Add.): "I just tried to get the highest score possible each round."

- P051 (Add.): "I decided to work equally strongly because I knew that higher percentage for any round means more change of winning the award."

In the additive-compensation scenario, a good number of participants (Coder 1: 16, Coder 2: 17) specifically commented on the inconsistency in robot speed, even though the instructions clearly stated that the robot's speed may be different in different rounds:

- P020 (Add.): "The robot seems inconsistent in its effort, and that frustrates me. In rounds where it gets above 20 projected score it seems competent, while other rounds where it gets 10-20 projected score it seems incompetent. The round where it had 8 projected score, it was just unacceptable and made me more frustrated at it."

- P049 (Add.): "In some rounds, robot did significantly bad, but in some rounds, it did quite well, even better than me. But overall, I won't be able to trust this robot since it's efforts are hardly consistent."

We did not see as many comments about the inconsistency in robot speed in the subtractive-compensation scenario (Coder 1: 8, Coder 2: 8).

Finally, some participants saw the robot as an intentional agent, for example P024 (Add.): "I noticed that some rounds the robot would take it easy and sometimes it would work at full capacity without lagging" or P038 (Add.): "It can clearly outperform me if it chooses to be but often or not it decides against it," while some looked at the robot's scores merely as numbers, for example P025 (Add.): "I tried to do my best every round and get at least double what the robots projected score was." We found similar responses in the subtractive-compensation experiment. These findings highlight that robots may or may not be treated as social agents in the workplace.