Journal of Experimental Social Psychology 116 (2025) 104699

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

People reward others based on their willingness to exert effort☆

Yang Xiang [a,*], Jenna Landy [b], Fiery A. Cushman [a], Natalia Vélez [c], Samuel J. Gershman [a,d,e]

[a] Department of Psychology, Harvard University, 52 Oxford St, Cambridge, MA, 02138
[b] Department of Psychology, New York University, 6 Washington Pl, New York, NY, 10003
[c] Department of Psychology, Princeton University, Peretsman Scully Hall, Princeton, NJ, 08540
[d] Center for Brain Science, Harvard University, 52 Oxford St, Cambridge, MA, 02138
[e] Center for Brains, Minds, and Machines, MIT, 43 Vassar St, Cambridge, MA, 02139

ARTICLE INFO

Keywords:
Collaboration
Effort
Reward
bonus
Prospective judgment
Retrospective judgment
Social cognition

ABSTRACT

Individual contributors to a collaborative task are often rewarded for going above and beyond—salespeople earn commissions, athletes earn performance bonuses, and companies award special parking spots to their employee of the month. How do we decide when to reward collaborators, and are these decisions closely aligned with how responsible they were for the outcome of a collaboration? In Experiments 1a and 1b (N = 360), we tested how participants give bonuses, using stimuli and an experiment design that has previously been used to elicit responsibility judgments (Xiang et al., 2023a). Past work has found that responsibility judgments are driven both by how much effort people actually contributed and how much they could have contributed (Xiang et al., 2023a). In contrast, here we found that participants allocated bonuses based *only* on how much effort agents actually contributed. In Experiments 2a and 2b (N = 358), we introduced agents who were instructed to exert a particular level of effort; participants still rewarded effort, but their rewards were more sensitive to the precise level of effort exerted when the agents decided how much effort to exert. Together, these findings suggest that people reward collaborators based on their *willingness* to exert effort, and point to a difference between decisions about how to assign responsibility to collaborators and how to incentivize them. One possible explanation for this difference is that responsibility judgments may reflect causal inference about past collaborations, whereas providing incentives may motivate collaborators to keep exerting effort in the future. Our work sheds light on the cognitive capacities that underlie collaboration.

1. Introduction

We often reward collaborators to recognize their contributions and encourage them to contribute their best efforts. For example, sports teams give Most Valuable Player (MVP) awards to the best-performing player, researchers who make substantial contributions to a research project tend to receive authorship credit, and employees in companies earn bonuses and commissions in proportion to their performance. What role do these performance bonuses play in collaboration, and how do we decide when, and how much, to incentivize collaborators?

Although prior work on this question varies in its details, much of it can be organized around a broad consensus that people are held responsible for aspects of their contribution that they can control (Lanzetta & Hannah, 1969; Rest et al., 1973; Weiner, 1972; Weiner, 1993; Weiner & Kukla, 1970). For example, people are punished more for failed actions when these result from a lack of effort, rather than a lack of ability, because effort is easier to control than ability (Weiner, 1993). Conceptually, however, there are many possible ways to measure what portion of a person's contribution was under their control.

The first possibility is that contributions are measured based on brute *output*—a person who applies 10 N of force to lift a box contributes, in a very literal sense, 10 N of output. Many companies provide monetary bonuses to employees ("pay for performance") in the form of piece rates or for achieving particular sales goals (Agarwal, 1998; Brown, 1990; Joseph & Kalwani, 1998; Kishore et al., 2013; Lazear, 2000; Milkovich & Wigdor, 1991; Van Herpen et al., 2005). Researchers have also found that children allocate more rewards to collaborators who generated more output (Baumard et al., 2012; Hamann et al., 2014; Kanngiesser & Warneken, 2012; Schäfer et al., 2023).

However, a direct mapping from outputs to rewards might be

☆ This paper has been recommended for acceptance by Paul Conway.

* Corresponding author.
E-mail address: yyx@g.harvard.edu (Y. Xiang).

https://doi.org/10.1016/j.jesp.2024.104699
Received 14 March 2024; Received in revised form 2 October 2024; Accepted 18 November 2024
Available online 26 November 2024

missing an important intuition: The same amount of work might require more effort from some people than others, because people have varying competencies (Xiang et al., 2023b) and tasks may vary in difficulty (Bigman & Tamir, 2016). Benching 70 kg will require an average untrained individual to put in an all-out effort but can be a piece of cake for top weightlifters. Because lifting the same weight is more effortful for some people than for others, they may incur higher effort costs, and thus need higher incentives to attempt it. Thus, a second possibility is that people may provide incentives based on how much effort people exerted. Indeed, previous work has found that people punish transgressors who exert more effort to bring about a harmful outcome (Jara-Ettinger et al., 2014). Conversely, people may also offer greater rewards for the same, helpful outcome if it takes more effort to bring it about.

Both output and actual effort result from the interaction between a controllable decision to exert effort and uncontrollable constraints, such as task difficulty and ability. In a deeper sense, effort is controllable to the extent that a person could have exerted more or less of it. Thus, a third possibility is that people assign incentives by making a *counterfactual* judgment about how much effort someone could have exerted. Counterfactual judgments reflect how much an outcome would have changed if the person had acted differently (Gerstenberg et al., 2012, 2017; Gerstenberg et al., 2021; Gerstenberg & Stephan, 2021; Hiddleston, 2005; Lewis, 2000; Schaffer, 2005; Woodward, 2011). Thus, the star striker on a soccer team may get a disproportionate share of the glory—because the team would have lost if they had not scored—even if their teammates expended more effort to run around the field (Miller & Komorita, 1995). This hypothesis is supported by prior work on how people mete out punishment for *harmful* actions; both adults and children between 5 and 11 years of age choose harsher punishments if a person played a causal role in bringing out a harmful outcome (Shultz et al., 1981, 1986).

These three possibilities are not mutually exclusive. For example, recent computational work suggests that people assign responsibility for the outcomes of a collaboration by considering both how much effort people actually contributed and how much they *could have* contributed (Xiang et al., 2023a). Thus, in the present work, we distinguished to what extent each of these three factors—output, actual effort, and counterfactual effort—contribute to people's judgments about how to incentivize collaborators. Additionally, we compared the influence of these factors on reward in the present study to their influence on responsibility judgments in Xiang et al., 2023a. This allowed us to examine whether reward is determined directly from judgments of responsibility, as suggested by prior theories (Lanzetta & Hannah, 1969; Rest et al., 1973; Weiner, 1972, 1993; Weiner & Kukla, 1970).

To this end, we adapted the materials from Xiang et al., 2023a collaborative box-lifting task with one change: Instead of asking participants how responsible each agent is for the outcome of the collaboration, we asked participants to give a bonus to each agent. Using the same task allowed us to (a) dissociate output, actual effort, and counterfactual effort and compare their effects on allocated bonus, and (b) simultaneously juxtapose participants' bonus allocations and responsibility attributions to understand if there is a direct mapping between the two.

To preview our results, in Experiments 1a and 1b ($N = 360$), we found that participants' bonus allocations do not align with models that assign rewards based on how much output agents contributed, nor based on how much they *could have* contributed; rather, they were best explained by how much effort agents *actually* exerted. Moreover, these reward decisions do not align well with responsibility judgments elicited using the same stimuli. In Experiments 2a and 2b ($N = 358$), we further examined how actual effort drives rewards by introducing agents who were instructed to exert a fixed amount of effort. We found that participants still rewarded actual effort, but their rewards were more sensitive to the precise level of effort exerted when the agents decided how much effort to exert. Together, these results suggest that people assign rewards to collaborators based on their willingness to exert effort, rather than responsibility, controllability, or brute output. We discuss the implications of these findings in our General Discussion.

2. Theoretical framework

In the experiments below, participants viewed vignettes where pairs of agents attempted to lift a box together. Participants were provided a bonus fund of $10 per contestant and told that none of the contestants were aware of this fund (therefore they could not behave strategically). Participants chose how much of the bonus fund to give each contestant. Each box has a weight W, and each agent a has a strength $S_a \in [1, 10]$ defined as the maximum degree of force that they can exert. Each agent produces force $F_a \in [0, S_a]$ by exerting a level of effort E_a defined as the proportion of their strength ($E_a = \frac{F_a}{S_a}$). Whether the team succeeds depends on whether the agents' combined force exceeds the box weight. In other words, $\sum_a F_a \geq W$ results in success ($L = 1$); $\sum_a F_a < W$ results in failure ($L = 0$). We use B_a to denote the bonus allocated to agent a after the lift attempt.

We adapted the models in Xiang et al., 2023a to the bonus allocation problem. The models are summarized in Table 1. These models decide how much bonus to allocate to one of the two agents—the *focal agent*—by considering different factors. These include three actual-contribution models that assign bonuses based only on the focal agent's actual contributions (Force, Strength, and Effort models), three counterfactual-contribution models that assign bonuses based on counterfactual judgments about how much effort the focal agent and their partner—the non-focal agent—could have contributed (Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models), and an Ensemble model that averages the Effort model and the Both-agent counterfactual model. In addition, for failed lift attempts, we reversed the model predictions such that more responsibility (i.e., more blame) leads to less bonus. This is because, while responsibility has two valences depending on the outcome (more responsibility means more blame for failures and more credit for successes), bonuses are only positive (more bonus is always better than less bonus). Below, we describe each model in detail.

2.1. Actual-contribution models

- **Force model**. The force (F) model allocates bonuses based on how much force an agent generates in the event. Agents who exert more force are rewarded more:

$$B_a^F \propto F_a \tag{1}$$

- **Strength model**. The strength (S) model allocates bonuses based on an agent's strength. Stronger agents are rewarded more for successes and rewarded less for failures:

$$B_a^S \propto \begin{cases} S_a & \text{if } L = 1 \\ 10 - S_a & \text{if } L = 0 \end{cases} \tag{2}$$

Table 1
Summary of the models.

Reasoning style	Model
Actual-contribution	Force
Assigns bonuses based on the focal agent's *actual* properties	Strength
	Effort
Counterfactual-contribution	Focal agent only
Assigns bonuses based on how much effort agent(s) could have exerted	Non-focal agent only
	Both agents
Ensemble	Ensemble
Averages the outputs of the Effort model and the Both-agent counterfactual model	

- **Effort model**. The effort (E) model allocates bonuses based on the level of effort an agent exerts. Agents who exert more effort are rewarded more:

$$B_a^E \propto E_a \tag{3}$$

2.2. Counterfactual-contribution models

The counterfactual-contribution models consider how the outcome could have been different if agents had exerted a different level of effort E'. As in prior work (Sanna & Turley, 1996), here we consider only directional counterfactuals (upward for failures, downward for successes). In other words, when agents fail, we consider what would have happened if they exerted more effort; when agents succeed, we consider what would have happened if they exerted less effort.

Each agent receives a bonus proportional to the probability that they could have changed the outcome by altering their effort allocation, defined as:

$$P_a = \begin{pmatrix} \sum_{E'_a} P(E'_a)\,\mathbb{I}\left[E'_a S_a + E_{/a} S_{/a} < W\right] & \text{if } L = 1 \\ \sum_{E'_a} P(E'_a)\,\mathbb{I}\left[E'_a S_a + E_{/a} S_{/a} \geq W\right] & \text{if } L = 0, \end{pmatrix} \tag{4}$$

where a indexes the focal agent, $/a$ indexes the non-focal agent, and $\mathbb{I}[\cdot] = 1$ if its argument is true (0 otherwise). Following Xiang et al., 2023a, we assume that counterfactual efforts (E'_a) are drawn from discrete uniform distributions in increments of 0.01, where $E'_a \in (E_a, 1]$ when agents fail and $E'_a \in [0, E_a)$ when agents succeed.

- **Focal agent only** The Focal-agent-only (FA) counterfactual model only considers counterfactual actions on the part of the focal agent. The model allocates bonuses based on the likelihood of the focal agent changing the outcome by altering her effort allocation, while holding the non-focal agent's effort allocation fixed:

$$B_a^{FA} \propto \begin{pmatrix} P_a & \text{if } L = 1 \\ 1 - P_a & \text{if } L = 0 \end{pmatrix} \tag{5}$$

In other words, the more likely the focal agent is able to change the outcome, the less bonus she gets for failures and the more bonus she gets for successes.

- **Non-focal agent only** The Non-focal-agent-only (NFA) counterfactual model only considers counterfactual actions of the non-focal agent. Holding the focal agent's effort allocation fixed, the more likely the non-focal agent is able to change the outcome, the more bonus the focal agent gets for failures and the less bonus the focal agent gets for successes:

$$B_a^{NFA} \propto \begin{pmatrix} 1 - P_{/a} & \text{if } L = 1 \\ P_{/a} & \text{if } L = 0 \end{pmatrix} \tag{6}$$

- **Both agents** The both-agent (BA) counterfactual model considers counterfactual actions of both the focal agent and the non-focal agent, i.e., a weighted combination of the Focal-agent-only model and the Non-focal-agent-only model. As in Xiang et al., 2023a, we assign equal weights to the two components for simplicity:

$$B_a^{BA} \propto \left(B_a^{FA} + B_a^{NFA}\right)/2 \tag{7}$$

Intuitively, when the lift attempt is successful, this model allocates more bonus to an agent when (a) she is more likely to change the outcome by adjusting her level of effort, and (b) the other agent is less likely to change the outcome by adjusting their effort. When the lift attempt fails, these patterns are reversed.

2.3. Ensemble model

Xiang et al., 2023a found that people's responsibility attributions (denoted by R_a) were best described by an Ensemble model that combines the Effort model and the Both-agent counterfactual model:

$$R_a^{EBA} \propto w R_a^E + (1 - w) R_a^{BA}, \tag{8}$$

where $w \in [0, 1]$ is the weight parameter. When $w = 0$, the Ensemble model recovers the Effort model, and when $w = 1$, the Ensemble model recovers the Both-agent counterfactual model. In past work, the two models were assigned equal weights ($w = 0.5$; an assumption backed up by empirical evidence). Here, as a point of comparison, we include the same Ensemble model:

$$B_a^{EBA} \propto \left(B_a^E + B_a^{BA}\right)/2 \tag{9}$$

In the next section, we put these seven models to the test.

3. Experiments 1a and 1b

In these experiments, we borrowed the task design and stimuli from Experiments 2a and 2b of Xiang et al., 2023a, with one change: Instead of asking for responsibility judgments, we asked participants to allocate bonuses to agents in a range of scenarios that elicit distinct judgments from the seven models. Experiments 1a and 1b differ in agents' "difference-making" ability. This was a manipulation in Xiang et al., 2023a to qualitatively test the predictions of the counterfactual models; an agent is a "difference-maker" if she is able to alter the outcome by changing her effort allocation, while holding the other agent's effort allocation fixed. In Experiment 1a, both agents were always difference-makers, whereas in Experiment 1b, only one agent was a difference-maker at a time.

3.1. Materials and methods

3.1.1. Participants

We recruited 181 participants for Experiment 1a and 179 participants for Experiment 1b via Amazon's Mechanical Turk platform (MTurk).[1] We used the same sample size as in Xiang et al., 2023a for these and subsequent experiments so that we could directly contrast bonus allocations with responsibility judgments. To confirm their understanding of the task, participants completed a comprehension check after reading the instructions. Only those who answered all comprehension check questions correctly were allowed to proceed with the experiment. Participants in Experiment 1a were compensated $3.00 to complete 30 trials, and participants in Experiment 1b were compensated $2.50 to complete 25 trials. To make sure that participants were paying attention, they completed two attention checks during the experiment. Participants who failed one attention check received a warning and were allowed to continue. Participants who failed both attention checks were asked to leave the experiment. These participants' data were not saved, so we do not know the precise count of participants who did not complete the experiment due to failing both attention checks. The experiments were carried out with appropriate institutional approval and pre-registered at https://aspredicted.org/D82_M76. In these and subsequent experiments, we report all measures and manipulations. As we pre-registered, we did not exclude any participant or observation.

3.1.2. Stimuli

Participants saw vignettes where two agents lifted a box together. In

[1] In Experiments 1a, 1b, and 2a, we stopped data collection once 180 participants submitted their responses on MTurk, as stated in the pre-registration, but received 181, 179, and 178 participants' data respectively due to a server error.

every scenario, the two agents were matched along one dimension—strength, effort, or force—in order to tease apart the three actual-contribution models from the different actual-contribution models. If people employed one of these as their metric for bonus allocation, we should expect equal bonus allocation when the agents are matched on that specific aspect. For instance, if bonus allocation was based on effort, then both agents should receive the same bonus when their effort is matched. For each of these three dimensions where agents were matched along, there were 5 unique strength and effort combinations, resulting in a total of 15 combinations; within each combination, we also manipulated the weight of the box to change whether agents were successful (Lift condition) or not (Fail condition), given the same strength and effort allocation.

In Experiment 1a, participants saw all variants of these combinations, for a total of 30 trials (15 combinations × 2 conditions). In Experiment 1b, we removed the 5 Lift trials where force was matched for a total of 15 Fail and 10 Lift trials. This is because the maximum force the two agents can reduce is the same, therefore they are either both difference-makers or no one is.

3.1.3. Procedure

Fig. 1 shows an example trial. Participants watched a game show where pairs of contestants could win prizes by lifting a heavy box together. On each trial, participants see two new contestants and a new box. They observe each contestant's strength, effort, and force, the box weight, and whether they succeeded or failed. Box weights and contestants' strengths were expressed using a 1 to 10 scale, where a box of weight 1 is so light that virtually anyone can lift it, while a box of weight 10 is so heavy that only the strongest humans can lift it by themselves. Contestants' strength determines the heaviest weight they can lift, and their effort determines how much force they actually applied to lift the box. For example, given an all-out, 100 % effort, a contestant of strength 6 would be able to lift a box of weight 6, whereas exerting 50 % of her effort, the contestant can only lift a box of weight 3 or less. Before moving on to the test trials, participants observed four examples of how a single contestant's strength, effort, and force and a box weight determine whether a lift is successful.

On each trial of the test phase, participants observed two contestants trying to lift a box together and assigned bonuses to each contestant individually from a bonus fund of $10 per contestant. We told participants that none of the agents knew about the bonus fund in order to prevent it from altering the agents' reward functions and behavior. Participants entered the bonus they would assign to each contestant using a text box that allowed integers between 0 and 10.

3.1.4. Data analysis

For these and subsequent experiments, we analyzed the data using R (version 4.3.2). To resolve convergence issues, we scaled all the continuous variables and used the BOBYQA optimizer (Powell et al., 2009) for all the linear mixed-effects models.

3.2. Results

Fig. 2 visualizes the data. Experiments 1a (leftmost column) and 1b (third column from the left) produced similar results. This challenges the Focal-agent-only and Non-focal-agent-only counterfactual models, which make qualitatively different predictions across experiments when agents' difference-making ability is manipulated.

Overall, we see an intercept shift between Fail and Lift conditions (see the "Exp 1a Bonus" and "Exp 1b Bonus" columns in Fig. 2). To validate this, we fit a linear mixed-effects model predicting participants' bonus allocations with Condition and Intercept, along with random intercept and random slope for Condition grouped by participants, and the results showed that participants assigned more bonus when the lift was successful [$t(180.0) = 17.60, p < .001$ in Experiment 1a and $t(178.0) = 18.04, p < .001$ in Experiment 1b]. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is 0.51 for Experiment 1a and 0.49 for Experiment 1b. See Table 2 for the full regression output. Note that this finding motivated us to deviate from our pre-registered plan and additionally control for Condition in the remaining regression analyses. This analysis decision was subsequently pre-registered for Experiments 2a and 2b.

From Fig. 2, it also seemed like people allocated different amounts of bonuses to agents when their strength or force was matched, but allocated similar bonuses to agents when their effort was matched. To formally check for these effects, we fit a separate linear mixed-effects model for each stimulus category ("Same strength", "Same force", and "Same effort"). Each regression model regressed participants' bonus allocation on Contestant (e.g., stronger or weaker on a Same Effort trial), Condition ("Lift" or "Fail"), and Intercept, with random intercept and random slope for Contestant and Condition grouped by participants. This was not possible for the "Same force" trials in Experiment 1b, which only had one condition ("Fail"), so for those we removed the Condition regressor. When agents had the same strength, participants assigned more bonus to the more effortful agent [$t(180.0) = 21.16, p < .001$ in Experiment 1a and $t(178.0) = 19.42, p < .001$ in Experiment 1b; note that we used the Satterthwaite approximation of the degrees of freedom throughout the paper]. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is 0.31 for Experiment 1a and 0.32 for Experiment 1b. When agents produced the same amount of force, participants gave more bonus to the weaker but more effortful agent [$t(180.0) = -17.24, p < .001$ in Experiment 1a and $t(178.0) = -13.12$, $p < .001$ in Experiment 1b]. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is -0.24 for Experiment 1a and $-$ 0.26 for Experiment 1b. When agents exerted the same level of effort, there was no significant difference in the bonus assigned to agents in Experiment 1a [$t(180.0) = 0.42, p = .678$], whereas participants assigned more bonus to the stronger agent in Experiment 1b [$t(178.0) = 2.72, p = .007$ in Experiment 1b]. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is 0.16 for Experiment 1a and 0.18 for Experiment 1b. See Table 3 for the full regression output. Juxtaposing these plots with results from Xiang et al., 2023a—the two "Xiang et al., 2023a Responsibility" columns in Fig. 2)[2]—the biggest difference is in the "Same effort" trials: the stronger agent is judged to be more responsible, but both agents receive similar bonuses.

We compared participants' bonus allocations to each of the seven computational models. For every computational model, we fit a linear mixed-effects model predicting participants' bonus allocations with the model prediction, Condition, and Intercept, with random intercept and random slope for Condition grouped by participants. The Condition regressor was included as an additive effect to control for its effect on bonus allocation. We did not include random slopes for the model prediction. This is because, for both Experiment 1a and Experiment 1b, regression on the strength model with a maximal random effects structure was near singular, meaning that the variances of one or more random effects were close to zero. A further principal component analysis on the random effects covariance matrix revealed that one of the components captured 0 % variance, suggesting that the regression model was overparameterized. Dropping either the random slope for Condition or the random slope for model prediction fixed the overparameterization, and since dropping the random slope for model prediction produced principal components with more total variance, we decided to drop the random slope for model prediction and keep the random slope for Condition. For consistency, we used the same

[2] We want to remind readers that the effect is flipped in the Fail condition because more responsibility (i.e., more blame) would lead to less bonus.

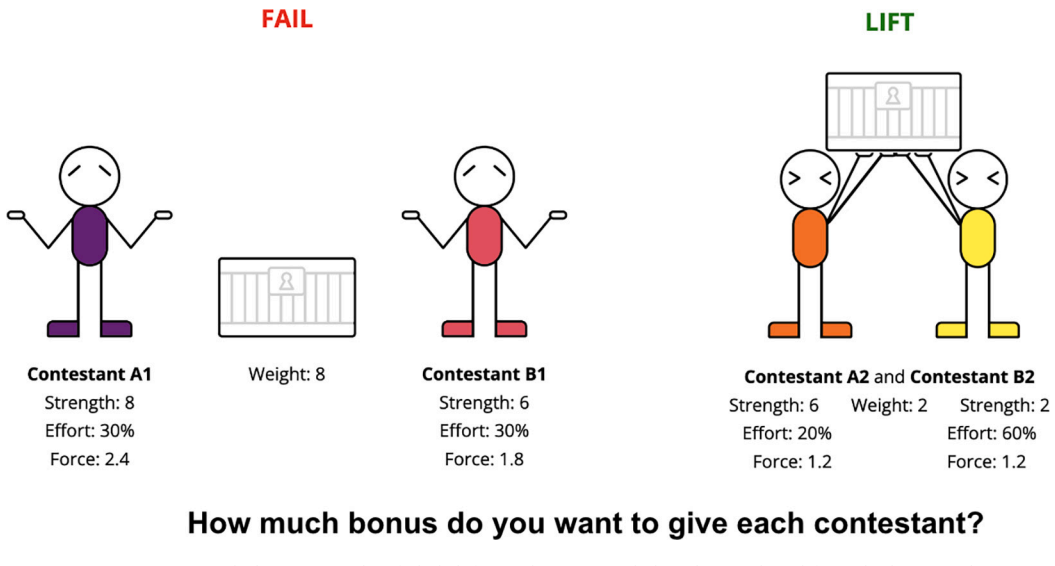

Fig. 1. Example contest in Experiments 1a and 1b. Agents either failed (left panel) or succeeded (right panel) in lifting the box together. Participants observed the box weight and each agent's strength, effort, and force. They assigned a bonus between $0–10 to each agent separately. In each contest, the two agents were matched on either strength, effort, or force. For each of these dimensions agents were matched along, there were 5 unique strength and effort combinations, with seven levels of strength ranging from 2 to 8, seven levels of effort ranging from 20 % to 80 %, and twelve levels of force ranging from 1.0 to 4.9.

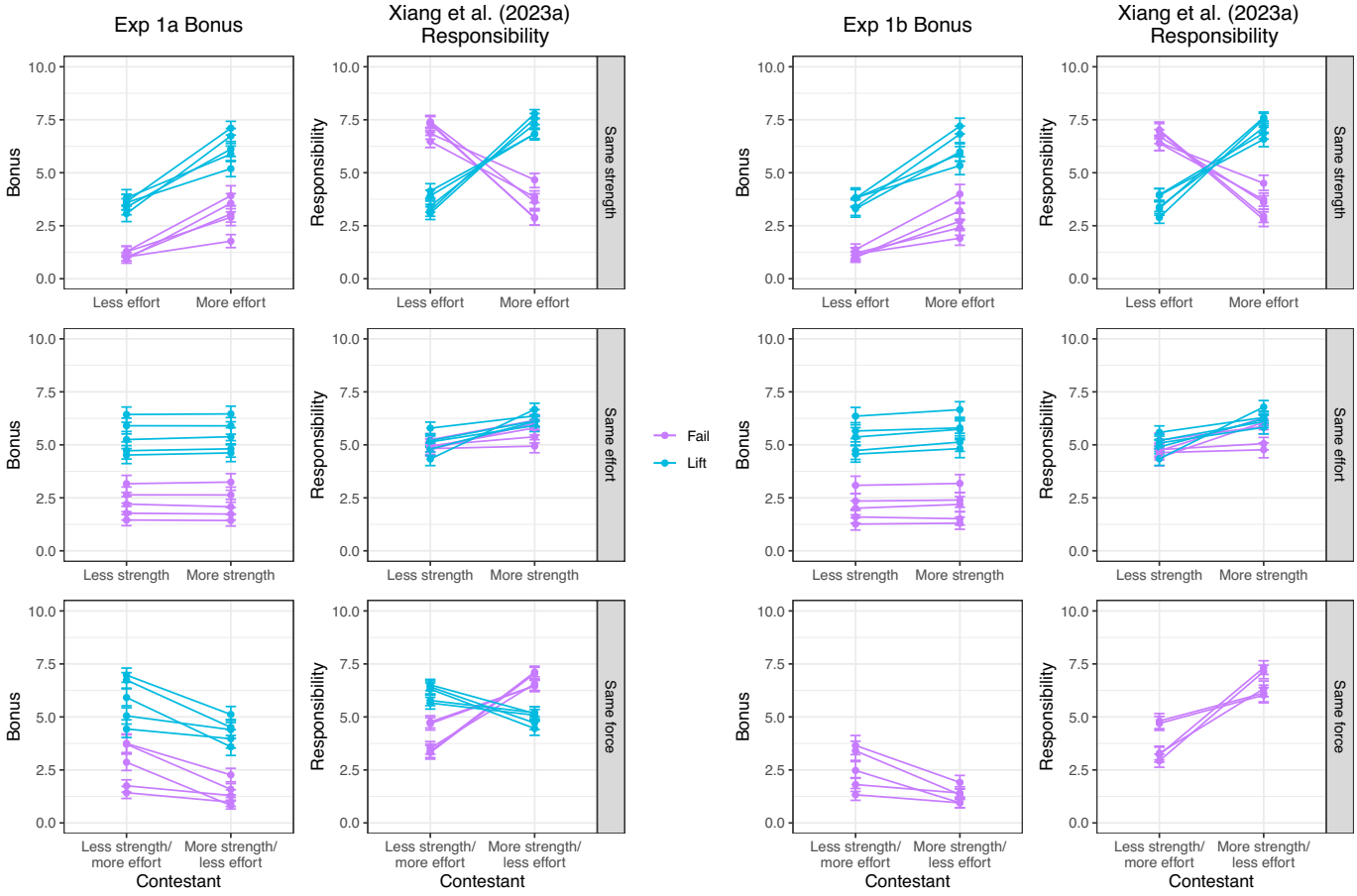

Fig. 2. Results from Experiments 1a and 1b ("Exp 1a Bonus" and "Exp 1b Bonus") and from Xiang et al., 2023a (the two corresponding "Xiang et al., 2023a Responsibility" columns). Each line corresponds to a scenario. Error bars indicate bootstrapped 95 % confidence intervals.

regression formula for all seven computational models.

We then conducted a formal model comparison using the Bayesian Information Criterion (BIC). Lower BIC values indicate better models. We found that the Effort model has the lowest BIC among all models in both experiments, followed by the Ensemble model and the Force model (see left panel of Fig. 3A). None of the counterfactual-contribution models are close to the Effort model in terms of BIC. This result provides a stark contrast to the responsibility attribution results in Xiang

Table 2
Bonus ~ Condition + (1 + Condition | Participant).

	Estimate	SE	df	t-statistic	*p*-value
Exp 1a.					
(Intercept)	2.118	0.137	180.0	15.49	< 0.001
Condition	2.989	0.170	180.0	17.60	< 0.001
Exp 1b.					
(Intercept)	2.004	0.137	178.0	14.63	< 0.001
Condition	3.203	0.178	178.0	18.04	< 0.001

Table 3
Bonus ~ Condition + Contestant + (1 + Condition + Contestant | Participant).

	Estimate	SE	df	t-statistic	p-value
Exp 1a. Same strength.					
(Intercept)	0.912	0.113	180.0	8.10	< 0.001
Condition	2.780	0.169	180.0	16.46	< 0.001
Contestant	2.320	0.110	180.0	21.16	< 0.001
Exp 1b. Same strength.					
(Intercept)	0.925	0.131	178.0	7.05	< 0.001
Condition	2.932	0.179	178.0	16.42	< 0.001
Contestant	2.153	0.111	178.0	19.42	< 0.001
Exp 1a. Same effort.					
(Intercept)	2.225	0.156	180.0	14.23	< 0.001
Condition	3.163	0.183	180.0	17.30	< 0.001
Contestant	0.023	0.055	180.0	0.42	0.678
Exp 1b. Same effort.					
(Intercept)	2.002	0.154	178.0	12.97	< 0.001
Condition	3.392	0.184	178.0	18.42	< 0.001
Contestant	0.174	0.064	178.0	2.72	0.007
Exp 1a. Same force.					
(Intercept)	2.749	0.156	180.0	17.57	< 0.001
Condition	3.024	0.173	180.0	17.48	< 0.001
Contestant	−1.407	0.082	180.0	−17.24	< 0.001
Exp 1b. Same force.					
(Intercept)	2.534	0.166	178.0	15.31	< 0.001
Contestant	−1.222	0.093	178.0	−13.12	< 0.001

Note. For the "Same force" trials in Experiment 1b, the Condition regressor was dropped because these trials could only produce a "Fail" outcome.

et al., 2023a, where the Ensemble model was found to best describe participants' responsibility judgments, followed by the Both-agent counterfactual model and Effort model (see middle panel of Fig. 3A).[3]

Plotting model predictions against the data ("Exp 1a Bonus" and "Exp 1b Bonus" columns in Fig. 4), we see that the Effort model also provides the closest quantitative fit to participants' judgments both when the collaboration failed [Pearson's $r(28) = .99, p < .001$ in Experiment 1a and $r(28) = .98, p < .001$ in Experiment 1b] and when the collaboration was successful [$r(28) = .97, p < .001$ in Experiment 1a and $r(18) = .96, p < .001$ in Experiment 1b]. The Both-agent counterfactual model provides a much worse fit to the data regardless of whether the collaboration failed [$r(28) = .57, p = .001$ in Experiment 1a and $r(28) = .59, p < .001$ in Experiment 1b] or was successful [$r(28) = .55, p = .002$ in Experiment 1a and $r(18) = .75, p < .001$ in Experiment 1b]. The Ensemble model does a decent job in the Fail condition [$r(28) = .88, p < .001$ in Experiment 1a and $r(28) = .91, p < .001$ in Experiment 1b] and the Lift condition [$r(28) = .95, p < .001$ in Experiment 1a and $r(18) = .96, p < .001$ in Experiment 1b], but still worse than the Effort model. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is $r = .21$ for both Experiments 1a and 1b. By contrast, the Ensemble model was the best-performing model for responsibility judgments, followed by the Both-agent counterfactual model and Effort model (see the two "Xiang et al., 2023a Responsibility" columns in Fig. 4).

The Ensemble model was introduced in Xiang et al., 2023a to deal with the inadequacy of the single-factor models: there, the Effort model and the Both-agent counterfactual model were the best at explaining responsibility attribution in each model category (actual-contribution or counterfactual-contribution, respectively), but neither provided a fully adequate account, so the Ensemble model—which averages the outputs of the Effort and Both-agent counterfactual models—was created to explore the possibility of people employing two types of reasoning. However, here we see that effort alone already captures participants' decisions about bonus allocation, and the Ensemble model doesn't provide a better fit than the Effort model.[4]

As an additional point of comparison, we checked if individual participants' responses were also best explained by the Effort model. We fit seven linear models for each participant, each predicting bonus allocations with one of the seven models, and compared the model BICs. We then counted how many participants' responses were best predicted by each of the seven models (see left panel of Fig. 3B). For this analysis, we deviated from our pre-registration and excluded three participants (one from Experiment 1a and two from Experiment 1b) who allocated 0 bonus to all contestants in all trials. This is because, without any variation in their data, we are unable to fit the models, and allocating 0 bonus regardless of contestants' strength, effort, force, or the outcome of the collaboration suggests that these participants might not have been totally compliant with the task. We found that the responses of the majority of participants were best captured by the Effort model (72.2 % of the participants in Experiment 1a and 50.8 % of the participants in Experiment 1b). Only 13.9 % of the participants in Experiment 1a and 19.2 % of the participants in Experiment 1b were best described by the Ensemble model. No participants in Experiment 1a and 5.6 % of the participants in Experiment 1b were best described by the Both-agent counterfactual model. By contrast, in Xiang et al., 2023a, the Ensemble model best explained the data of the majority of participants (56.7 % and 45.8 %, respectively), followed by the Both-agent counterfactual model which best explained 14.4 % and 23.2 % of the participants, and the Effort model only explained 11.7 % and 11.9 % of the participants (see middle panel of Fig. 3B). This contrast again demonstrates that people might consider different factors when making different judgments.

Finally, we were curious whether there were changes in bonus allocations over time. To that end, we fit a linear mixed-effects model predicting participants' bonus allocations with the trial number ("Trial") and Intercept, along with random intercept and random slope for Trial. This exploratory analysis was not pre-registered. We did not find a significant effect of Trial in either experiment [$t(180.0) = 0.98, p = .328$ in Experiment 1a and $t(178.0) = 0.94, p = .347$ in Experiment 1b]. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is 0.11 for Experiment 1a and 0.12 for Experiment 1b. See Table 4 for the full regression output.

[3] For consistency, we also used the same regression formula in the re-analyses of (Xiang et al., 2023a).

[4] In an exploratory analysis where we allowed the Ensemble model's weighting factor to vary, we found that the best-fitting model to participants' bonus allocations predominantly considers Effort ($w \approx 0.9$); incorporating counterfactuals slightly improves the model's quantitative fit to the data, but does not make predictions that are qualitatively distinct from a model that considers Effort alone.

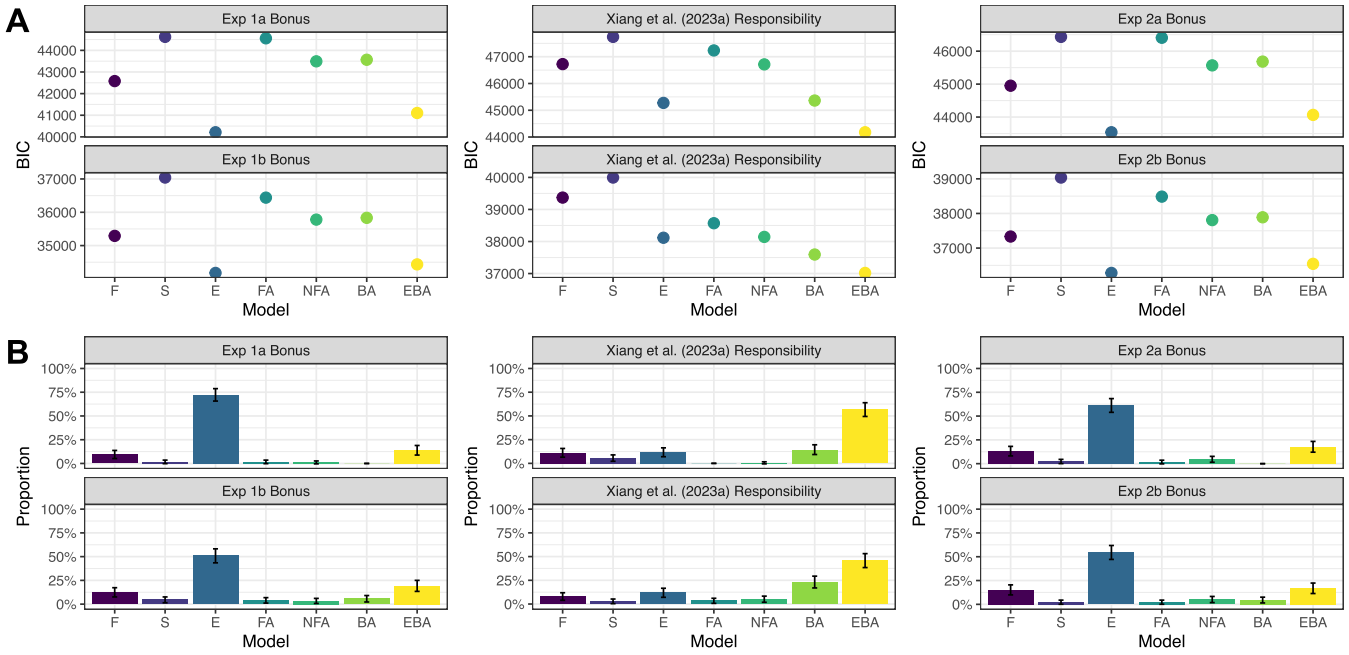

Fig. 3. Model comparison in Experiments 1a and 1b ("Exp 1a Bonus" and "Exp 1b Bonus"), in Xiang et al., 2023a (the two corresponding "Xiang et al., 2023a Responsibility" columns), and in Experiments 2a and 2b ("Exp 2a Bonus" and "Exp 2b Bonus"). (A) Bayesian information criterion (BIC) for each mixed-effects regression model. (B) Proportion of participants best explained by each model. Error bars indicate 95 % confidence intervals of proportions. F = Force model, S = Strength model, E = Effort model, FA = Focal-agent-only counterfactual model, NFA = Non-focal-agent-only counterfactual model, BA = Both-agent counterfactual model, EBA = Ensemble model.

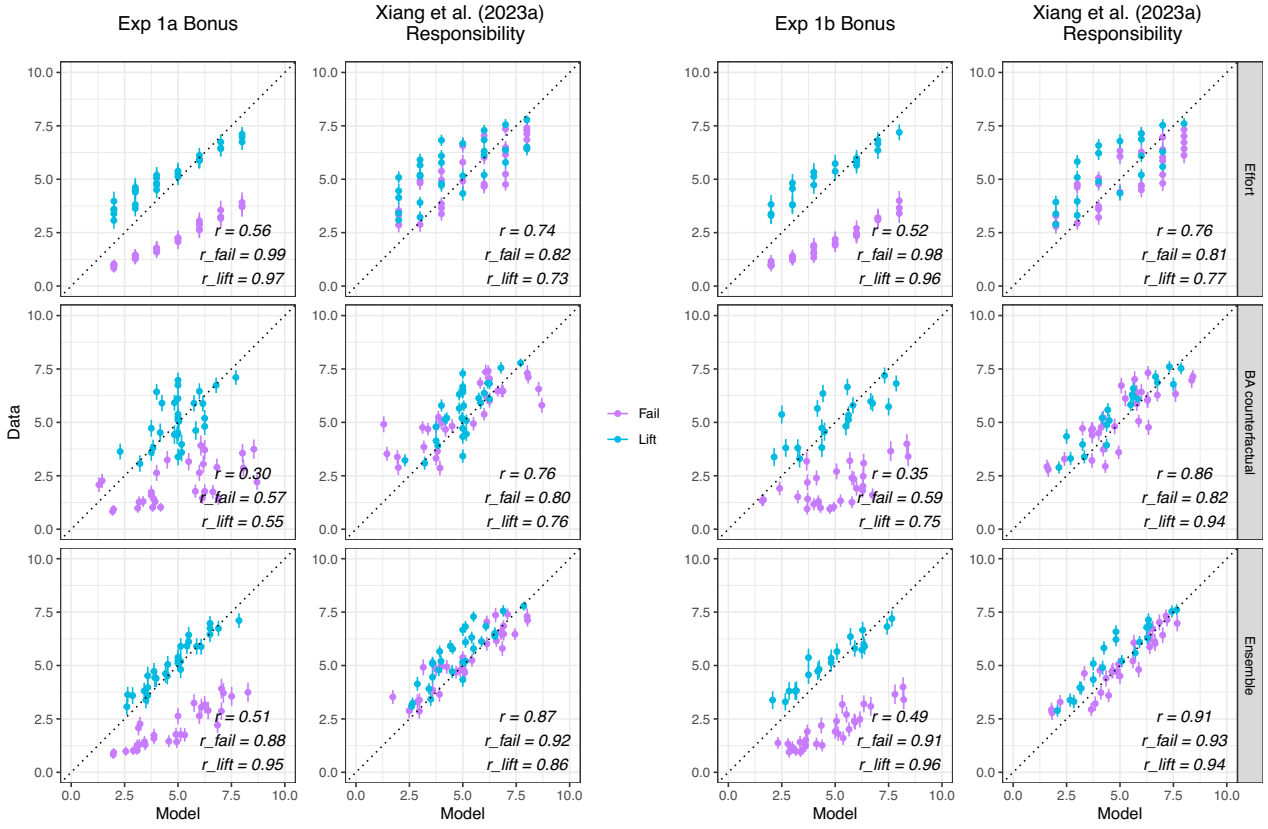

Fig. 4. Data-model comparison in Experiments 1a and 1b ("Exp 1a Bonus" and "Exp 1b Bonus") and in Xiang et al., 2023a (the two corresponding "Xiang et al., 2023a Responsibility" columns). Pearson correlation coefficients for conditions combined and separate are shown at the bottom right of each subplot. Each point denotes one agent in one scenario; error bars indicate 95 % confidence intervals.

Table 4
Bonus ~ Trial + (1 + Trial | Participant).

	Estimate	SE	df	t-statistic	p-value
Exp 1a.					
(Intercept)	3.612	0.127	180.0	28.36	<0.001
Trial	0.036	0.037	180.0	0.98	0.328
Exp 1b.					
(Intercept)	3.285	0.123	178.0	26.61	<0.001
Trial	0.043	0.046	178.0	0.94	0.347

3.3. Discussion

In Experiments 1a and 1b, we compared participants' bonus allocations to seven models: three actual-contribution models (Force, Strength, and Effort), three counterfactual-contribution models (Focal-agent-only, Non-focal-agent-only, and Both-agent), and an Ensemble model which assigns bonuses based on the average of the Effort model and the Both-agent counterfactual model. Model comparison and correlation analysis showed that the Effort model provided the best quantitative fit to the behavioral data. The Effort model also best explained the responses of the majority of participants. These results suggest that bonus allocations are closely tied to subjective effort, rather than actual force or counterfactual reasoning.

Comparing our results to Xiang et al., 2023a, we see that, while people reason about both agents' actual and potential effort when they attribute responsibility, they mostly consider agents' actual effort alone when they allocate bonuses. This shows that participants likely do not base their reward assignments on responsibility. We return to this discrepancy in the General discussion.

4. Experiments 2a and 2b

Experiments 1a and 1b showed that people's decision to reward collaborators is closely tied to how much effort they exerted, and less to who caused the outcome. But why do people reward effort? Are they merely rewarding agents for the fact that they exerted effort (which comes with a cost), or are they rewarding something deeper, for example their dispositions and mental states?

To answer these questions, in Experiments 2a and 2b we added a secrecy manipulation. On every trial, we told participants that one agent was secretly instructed to exert a certain level of effort, marked by a symbol. If participants provide bonuses based just on the level of exerted effort, then whether or not an agent is secretly instructed should not affect the relation between bonus and effort. On the other hand, if participants base their bonus allocations on agents' reasons to exert effort (i.e., rewarding an agent's desire to successfully complete a collaborative task), then we should expect to see an interaction between secrecy and effort, where changes in effort levels produce smaller changes in bonus when an agent is secretly instructed. Note that an interaction between secrecy and effort is potentially also compatible with an account that bases rewards on perceived responsibility, as agents who are instructed what to do are likely less responsible for the effort they contribute. However, if that is the case, we should see that the secret agents get rewarded moderately across different scenarios, and their bonuses should be relatively stable regardless of the outcome of the collaboration.

4.1. Materials and methods

4.1.1. Participants

We recruited 178 participants for Experiment 2a and 180 participants for Experiment 2b via Amazon's Mechanical Turk platform (MTurk). As in Experiments 1a and 1b, participants completed a comprehension check after reading the instructions and two attention checks during the experiment. Participants who failed both attention checks were asked to leave the experiment early, and since their data were not saved, we do not know how many failed both attention checks and left. Participants in Experiment 2a were compensated $3.00 to complete 30 trials, and participants in Experiment 2b were compensated $2.50 to complete 25 trials. The experiments were carried out with appropriate institutional approval and pre-registered at https://aspredicted.org/LTD_XPK.

4.1.2. Stimuli

The stimuli were the same as in Experiments 1a and 1b, except that on each trial, one agent is randomly selected to receive a secret instruction about exactly how much effort to exert. This is indicated by a spy symbol shown right below the agent.

4.1.3. Procedure

The procedure was the same as Experiments 1a and 1b, except that participants also saw which agent was given the secret instruction about exactly how much effort to exert (marked by a spy symbol).

4.2. Results

To test the robustness of the previous findings, we compared participants' bonus allocations to each of the seven computational models. As in Experiments 1a and 1b, for every computational model, we fit a linear mixed-effects model predicting participants' bonus allocations with the model prediction, Condition, and Intercept, with random intercept and random slope for Condition grouped by participants. We then conducted a formal model comparison using the Bayesian Information Criterion (BIC). Lower BIC values indicate better models. We found that the Effort model has the lowest BIC among all models in both Experiment 2a and Experiment 2b, followed by the Ensemble model and the Force model (see right panel of Fig. 3A). Comparing the left and right panels of Fig. 3, we see that the patterns look very similar, suggesting that introducing the secret agent likely did not change the underlying process.

We also conducted the same single-participant analysis as in Experiments 1a and 1b. We fit seven linear models for each participant, each predicting bonus allocations with one of the seven models, and compared the model BICs. Three participants from Experiment 2a and two participants from Experiment 2b were excluded in this analysis because they allocated 0 bonus to all contestants in all trials and hence had no variation in their data. We counted how many participants' responses were best predicted by each of the seven models (see right panel of Fig. 3B). Two participants, one in each experiment, were fit equally well by the Focal-agent-only counterfactual model and Non-focal-agent-only counterfactual model. We took a conservative approach and counted them as both best fit by the Focal-agent-only counterfactual model and best fit by Non-focal-agent-only counterfactual model, to be the most flattering to the two models. Even so, very few participants were best fit by these two models—1.7 % of the participants in Experiment 2a and 2.2 % of the participants in Experiment 2b were best fit by the Focal-agent-only counterfactual model, and 4.6 % of the participants in Experiment 2a and 5.1 % of the participants in Experiment 2b were best fit by the Non-focal-agent-only counterfactual model. As in Experiments 1a and 1b, we found that the responses of the majority of participants were best captured by the Effort model (61.1 % of the participants in Experiment 2a and 54.5 % of the participants in Experiment 2b). Only 17.7 % of the participants in Experiment 2a and 16.9 % of the participants in Experiment 2b were best described by the Ensemble model. No participants in Experiment 2a and 4.5 % of the participants in Experiment 2b were best described by the Both-agent counterfactual model.

Our main goal of Experiments 2a and 2b was to examine whether agents' motivation to exert effort affected the bonuses they received. To that end, we fit a linear mixed-effects model predicting participants'

bonus allocations with agents' exerted effort, secrecy (i.e., whether an agent was secretly instructed), the interaction between effort and secrecy, Condition (Fail or Lift), and Intercept, with random intercept and random slopes for effort, secrecy, the interaction between effort and secrecy, and Condition grouped by participants. The full regression output is shown in Table 5. As with Experiments 1a and 1b, we saw a significant main effect of Condition [$t(177.0) = 16.85, p < .001$ in Experiment 2a and $t(179.0) = 14.87, p < .001$ in Experiment 2b], suggesting that participants assign more bonus to agents when they succeed. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is 0.49 for Experiment 2a and 0.47 for Experiment 2b. We also saw a significant main effect of effort [$t(177.3) = 18.28, p < .001$ in Experiment 2a and $t(179.8) = 19.16, p < .001$ in Experiment 2b], suggesting that bonuses correlate with the level of effort exerted; specifically, more effort is associated with more bonus. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is 0.16 for Experiment 2a and 0.17 for Experiment 2b.

We observed a statistically significant interaction effect between effort and secrecy [$t(177.8) = -3.17, p = .002$ in Experiment 2a and $t(178.6) = -3.89, p < .001$ in Experiment 2b], meaning that changes in effort levels produce greater changes in bonus when agents are not secretly instructed. The minimum effect size detectable by this analysis under standard criteria (80 % power and 5 % false-positive rate) given our sample size is −0.13 for Experiment 2a and − 0.14 for Experiment 2b. This effect is visualized in Fig. 5. This suggests that whether an agent voluntarily decided their effort allocation or was instructed to exert a certain level of effort affects how much bonus people give them: People do not provide bonuses based just on agents' effort, but also on their motivation to exert effort.

The significant interaction effect might seem to be compatible with rewarding others based on responsibility, since the secret agents were being told what to do and thus were less responsible for the level of effort they exerted. If this was true, then participants should be rewarding the secret agents a relatively stable bonus throughout the scenarios, regardless of the outcome of the collaboration. However, an exploratory analysis (see Fig. 6) showed that participants overall assigned similar bonuses to the secret agents and non-secret agents, and for both types of agents, the bonuses were sensitive to the outcome of the collaboration. This indicates that lack of responsibility in deciding how much effort to exert doesn't exculpate the secret agents when the collaboration fails, nor does it diminish the reward they deserve for their effort exertion when the collaboration succeeds.

4.3. Discussion

In Experiments 2a and 2b, we further investigated why people rewarded agents for their effort by adding a secrecy manipulation. We found that the relation between effort and bonus was moderated by secrecy, i.e., the relation between bonus and effort was stronger when agents decided how much effort to exert. This finding suggests that participants were not merely rewarding agents for the level of effort they exerted; they cared about *why* the agents exerted that effort. In addition, participants rewarded the secret agents differently for different outcomes, but similar to the non-secret agent, suggesting that the interaction wasn't from the secret agents' lack of responsibility to decide how much effort to exert.

Table 5
Bonus ~ Effort * Secrecy + Condition + (1 + Effort * Secrecy + Condition | Participant).

	Estimate	SE	df	t-statistic	p-value
Exp 2a.					
(Intercept)	2.618	0.156	177.0	16.73	<0.001
Effort	1.030	0.056	177.3	18.28	<0.001
Secrecy	0.068	0.076	177.2	0.89	0.374
Condition	2.849	0.169	177.0	16.85	<0.001
Effort × Secrecy	−0.136	0.043	177.8	−3.17	0.002
Exp 2b.					
(Intercept)	2.849	0.162	179.0	17.64	<0.001
Effort	1.105	0.058	179.8	19.16	<0.001
Secrecy	0.085	0.072	179.7	1.17	0.243
Condition	2.397	0.161	179.0	14.87	<0.001
Effort × Secrecy	−0.185	0.048	178.6	−3.89	<0.001

5. General discussion

Rewarding the right collaborators can encourage desired behaviors and increase chances of success in the future, but it is unclear how we make these decisions. In Experiments 1a and 1b, we compared participants' bonus allocations to seven models reported in Xiang et al., 2023a using the same design and stimuli. This allowed us to directly contrast bonus allocations with responsibility judgments, and compare the contributions of output (i.e., force), actual effort, and counterfactual effort on bonuses. While participants' responsibility attributions were best predicted by the Ensemble model—which makes judgments based on an equal weighting between actual and counterfactual effort—here we found that participants' bonus allocations were best predicted by actual effort alone. These results suggest that responsibility judgments and bonus allocations differ in meaningful ways, and that participants do not reward agents based on how much output they actually contributed or how much effort they *could have* exerted. Instead, participants rewarded agents based on the effort they *actually* exerted.

In Experiments 2a and 2b, we investigated how effort drives reward assignments by introducing agents who were secretly instructed to exert a certain level of effort. We found a significant interaction effect between effort and secrecy showing that the relation between effort and bonus was stronger when agents themselves decided how much effort to exert. These results deepen our understanding from Experiments 1a and 1b: People reward actual effort not for the effort itself—they also care about *why* a person exerted the effort. When collaborators can decide how much effort to exert, participants are more sensitive to the precise level of effort they choose to exert and reward them more for exerting more effort. Rewarding collaborators based on their *willingness* to exert effort may reflect deeper attributions about each collaborator's motivations. In particular, past work has shown that effort is rewarded, valued, and praised because it reflects the importance of the goal to the agent (Bigman & Tamir, 2016) and serves as a signal about an agent's underlying character (Anderson et al., 2020; Celniker et al., 2023). Our results provide causal evidence that complements this line of work by directly manipulating the reasons behind agents' effort exertion.

Rewarding collaborators' willingness to exert effort may have additional benefits to collaboration. One benefit is that willingness to exert effort increases the chances of success in future collaborations. Recent work has found evidence that rewarding people for their willingness to exert effort can increase their willingness to choose harder tasks when rewards are no longer offered (Lin et al., 2021). In addition, by exerting more effort, agents can also get better at the task and become better collaborators over longer timescales (Xiang et al., 2024). Another benefit is that it encourages qualities of good collaborators such as a strong commitment to the team's goals, which are stable across task domains and contexts (Hackel et al., 2015; Heider, 1958) where agents' competence may vary. Incentivizing traits that benefit the group may also have long-lasting benefits by promoting collaboration (Henrich, 2009; Henrich & Boyd, 2016).

We observed that bonus allocations are determined by effort alone, in contrast to prior work that has found that responsibility attributions also depend on counterfactuals. Both studies used the same stimuli, and both have been replicated across multiple experiments, so it appears that merely prompting participants differently may affect their judgments.

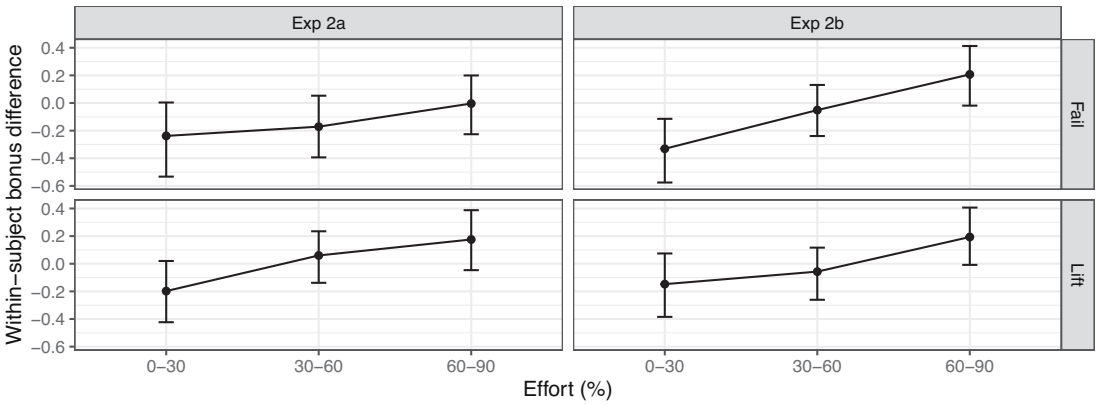

Fig. 5. Within-subject bonus difference (non-secret agent's bonus minus secret agent's bonus). Agents' effort was binned to 0–30 %, 30 %–60 %, and 60 %–90 % for visualization (the highest effort level in the stimuli is 80 %). Error bars indicate bootstrapped 95 % confidence intervals.

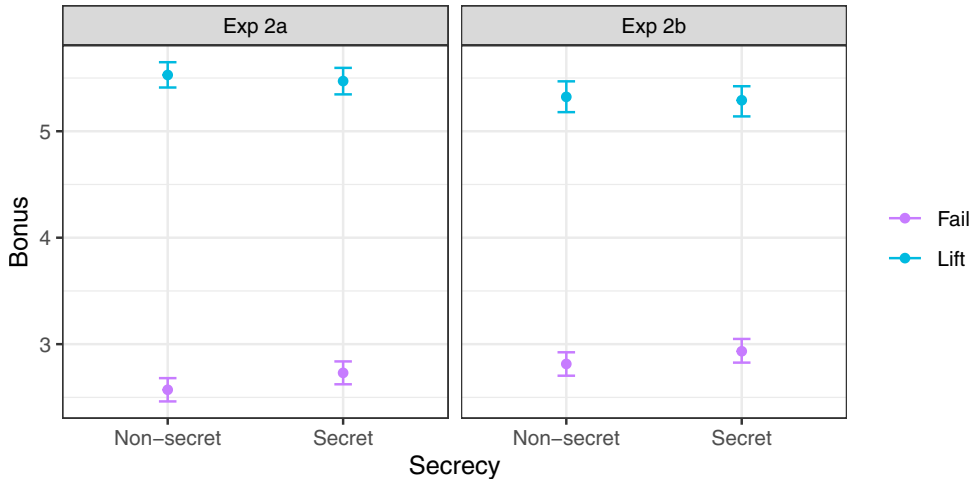

Fig. 6. Bonus as a function of secrecy and Condition. Error bars indicate bootstrapped 95 % confidence intervals.

One possibility for this discrepancy is that judgments about responsibility and bonus allocation may rely on different computations because they are distinct in more fundamental ways. Gerstenberg (2022) made a distinction between counterfactuals and hypotheticals, arguing that counterfactuals are thoughts about changes that lie in the past (e.g., after an event happened, wondering whether it would have happened if someone hadn't been present), whereas hypotheticals are thoughts about changes that lie in the future (e.g., before an event happens, wondering whether it would happen if someone wasn't present). In a similar vein, responsibility attribution seems to be a type of *retrospective judgment* that is sensitive to causal reasoning—understanding who and what led to the success or failure (Gerstenberg, 2022; Gerstenberg et al., 2015; Gerstenberg & Lagnado, 2010; Langenhoff et al., 2021)—and reflects who we think is a better partner (such as more competent and more willing to exert effort in the current setup, or more wealthy and fair in Raihani & Barclay, 2016). By contrast, bonus allocation seems to be a type of *prospective judgment*, based primarily on how to motivate partners to contribute more in the future (Chung et al., 2014; Deversi & Spantig, 2023; Joseph & Kalwani, 1998; Lazear, 2000). Responsibility attributions and bonus allocations might thus reflect two different kinds of judgment: The former situates the agent in the event and considers both the agent's behavior and the environment (e.g., other agents' behavior; Zultan et al., 2012; Xiang et al., 2023a), whereas the latter cares about how to motivate the agent to do the best they can (Bun & Huberts, 2018). It is worth noting that our speculation is more so based on the rationale that a major goal of rewards is to reinforce behaviors in the future. Since our task didn't involve collaborating with the same agents in the future, it is possible that the bonus allocations are also somewhat retrospective but vary meaningfully from responsibility attributions; or that they are a kind of heuristic we apply even in the absence of future collaborations, but the heuristic itself is learned through the need to reinforce successful future collaborations.

Other fields make similar distinctions when evaluating past and future behavior. For example, in the law, this distinction is also reflected in the two standard justifications for state punishment: backward-looking, retributivist justifications, whose principle is to punish people engaging in criminal behavior, versus forward-looking, consequentialist justifications, which justify punishment by its future beneficial effects (Greene & Cohen, 2004; Lacey, 1988). Organizations also distinguish between the two when allocating base pay versus bonus pay: Base pay mostly has to do with supply and demand in the labor market—finding the collaborators that will lead to success (for example, in the context of our experiments, partners who are both competent and willing to exert effort), whereas bonus pay mostly has to do with effort—incentivizing existing collaborators to increase their effort (Bun & Huberts, 2018; Lazear, 2000; Ramos, 2022). Thus, the discrepancies between responsibility attribution and bonus allocation raise the possibility that people are motivated by different goals when they make retrospective versus prospective judgments in collaborations.

This logic might be able to explain why we observed a consistent and significant effect of the collaboration outcome on bonus allocations—namely, participants reward collaborators more when they succeed than when they fail—while Xiang et al., 2023a did not find a similar effect in responsibility attribution. In fact, this pattern was also

found in studies conducted decades ago (Rest et al., 1973; Weiner & Kukla, 1970). While the relative reward-worthiness is preserved within the same outcome condition (i.e., agents who were more willing to exert effort were still rewarded more), a failure signals that the teammates likely need to contribute more to succeed in the future, so motivating teammates becomes especially important. A closer look at Fig. 2 ("Exp 1a Bonus" and "Exp 1b Bonus" columns) reveals that bonus allocations align with model predictions in magnitude when agents succeed (points fall along the dashed line), whereas when agents fail, participants' responses are smaller than model predictions (points fall under the dashed line). Because bonuses cannot be negative, lesser bonuses may have served as a form of punishment for failed collaborations. If the goal of bonus allocation is to increase chances of success in the future by motivating collaborators to contribute more, then punishing failures and signaling a preference for willingness to exert effort seem to work in the same direction. By contrast, the outcome of a collaboration might not matter as much for a retrospective judgment such as responsibility attributions aiming at understanding who caused the outcome.

An open question for future research is how closely these findings resemble the way people reward collaborators in real-world situations. Our approach allowed us to examine lay intuitions about who deserves a reward in finely controlled experiments where people's competence, effort, and force were directly manipulated and observed by participants. However, information about someone's competence and effort is often difficult to tease apart in real-world organizations, and that could be a reason why there exist formal incentive schemes that reward output (e.g., piece-rate pay) in lieu of effort. Additionally, real-world situations might bring challenges not considered in our setup, and people might subsequently deviate from rewarding effort. For instance, they may reward their in-group more than out-group (Brewer & Silver, 1978; Vaughan et al., 1981), adjust their reward allocations to restore equity in a dyad (Leventhal et al., 1969), or base their rewards on other cues when information about effort is less reliable (e.g., someone might lie about how hard they tried) or when effort judgments are biased by the context of the task (Ibbotson et al., 2019) and reward magnitude (Rollwage et al., 2020). Future work is needed to understand how well laypeople's intuitions and the current theory apply to these naturalistic settings. Another open question is whether rewarding collaborators based on their willingness to exert effort *actually* motivates them to contribute more in the future. Because participants in our experiments only played the role of handing out bonuses, it is unclear if people receiving the bonuses interpret them the same way as givers would expect.

Finally, although we focused solely on physical tasks, it is reasonable to think that our findings generalize to situations involving cognitive effort. Past work has shown that people's intuitive theories of competence and effort are similar across physical tasks (such as box-lifting) and cognitive tasks (such as solving math problems) (Xiang et al., 2024). Further investigation is needed to confirm that the same findings hold in cognitive tasks.

In summary, we showed that participants reward collaborators based on their willingness to exert effort, rather than their responsibility for the outcome of the collaboration. This suggests that rewarding partners and judging their responsibility are likely driven by different computations. We propose that responsibility attributions evoke retrospective judgments to understand the past, whereas bonus allocations entertain prospective judgments of how can we motivate collaborators in the future. In doing so, the current work sheds light on how we understand and formalize the cognitive capacities that underlie collaboration. Future work is needed to formally test this theory, understand more generally what ought to influence retrospective versus prospective representations, and which sorts of judgments (in addition to responsibility and bonus) ought to reflect the kind of representations evoked.

Author note

Our data and code are publicly available at https://github.com/yyyxiang/bonus_allocation.

A preprint of the manuscript has been posted on PsyArXiv: https://osf.io/preprints/psyarxiv/7vtex.

Open practices

Our data and code are publicly available at https://github.com/yyyxiang/bonus_allocation. Experiments 1a and 1b were pre-registered at https://aspredicted.org/D82_M76. Experiments 2a and 2b were pre-registered at https://aspredicted.org/LTD_XPK.

CRediT authorship contribution statement

Yang Xiang: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jenna Landy:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Fiery A. Cushman:** Writing – review & editing, Validation, Supervision, Conceptualization. **Natalia Vélez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization. **Samuel J. Gershman:** Writing – review & editing, Writing – original draft, Validation, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Joshua Greene for helpful discussions. This research was supported by the Center for Brains, Minds, and Machines (CBMM), funded by an NSF STC award (award number CCF-1231216 to S.J.G., and an NIMH K00 award (award K00MH125856) to N.V.

References

Agarwal, N. C. (1998). Reward systems: Emerging trends and issues. *Canadian Psychology/Psychologie Canadienne, 39*(1–2), 60.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences, 24*(9), 694–703.

Baumard, N., Mascaro, O., & Chevallier, C. (2012). Preschoolers are able to take merit into account when distributing goods. *Developmental Psychology, 48*(2), 492.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General, 145*(12), 1654–1669.

Brewer, M. B., & Silver, M. (1978). Ingroup bias a function of task characteristics. *European Journal of Social Psychology, 8*(3).

Brown, C. (1990). Firms' choice of method of pay. *ILR Review, 43*(3), 165–S.

Bun, M. J., & Huberts, L. C. (2018). The impact of higher fixed pay and lower bonuses on productivity. *Journal of Labor Research, 39*, 1–21.

Celniker, J. B., Gregory, A., Koo, H. J., Piff, P. K., Ditto, P. H., & Shariff, A. F. (2023). The moralization of effort. *Journal of Experimental Psychology: General, 152*(1), 60.

Chung, D. J., Steenburgh, T., & Sudhir, K. (2014). Do bonuses enhance sales productivity? A dynamic structural analysis of bonus-based compensation plans. *Marketing Science, 33*(2), 165–187.

Deversi, M., & Spantig, L. (2023). *Incentive and signaling effects of bonus payments: An experiment in a company*. Technical report. CESifo.

Gerstenberg, T. (2022). What would have happened? Counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B, 377*(1866), Article 20210339.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2012). Noisy newtons: Unifying process and dependency accounts of causal attribution. In *, Vol. 34. Proceedings of the annual meeting of the cognitive science society*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *, Vol. 37. Proceedings of the annual meeting of the cognitive science society* (pp. 782–787).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review, 128*(5), 936.
Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115*(1), 166–171.
Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731–1744.
Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition, 216*, Article 104842.
Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 359* (1451), 1775.
Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18* (9), 1233–1235.
Hamann, K., Bender, J., & Tomasello, M. (2014). Meritocratic sharing is based on collaboration in 3-year-olds. *Developmental Psychology, 50*(1), 121.
Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc.
Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior, 30*(4), 244–260.
Henrich, J., & Boyd, R. (2016). How evolved psychological mechanisms empower cultural group selection. *Behavioral and Brain Sciences, 39*.
Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs, 39*(4), 632–657.
Ibbotson, P., Hauert, C., & Walker, R. (2019). Effort perception is made more accurate with more effort and when cooperating with slackers. *Scientific Reports, 9*(1), Article 17491.
Jara-Ettinger, J., Kim, N., Muetener, P., & Schulz, L. (2014). Running to do evil: Costs incurred by perpetrators affect moral judgment. In , *Vol. 36. Proceedings of the annual meeting of the cognitive science society* (pp. 684–688).
Joseph, K., & Kalwani, M. U. (1998). The role of bonus pay in salesforce compensation plans. *Industrial Marketing Management, 27*(2), 147–159.
Kanngiesser, P., & Warneken, F. (2012). Young children consider merit when sharing resources with others. *PLoS One, 7*.
Kishore, S., Rao, R. S., Narasimhan, O., & John, G. (2013). Bonuses versus commissions: A field study. *Journal of Marketing Research, 50*(3), 317–333.
Lacey, N. (1988). *State punishment: Political principles and community values*.
Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology, 129*, Article 101412.
Lanzetta, J. T., & Hannah, T. (1969). Reinforcing behavior of "naive" trainers. *Journal of Personality and Social Psychology, 11*(3), 245.
Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review, 90* (5), 1346–1361.
Leventhal, G. S., Weiss, T., & Long, G. (1969). Equity, reciprocity, and reallocating rewards in the dyad. *Journal of Personality and Social Psychology, 13*(4), 300.
Lewis, D. (2000). Causation as influence. *The Journal of Philosophy, 97*(4), 182–197.
Lin, H., Westbrook, A., Fan, F., & Inzlicht, M. (2021). *An experimental manipulation increases the value of effort*. Nature Human Behaviour.
Milkovich, G. T., & Wigdor, A. K. (1991). *Pay for performance: Evaluating performance appraisal and merit pay*. National Academy Press.
Miller, C. E., & Komorita, S. S. (1995). Reward allocation in task-performing groups. *Journal of Personality and Social Psychology, 69*(1), 80.
Powell, M. J., et al. (2009). *The bobyqa algorithm for bound constrained optimization without derivatives*. Cambridge NA report NA2009/06 (Vol. 26, pp. 26–46). Cambridge: University of Cambridge.
Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society Open Science, 3*(11), Article 160510.
Ramos, A. (2022). *A comparison of fixed pay, piece-rate pay, and Bonus pay when performers receive tiered goals*. Western Michigan University.
Rest, S., Nierenberg, R., Weiner, B., & Heckhausen, H. (1973). Further evidence concerning the effects of perceptions of effort and ability on achievement evaluation. *Journal of Personality and Social Psychology, 28*(2), 187–191.
Rollwage, M., Pannach, F., Stinson, C., Toelch, U., Kagan, I., & Pooresmaeili, A. (2020). Judgments of effort exerted by others are influenced by received rewards. *Scientific Reports, 10*(1), 1868.
Sanna, L. J., & Turley, K. J. (1996). Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin, 22*(9), 906–919.
Schäfer, M., Haun, D. B., & Tomasello, M. (2023). Children's consideration of collaboration and merit when making sharing decisions in private. *Journal of Experimental Child Psychology, 228*, Article 105609.
Schaffer, J. (2005). Contrastive causation. *The Philosophical Review, 114*(3), 327–358.
Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science, 13* (3), 238.
Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, 177–184.
Van Herpen, M., Van Praag, M., & Cools, K. (2005). The effects of performance measurement and compensation on motivation: An empirical study. *De Economist, 153*, 303–329.
Vaughan, G. M., Tajfel, H., & Williams, J. (1981). Bias in reward allocation in an intergroup and an interpersonal context. *Social Psychology Quarterly*, 37–42.
Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of Educational Research, 42*(2), 203–215.
Weiner, B. (1993). On sin versus sickness: A theory of perceived responsibility and social motivation. *American Psychologist, 48*(9), 957.
Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology, 15*(1), 1.
Woodward, J. (2011). Mechanisms revisited. *Synthese, 183*, 409–427.
Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023a). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition, 241*, Article 105609.
Xiang, Y., Vélez, N., & Gershman, S. J. (2023b). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology. General, 152*(6), 1565–1579.
Xiang, Y., Vélez, N., & Gershman, S. J. (2024). Optimizing competence in the service of collaboration. *Cognitive Psychology, 150*, Article 101653.
Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Causality and counterfactuals in group attributions. *Cognition, 125*(3), 429–440.