Cognition 241 (2023) 105609

Contents lists available at ScienceDirect

Cognition

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

Original articles

Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks

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

[a] Department of Psychology, Harvard University, United States of America
[b] College of Human Ecology, Cornell University, United States of America
[c] Center for Brain Science, Harvard University, United States of America
[d] Center for Brains, Minds, and Machines, MIT, United States of America

ARTICLE INFO

Dataset link: https://github.com/yyyxiang/responsibility_attribution

Keywords:
Responsibility attribution
Moral psychology
Causal reasoning
Social cognition

ABSTRACT

How do people judge responsibility in collaborative tasks? Past work has proposed a number of metrics that people may use to attribute blame and credit to others, such as effort, competence, and force. Some theories consider only the actual effort or force (individuals are more responsible if they put forth more effort or force), whereas others consider counterfactuals (individuals are more responsible if some alternative behavior on their or their collaborator's part could have altered the outcome). Across four experiments (N = 717), we found that participants' judgments are best described by a model that considers both actual and counterfactual effort. This finding generalized to an independent validation data set (N = 99). Our results thus support a dual-factor theory of responsibility attribution in collaborative tasks.

1. Introduction

When humans collaborate, we often face the problem of apportioning responsibility for the outcome of the collaboration. When a scientific team makes a new discovery, who gets the credit? When a new product tanks, who in the company shoulders the blame? Responsibility attributions are not only important in splitting the spoils of past collaborations; they also help teams better adapt to new ones, by informing decisions about whom to recruit for a future collaboration, or about how to structure teams to increase the chances of success in the future. However, compared to judging an *individual's* responsibility for their own, isolated actions, responsibility attributions in *collaborative* settings pose a unique challenge: When multiple agents bring about a single outcome, how do we judge the responsibility of each agent?

One simple solution to this problem would be to simply share blame and credit evenly among all collaborators. However, responsibility attributions in collaborative tasks are often uneven. For example, suppose that you and your friend are trying to lift a couch up a flight of stairs. You both grab the underside of the couch and try to pull the couch up. You are not very strong, so you decide to put in an all-out effort and pull with all your might. Your friend gives the couch a gentle pull, but she is a champion powerlifter, so even a gentle effort from her outstrips the force you have applied. The couch refuses to budge. Even though you *actually* contributed less, your friend may receive a greater share of the blame. Because your friend is stronger and put in less effort, she *could* have contributed more and potentially changed the outcome. Indeed, past work has found that young children give a greater share of the rewards of successful collaborations to collaborators who did more (Schäfer et al., 2023), and adults give a greater share of the blame for failed collaborations to collaborators who could have changed the outcome (Allen et al., 2015).

Currently the literature offers several conflicting accounts of the precise computations that drive these intuitions. These various accounts agree, however, on one central premise: Responsibility judgments are deeply tied to the process of causal attribution (Weiner, 1995). People hold others responsible for outcomes they cause, and exculpate them from responsibility otherwise. However, past accounts have proposed a number of metrics that people may use to attribute causation. They largely fall under two styles of reasoning: production style and counterfactual style (Hall, 2004).

Classical production-style theories attribute causation based on the force that one object or agent exerts on another. Dowe (2000) described causal interactions as exchanges of conserved quantities, such as force passed from one object or agent to another. Following Talmy's (1988) analysis of linguistic notions of causation, Wolff (2007) developed a *force dynamics model*, characterizing causation as a pattern of forces and a position vector. According to this account, the representation of causal events consists of the magnitude and direction of the forces of a patient and an affector. An affector *causes* a patient to approach an end

* Correspondence to: 52 Oxford St., Cambridge MA 02138, United States of America.
E-mail address: yyx@g.harvard.edu (Y. Xiang).

https://doi.org/10.1016/j.cognition.2023.105609
Received 8 March 2023; Received in revised form 14 July 2023; Accepted 28 August 2023
Available online 12 September 2023

state only if the patient lacks the tendency to approach the end state but ends up doing so under the impact of the affector. For example, if a woman walks towards a man unwillingly only because a police officer directs her to do so, this scene would be interpreted as "the officer caused the woman to walk to the man". Causal judgments are thus made by combining the forces produced by agents, making force a possible metric for responsibility attributions. Production-based models have also been applied to moral judgments (Greene et al., 2009; Nagel & Waldman, 2012; Waldmann & Dieterich, 2007).

While production-style models typically involve a chain of events where some quantity is transmitted from cause to effect, they are just one canonical, well-studied example of a broader family of models that track agents' actual contributions to an event. Here we use the term *actual-contribution models* to refer to a broader class of theories that hold an agent responsible for an event based on *actual properties* of that agent (as opposed to counterfactual ones). Our category of actual-contribution models is thus a) broader in the sense that they encompass more than force, and b) apply to judgments of responsibility rather than causation. Specifically, in addition to force, we also consider the *competence* of agents (i.e., the maximum amount of force they can generate) and the *effort* actually exerted (i.e., the proportion of competence exerted—that is, force divided by competence). Related to our notion of competence, Gerstenberg et al. (2011) found that responsibility judgments are closely related to agents' skill levels: Skilled players receive more blame for losses than unskilled players, but receive similar credit for wins compared to unskilled players. And, related to our notion of effort, some studies argue that moral judgments are based on the degree of effort exerted in performing the act: Greater effort in performing immoral acts would lead to more blame, whereas greater effort in performing moral acts would lead to more credit (Bigman & Tamir, 2016; Jara-Ettinger et al., 2014). Individuals also tend to be punished more if they fail for lack of effort, rather than lack of ability (Weiner, 1993).

In contrast, counterfactual-style accounts propose that causal judgments are made by considering whether the outcome would be different in an alternative world where the agents had acted differently. An agent bears responsibility to the extent that their actions were necessary to the outcome. This form of counterfactual dependence is especially important for causal attributions in moral judgments (Lombrozo, 2010). Chockler and Halpern (2004) proposed a model of responsibility which assigns responsibility based on the minimal number of changes that have to be made to obtain a contingency where an outcome depends on an event. This model has since been supported by experiments that manipulated the contributions of different agents to a group outcome (Gerstenberg & Lagnado, 2010; Zultan et al., 2012).

In this paper, we explore how reasoning about agents' actual contribution and counterfactual contribution contribute to responsibility judgments in a collaborative box-lifting task. Across five experiments, we compared participants' judgments to seven models: three actual-contribution models (which map agents' actual force, strength, or effort directly onto responsibility judgments), three counterfactual-contribution models (which base their judgments on different effort counterfactuals), and an ensemble model that combines aspects of the winning actual- and counterfactual-contribution models. Here, force is operationalized as the physical force exerted by each agent on the box; competence, or strength, is operationalized as the maximum force an agent could exert; and effort is operationalized as the proportion of strength exerted (i.e., force divided by strength). Informed by these models, we designed a range of scenarios involving different levels of competence, effort, force, and box weight. Below, we describe our theoretical framework in more detail.

2. Theoretical framework

In the experiments below, participants viewed vignettes where pairs of agents attempted to lift a box together, and participants apportioned credit (when the lift was successful) or blame (when it failed) to individual agents. Each box has some 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 chooses how much force to exert ($F_a \in [0, S_a]$); the subjective effort needed to produce this force is given by the proportion $E_a = \frac{F_a}{S_a}$. Finally, the two agents successfully lift the box if their combined force is greater than or equal to the weight of the box. In other words, the outcome is determined by $L = \mathbb{I}[\sum_a F_a \geq W]$, where $\mathbb{I}$ is an indicator function; $L = 1$ indicates that the lift was successful ($L = 0$ otherwise). We use R_a to denote the responsibility (blame or credit) assigned to each agent after the lift attempt.

Table 1
Summary of the models. The best-fitting model in each reasoning style is in boldface.

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

Below we describe how each model computes responsibility based on a description of the lifting event. Table 1 summarizes the models, organized by reasoning style. At the end of the Theoretical framework section, we apply the models to a concrete example to illustrate how the models compute responsibility judgments.

2.1. Actual-contribution models

- **Force model**. The force (F) model judges responsibility based on how much force an agent generates in the event. Agents who exert more force are credited more for successes and blamed less for failures:

$$R_a^F \propto \begin{cases} F_a & \text{if } L = 1 \\ 10 - F_a & \text{if } L = 0 \end{cases} \quad (1)$$

- **Strength model**. The strength (S) model considers agents' strength as the metric for responsibility judgments. Gerstenberg et al. (2011) found that stronger agents are blamed more for failures. Although they did not find the same for credit assignment, intuitively, in a collaborative setting where each person's contribution is not directly observable, stronger people are blamed more for failures and credited more for successes. Imagine a scenario where an adult and a toddler lift a heavy box together. Their force and effort are unknown, although greater force on the part of the grownup might be inferred due to their superior strength. It would be natural to attribute the success mostly to the adult, rather than to the toddler. Following these intuitions, we design our Strength model in such a way that stronger agents are credited more for successes and blamed more for failures:

$$R_a^S \propto S_a \quad (2)$$

- **Effort model**. The effort (E) model attributes blame and credit based on the level of effort an agent exerts. Agents who exert more effort are credited more for successes and blamed less for failures:

$$R_a^E \propto \begin{cases} E_a & \text{if } L = 1 \\ 1 - E_a & \text{if } L = 0 \end{cases} \quad (3)$$

2.2. Counterfactual-contribution models

We adapt Icard et al.'s (2017) quantitative measure of causal strength to construct our counterfactual-contribution models. The key point of their theory is that people sample counterfactuals when they are making causal judgments, and the causal strength is a combination of actual necessity (whether changes in the focal agent could change the outcome) and robust sufficiency (whether changes in background conditions could change the outcome). When applied to our setup, the focal agent refers to the agent people are judging, and background conditions refer to the non-focal agent (the other agent).

Our counterfactual-contribution models consider whether the outcome would have been different if one or more agents had acted differently. According to Kahneman and Miller's (1986) norm theory, people are more likely to simulate modifications of variables that are more *mutable* (i.e., variables that could plausibly have been different). Similarly, Girotto et al. (1991) found that people tend to mutate controllable events. In our setup, effort is more controllable and mutable than strength—it is more plausible that agents could have exerted more or less effort, compared to agents having more or less strength (at least on a short timescale). The key implication is that agents should be credited or blamed more based on how much effort they could have exerted. This idea is broadly consistent with the account of negligence developed by Sarin and Cushman (2022), who argued that people punish negligence when others could have exerted more mental effort (e.g., bringing to mind information useful for avoiding important risks).

In light of these points, our counterfactual-contribution models consider whether a counterfactual effort allocation (denoted as E') could have altered the outcome. If agents fail, we consider whether they could have done more to change the outcome (upwards counterfactuals); if they succeed, we consider whether doing less would have changed the outcome (downwards counterfactuals). These assumptions agree with studies showing that people engage in upward counterfactual thinking after failures and downward counterfactual thinking after unexpected successes (Sanna & Turley, 1996).

Each agent's probability of changing the outcome is defined as:

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

where $/a$ indexes the agent other than agent a. For simplicity, we assume counterfactual efforts (E'_a) are drawn from discrete uniform distributions in increments of 0.01, ranging between 0 and E_a when $L = 1$, and between E_a and 1 when $L = 0$.

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

  $$R_a^{FA} \propto P_a \tag{5}$$

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

  $$R_a^{NFA} \propto 1 - P_{/a} \tag{6}$$

- **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. For simplicity, we assign equal weights to the two:

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

  Intuitively, this model assigns more responsibility 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.

2.3. Ensemble model

The Ensemble (EBA) model combines the best-fitting actual- and counterfactual-contribution models. We designed this model to explore the possibility that people employed two styles of reasoning simultaneously. To foreshadow our results, we found that the Effort model and the Both-agent counterfactual model provided the best fit to the behavioral data, therefore the Ensemble model is a weighted combination of the Effort model and the Both-agent counterfactual model:

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

where $w \in [0, 1]$ is the weight parameter. When $w = 0$, the Ensemble model is essentially the Effort model, and when $w = 1$, the Ensemble model is equivalent to the Both-agent counterfactual model.

For simplicity, we assign equal weights to both models ($w = 0.5$), which results in:

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

We are not making strong claims about the weights. Instead, we care about whether a combination of the two models captures the qualitative patterns of the data better and provides a better quantitative fit than individual models, without adding free parameters to the model. The equal-weighting design is also supported by regression results showing that the two models have similar coefficients when they are used to predict participants' judgments (see the Results section of Experiments 1a and 1b). Additionally, we explored unequal-weighting models, which produced similar predictions visualized in Figure S1 in the Supplement.

2.4. Toy scenario

To better understand how the models generate predictions and compute counterfactual probabilities, imagine a toy scenario where Agent A has a strength of 8, exerted 40% effort, and produced a force of $8 \times 0.4 = 3.2$. Agent B has a strength of 8, exerted 20% effort, and produced a force of $8 \times 0.2 = 1.6$. They failed to lift a box of Weight 8 since their combined force was $3.2 + 1.6 = 4.8 < 8$. Expressed in mathematical form, that is: $F_A = 3.2, S_A = 8, E_A = 0.40, F_B = 1.6, S_B = 8, E_B = 0.20$. Suppose that we are judging Agent A's responsibility, i.e., how much blame Agent A should receive for the team's failure on a scale of 0 to 10. In other words, Agent A is the focal agent.

The actual-contribution models assign blame based on the actual properties of the agent:

- The Force (F) model predicts that Agent A's blame equals 10 minus the force she exerted, that is, $R_A^F = 10 - 3.2 = 6.8$.
- The Strength (S) model predicts that Agent A's blame equals her strength, that is, $R_A^S = 8$.
- The Effort (E) model predicts that Agent A's blame is equal to 100% minus the level of effort she exerted, then rescaled a 0–10 range, that is, $R_A^E = 6$.

The counterfactual-contribution models assign blame based on how likely counterfactual effort allocations would have changed the outcome. Intuitively: when agents fail, counterfactual models consider whether agents could have succeeded if they had put in more effort; conversely, when agents succeed, counterfactual models consider whether agents would have failed if they had put in less effort. Since agents failed in this toy scenario, we consider upward counterfactuals; for example, Agent A's counterfactual effort allocations range from 41% to 100% with increments of 1%. Since the models have a uniform prior over these upward counterfactual effort allocations, the probability of Agent A changing the outcome is computed by plugging each counterfactual effort allocation (E'_A) into the equation

$$E'_A \times S_A + E_B \times S_B \geq W, \tag{10}$$

and taking the mean. For example, if Agent A had exerted 50% effort, then $E'_A \times S_A + E_B \times S_B = 0.5 \times 8 + 0.2 \times 8 = 5.6$ which is less than 8, returning a 0 (false). If Agent A had exerted 90% effort, then $E'_A \times S_A + E_B \times S_B = 0.9 \times 8 + 0.2 \times 8 = 8.8$ which is greater than 8, returning a 1 (true). Thus, unless Agent A had exerted an effort between 80% and 100%, she would not have overturned the outcome. If we take the average of all the counterfactual effort allocations applied to Eq. (10), we get the probability of Agent A changing the outcome by altering her effort allocation: $P_A = 0.35$.

Similarly, Agent B's counterfactual effort allocations range from 21% to 100% with increments of 1%. Following the same calculations, we get the probability of Agent B changing the outcome by altering her effort allocation: $P_B = 0.51$.

The counterfactual-contribution models use P_A and P_B to make predictions of responsibility:

- The Focal-agent-only (FA) counterfactual model predicts that Agent A should receive blame that is proportional to P_A, that is, $R_A^{FA} = 3.5$.
- The Non-focal-agent-only (NFA) counterfactual model predicts that Agent A should receive blame that is proportional to $1 - P_B$, that is, $R_A^{NFA} = 4.88$.
- The Both-agent (BA) counterfactual model predicts that Agent A's blame should be the average of R_A^{FA} and R_A^{NFA}, that is, $R_A^{BA} = (3.5 + 4.88)/2 = 4.19$.

Finally, the Ensemble (EBA) model predicts that Agent A's blame should be the average of R_A^{E} and R_A^{BA}, that is, $R_A^{EBA} = (6 + 4.19)/2 = 5.09$. Note that, put together, the Ensemble model makes a more lenient responsibility judgment than a model that considers an agent's actual effort alone—though the agent did not contribute very much effort, there were relatively few effort allocations that would have changed the outcome.

3. Experiments 1a and 1b

In these experiments, participants apportioned responsibility to agents in a range of scenarios that were designed to elicit quantitatively distinct judgments from the models described above. Additionally, across the two experiments, we manipulated agents' "difference-making" ability to qualitatively test the predictions of the counterfac tual-contribution models. "Difference-making" refers to whether an agent is able to make a difference to the outcome by changing her effort allocation. In Experiment 1a, both agents are always difference-makers, whereas in Experiment 1b, only one agent is a difference-maker at a time. Difference-making is a purely counterfactual concept; thus, we would expect the data to be qualitatively different across the two experiments if people only consider whether a single agent could have acted differently to change the outcome, as predicted by the Focal-agent-only counterfactual model and Non-focal-agent-only counterfactual model. The Both-agent counterfactual model does not make this prediction.

3.1. Materials and methods

3.1.1. Participants

We recruited 180 participants for each experiment via Amazon's Mechanical Turk platform (MTurk). Participants' demographic information was not collected. To make sure that they understood the task, participants completed a comprehension check following the instructions. The comprehension check questions tested participants' knowledge of the structure of the task and agents' reward function, their understanding of strength, effort, and force, etc. We include in the Supplement a list of the comprehension check questions we used (see Comprehension check questions used in the experiments). Participants were allowed to proceed to the experiment only after they answered all the comprehension check questions correctly. Participants in Experiment 1a received $3.00 to complete 30 trials with an estimated completion time of 15 min. Participants in Experiment 1b were paid $2.50 to complete 25 trials with an estimated completion time of 12 min. To ensure data quality, participants completed two attention checks during the experiment. Participants received a warning after failing an attention check for the first time. Participants who failed only one attention check were allowed to finish the experiment, and their data were saved and not excluded from analyses. Participants who failed both attention checks were asked to leave the study before completion. The experiments were approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/MCX_M9K; we explain deviations from the preregistered analysis plan in the Supplement (see Deviations from pre-registrations).

3.1.2. Stimuli

We designed the stimuli such that in each scenario, the two agents either have the same strength, same effort, or same force (5 trials for each type). This allowed us to differentiate between the actual-contribution models: If people use any of these as their metric for judging responsibility, we would expect that the same responsibility be attributed to the two agents when they match on that dimension. For example, if people attribute responsibility based on effort, we should see the same responsibility assigned to both agents when their effort is matched.

To minimize changes across conditions and make them directly comparable, we used the same strength and effort combinations for Fail and Lift conditions. This was achieved by adjusting the box weight. The strength and effort combinations were also kept the same across Experiment 1a and Experiment 1b by modifying the box weight.

This would result in 30 trials in each experiment. However, in the Lift condition, when the two agents apply the same force, their ability to change the outcome is the same. That is to say, either both agents are difference makers, or neither one is. To give a concrete example: If both agents exerted a force of 4 and lifted a box of Weight 6, then both agents are difference makers because they would have failed if either agent withdrew their contribution. Similarly, if both agents exerted a force of 4 and lifted a box of Weight 2, then neither agent is a difference maker, since the box will still be lifted if either agent's contribution was removed. Therefore, when agents' force is matched, it is impossible to have Lift trials where one agent is a difference maker and the other is not. Consequently, it is impossible to have Lift trials where the agents' force is matched in Experiment 1b. As a result, Experiment 1a contains 30 trials (15 Fail trials, 15 Lift trials), while Experiment 1b contains 25 trials (15 Fail trials, 10 Lift trials). A complete list of the stimuli we used can be found in the Supplement (Tables S1 and S2).

3.1.3. Procedure

Fig. 1 shows the task setup. Participants read vignettes about a game show where contestants could win prizes by lifting boxes of varying weights. In each contest, new contestants of varying strengths and a new box were brought out. Participants were shown the weight of the box, the strength of each contestant, the subjective effort that each

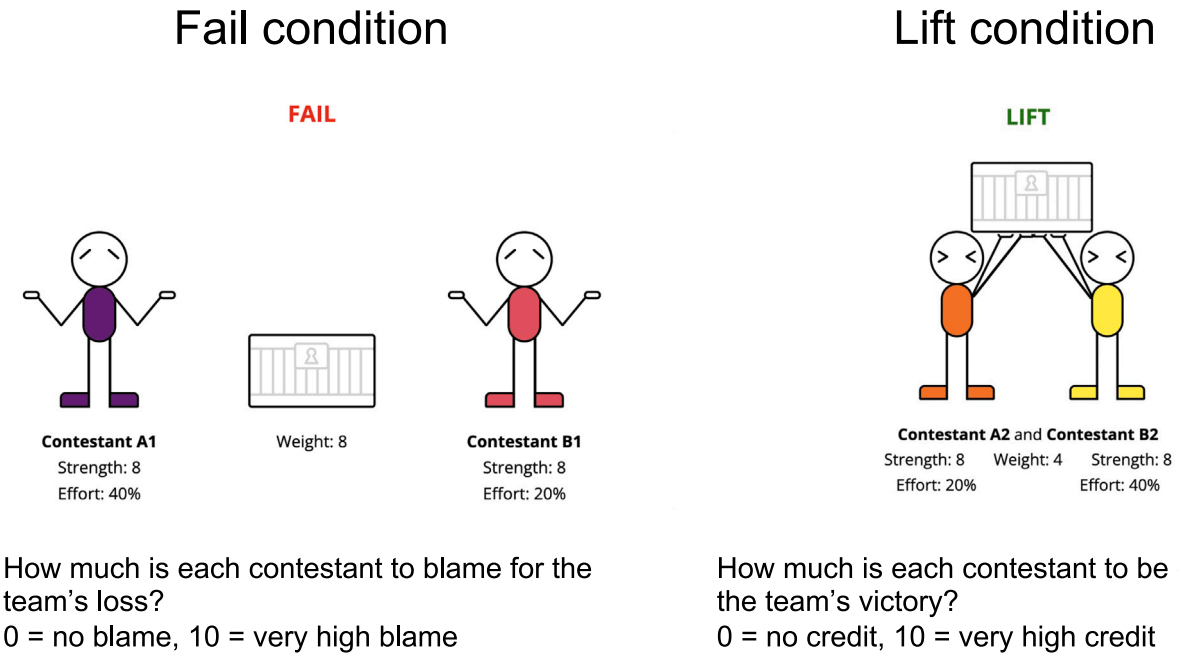

Fig. 1. Experimental setup used in Experiments 1a, 1b, and 3. Participants observed each contestant's strength and effort, the weight of the box, and whether the contestants successfully lifted the box together (Lift condition) or not (Fail condition), and they then assigned blame or credit to each contestant. In Experiments 2a and 2b, participants also observed the force exerted by each contestant.

contestant put into the lift, and the outcome of the lift (i.e., whether they succeeded or failed in lifting the box). In order to make these quantities intuitive to participants, we expressed the weight of the box and each contestant's strength using a 1 to 10 scale. For example, participants were told that a box with a weight of 1 is so light that virtually anyone can lift it, while a box of 10 is so heavy that only the strongest humans can lift it. Similarly, each contestant's strength determines the heaviest weight they can lift given an all-out effort; a contestant with a strength of 5 would fail to lift boxes with a weight of 6 or higher on their own, no matter how much effort they exerted. In order to ensure that participants understood this task setup, they first completed four training trials where they observed a single contestant attempt to lift a box alone, and they were shown how the contestant's strength and effort determine whether she succeeds or fails in lifting the box.

In the critical test trials, participants then observed multiple contests (30 contests in Experiment 1a and 25 in Experiment 1b) where two contestants applied force to a box simultaneously to try to lift it. Participants did not receive information about the contestants beyond their properties (strength, effort exerted, and force applied). This was an intentional choice with the hope that participants could base their judgments solely on the agent properties that change from trial to trial. Participants then indicated how much they thought each contestant was responsible for the outcome of the lift,i.e., how much they think each contestant is to blame for the team's loss or to be credited for the team's victory. Participants reported their judgments for each contestant separately on a scale from 0 to 10, where 0 means no blame/credit and 10 means very high blame/credit. The responses were provided using individual text boxes which allowed only entries of integers between 0 and 10. Links to the experiments can be found at https://github.com/yyyxiang/responsibility_attribution.

3.2. Results

Participants made similar judgments in Experiments 1a and 1b, as shown in the leftmost columns of Fig. 2 and Fig. 3, respectively. This is evidence against a subset of counterfactual-contribution models. Specifically, the Focal-agent-only and Non-focal-agent-only counterfactual models predict qualitatively different patterns of results across both experiments, depending on whether both agents are difference-makers or only one agent is a difference-maker; by contrast, the Both-agent counterfactual model predicts qualitatively similar results for both experiments.

We assessed which factors shaped participants' responsibility attributions by constructing a separate linear mixed-effects model for each type of trial ("Same strength", "Same effort", and "Same force"). Each regression predicted participants' responsibility judgments (i.e., blame and credit judgments) as a function of Contestant (e.g., "Less Effort" or "More Effort" in the Same Strength condition), Condition ("Lift" or "Fail"), and the interaction between Contestant and Condition, along with random effects for every regressor grouped by participants. When the contestants' strengths were matched, we observed a significant interaction between the Contestant and Condition [$t(179.0) = 31.06, p < .0001$ in Experiment 1a and $t(179.0) = 30.67, p < .0001$ in Experiment 1b; note that here and elsewhere we use the Satterthwaite approximation of the degrees of freedom]. As shown in Figs. 2 and 3, contestants who exerted more effort received less blame for failures and more credit for successes. When the contestants' efforts were matched, there was a significant interaction between the Contestant and Condition [$t(179.0) = -2.57, p < .05$ in Experiment 1a and $t(179.0) = -2.01, p < .05$ in Experiment 1b]. When both contestants exerted the same subjective effort, the stronger of the two contestants received more blame for failures and more credit for successes, with failures having slightly larger differences in contestants' responsibility. When the contestants' force levels were matched, we again saw a significant interaction between Contestant and Condition in Experiment 1a [$t(179.0) = -21.33, p < .0001$]. Fig. 2 visualizes the effect: The stronger contestant who exerted less effort received more blame for failures and less credit for successes. As explained earlier, we do not have "Same force" trials in the Lift condition of Experiment 1b. However, the trend of trials in the Fail condition is similar to that in Experiment 1a, with the stronger but lazier contestant receiving more blame for failures [$t(179.0) = 23.48, p < .0001$].

We compared participants' responsibility attributions to each of the models described above. For every computational model, we fit a linear

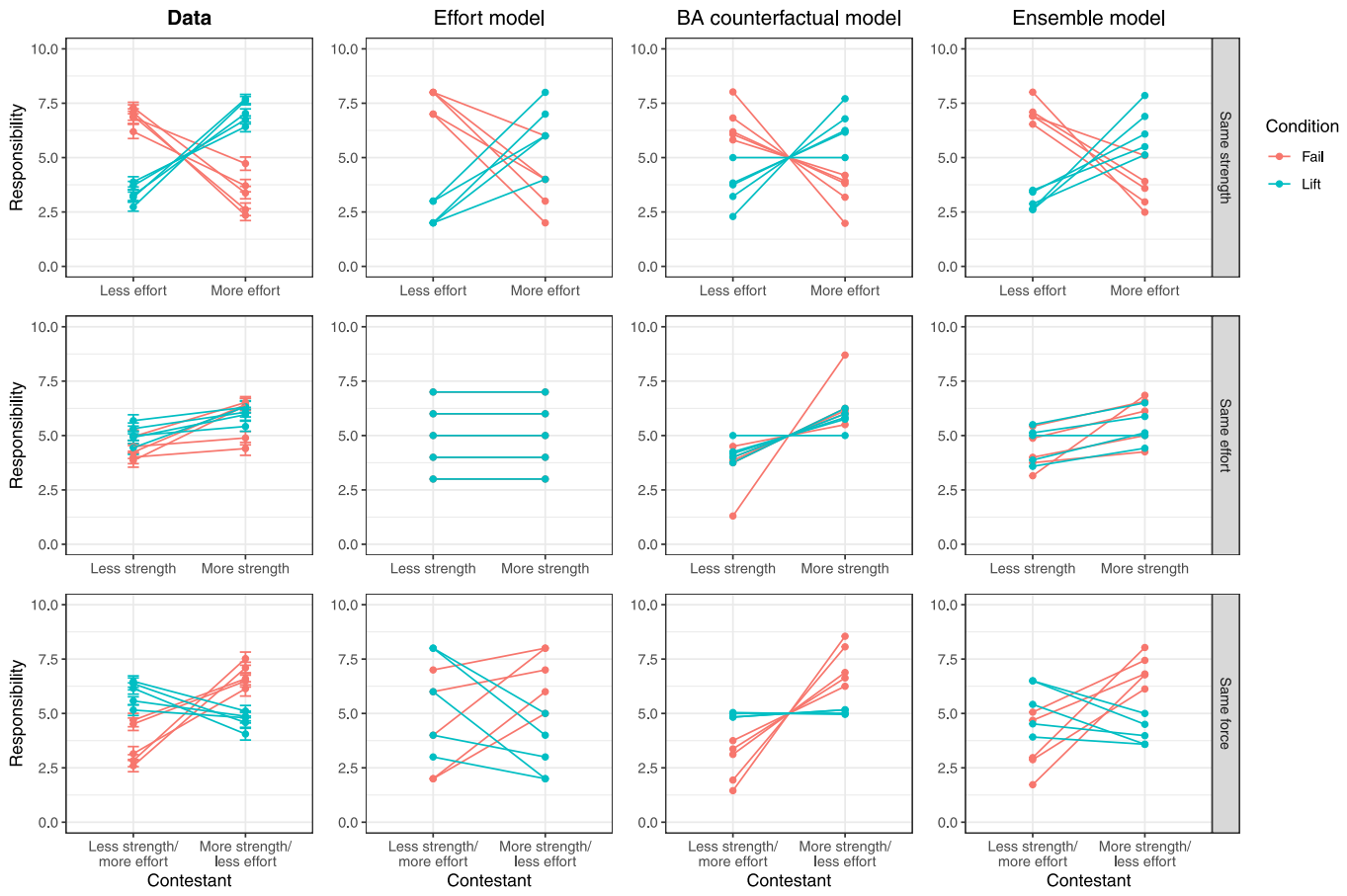

Fig. 2. Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 1a. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

mixed-effects model with participants' responsibility judgments as the response variable and the model prediction as the predictor variable, with random effects grouped by participants. We then compared models using the Bayesian Information Criterion (BIC). Lower BIC values indicate better (more probable) models. As shown in Fig. 4A, the Effort model had the lowest BIC among the three actual-contribution models and the Both-agent counterfactual model had the lowest BIC among the three counterfactual-contribution models, indicating that the Effort model and the Both-agent counterfactual model are the best at explaining the behavioral data in their respective model classes. This finding was consistent across Experiments 1a and 1b (see Figures S2 and S3 in the Supplement for a comparison between the three actual-contribution models and data, and Figures S4 and S5 in the Supplement for a comparison between the three counterfactual-contribution models and data).

To further confirm that effort, rather than force, is driving reasoning about agents' actual contributions, we fit another linear mixed-effects regression model with participants' responsibility judgments as the response variable. The predictor variables were the predictions from the Effort and Force models, along with random effects of each regressor clustered around participants. We found that the Effort model's predictions were positively correlated with participants' responsibility judgments [$t(182.2) = 29.12, p < .0001$ in Experiment 1a and $t(181.3) = 30.11, p < .0001$ in Experiment 1b], while the Force model's predictions were negatively correlated with participants' responsibility judgments [$t(182.3) = -10.89, p < .0001$ in Experiment 1a and $t(181.2) = -11.85, p < .0001$ in Experiment 1b]. This shows that effort indeed explains away the effects of force; in other words, when effort-driven responsibility attribution is accounted for, force-driven responsibility attribution fails to add explanatory value.

Figs. 2 and 3 juxtapose the Effort model and the Both-agent counterfactual model with the behavioral data. By visually comparing each of the model prediction columns against the Data column, we can see that the Effort model and the Both-agent counterfactual model each captures some aspects of the data, but neither is a fully adequate account. For example, the Effort model predicts that contestants should receive equal blame and credit when their effort is matched; however, the data show that the stronger contestant receives more responsibility [$t(179.0) = 10.25, p < .0001$ in Experiment 1a and $t(179.0) = 9.33, p < .0001$ in Experiment 1b]. Another example is that the Both-agent counterfactual model's prediction stands out in one of the "Same effort" trials, when the contestants' strengths differ greatly (one with strength 2 and one with strength 8). This suggests that responsibility judgments might be a combination of reasoning about agents' actual and counterfactual contributions. To test this hypothesis, we fit a linear mixed-effects model regressing participants' responsibility judgments on predictions of the Effort model, the Focal-agent-only counterfactual model, and the Non-focal-agent-only counterfactual model, with random effects of each regressor grouped by participants. The regression results reveal that the Effort model's predictions positively correlated with participants' responsibility judgments [$t(193.6) = 19.66, p < .0001$ in Experiment 1a and $t(197.6) = 13.56, p < .0001$ in Experiment 1b], as do the Focal-agent-only counterfactual model's predictions [$t(179.3) = 23.24, p < .0001$ in Experiment 1a and $t(221.7) = 16.54, p < .0001$ in Experiment 1b], and the Non-focal-agent-only counterfactual model's predictions [$t(192.6) = 15.69, p < .0001$ in Experiment 1a and $t(228.0) = 15.66, p < .0001$ in Experiment 1b]. This shows that predictions of the actual- and counterfactual-contribution models all made distinct contributions to predicting participants' responsibility attributions.

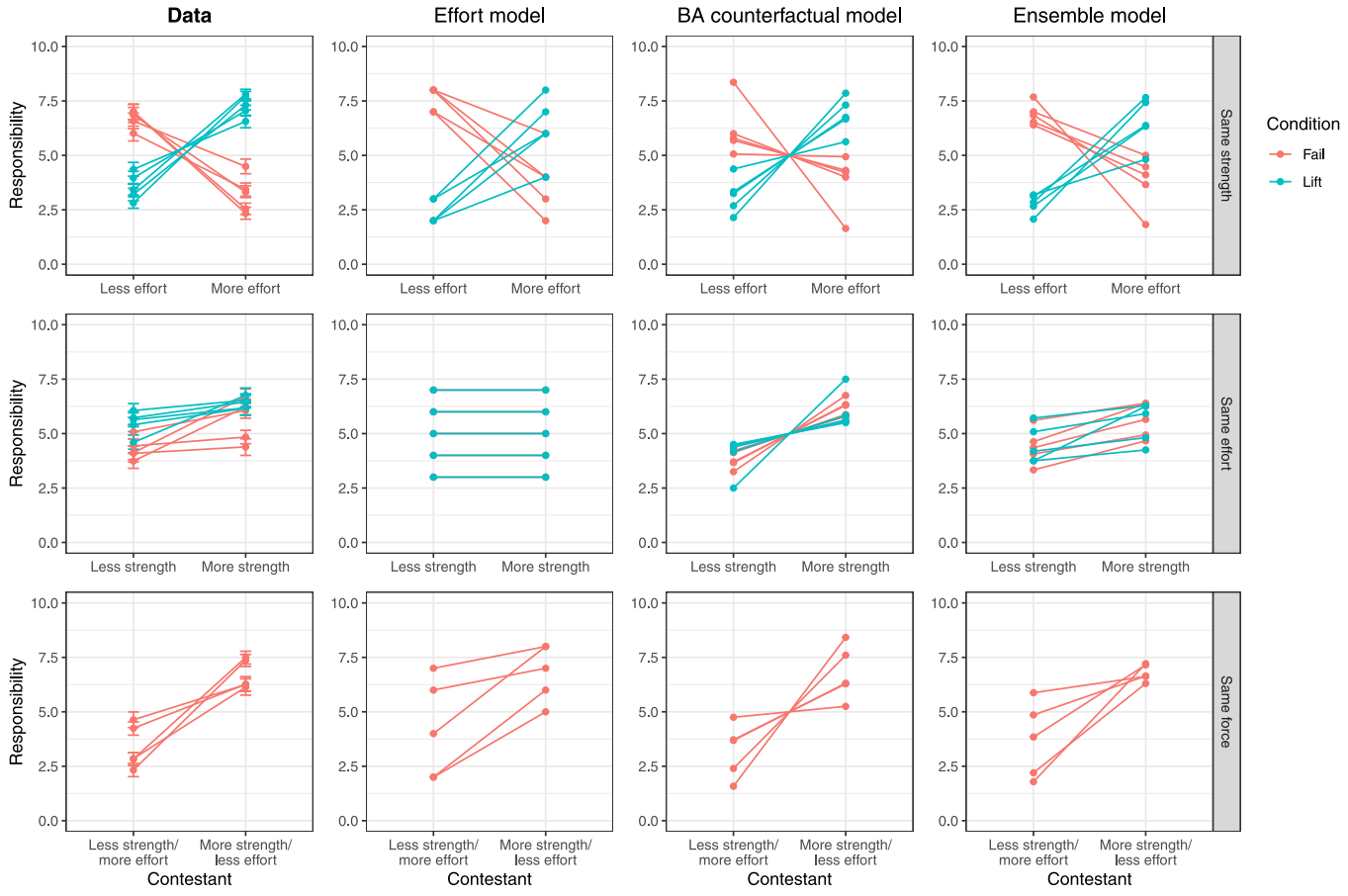

Fig. 3. Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 1b. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

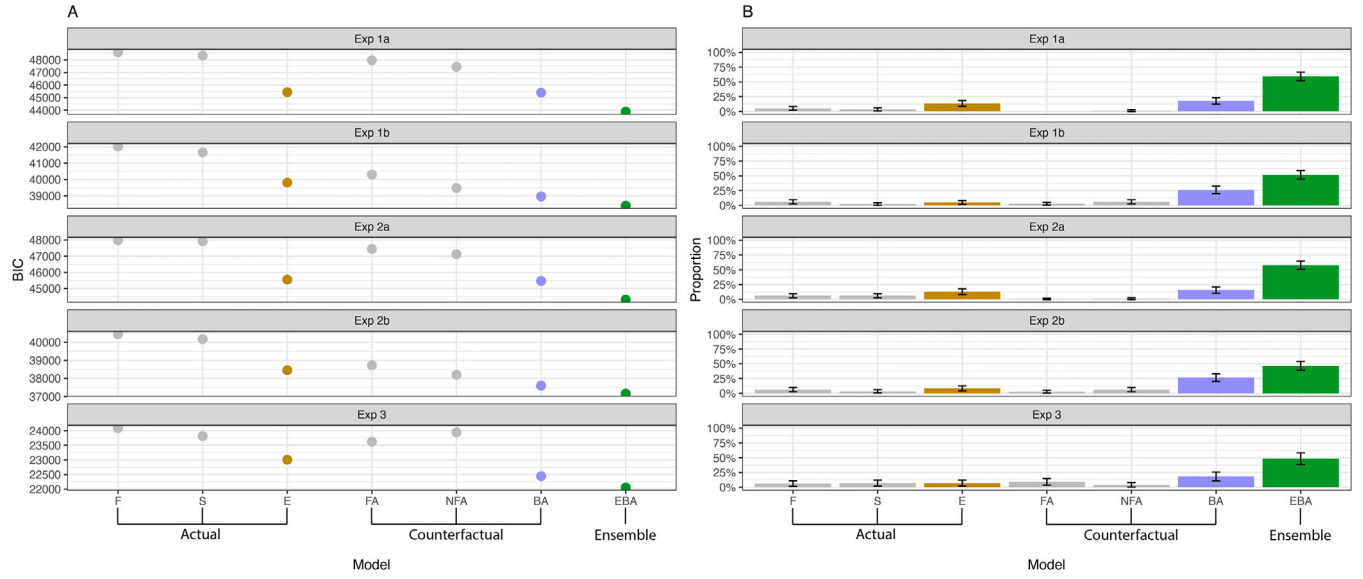

Fig. 4. (A) Bayesian information criterion (BIC) for each mixed-effects regression model. (B) Proportion of participants best described by each model. Error bars indicate 95% confidence intervals of proportions. F = Force model, S = Strength model, E = Effort model, FA = Focal-agent-only counterfactual model, NFA = Non-focal-agent-only counterfactual model, BA = Both-agent counterfactual model, EBA = Ensemble model (weighted combination of Effort model and Both-agent counterfactual model).

These results motivated us to create an Ensemble model, which is a weighted combination of the Effort model's predictions and the Both-agent counterfactual model's predictions. From a linear mixed-effects model predicting participants' judgments based on the Effort model and the Both-agent counterfactual model, with subject-level random effects, we found that both models have similar coefficients (0.38 for the Effort model and 0.47 for the Both-agent counterfactual model in Experiment 1a, and 0.31 for the Effort model and 0.59 for the Both-agent counterfactual model in Experiment 1b), corresponding to w of 0.4 and 0.3, respectively. This gave us confidence in giving equal weights to both models for simplicity (i.e., $w = 0.5$) since we are not making strong claims about the weights, as explained in the theoretical framework. We also explored unequal-weighting models, which gave similar predictions (see Figure S1), but focus on the equal-weighting Ensemble model in the main text in the interest of parsimony. The Ensemble model has the lowest BIC compared to all the other six models (see Fig. 4A). From Fig. 2 and Fig. 3, we can also see that the Ensemble model resembles the data patterns the best. Note that the counterfactual-contribution models have lower BICs on average than actual-contribution models in Experiment 1b, but not in Experiment 1a. This is perhaps due to difference-making being relevant for counterfactual-contribution models only. Even so, the Ensemble model still has the lowest BIC. Further, the Ensemble model provides a close quantitative fit to participants' judgments (Pearson's $r = 0.91, p < .0001$ in Experiment 1a, $r = 0.89, p < .0001$ in Experiment 1b), compared to the Effort model ($r = 0.77, p < .0001$ in Experiment 1a, $r = 0.73, p < .0001$ in Experiment 1b) and Both-agent counterfactual model ($r = 0.79, p < .0001$ in Experiment 1a, $r = 0.84, p < .0001$ in Experiment 1b). See Fig. 5 for a comparison between participants' judgments and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model. Figures S6 and S7 in the Supplement show comparisons between data and the three actual-contribution models and between data and the three counterfactual-contribution models, respectively.

Finally, we conducted a more detailed interrogation of the Ensemble model's success. Specifically, we asked whether there is evidence that both actual-contribution and counterfactual-contribution reasoning contribute to the model's fit to each individual participant, or whether instead some participants were best fit by an actual-contribution model alone, while others were best fit by a counterfactual-contribution model alone. In other words, is the Ensemble model favored because everybody uses both styles of reasoning, or because some people reason about agents' actual properties, and others reason about counterfactuals?

To answer this question, we ran seven linear models for each participant, each predicting responsibility judgments with one of the seven models we have. We compared the BICs of the seven models for every participant and counted the number of participants best explained by each model (indicated by the lowest model BIC). As shown in Fig. 4B, the Ensemble model explains the data of the largest number of participants (59.4% of the participants in Experiment 1a, 51.7% of the participants in Experiment 1b), whereas the Effort model explains the data of 13.3% of the participants in Experiment 1a and 5.0% of the participants in Experiment 1b, and the Both-agent counterfactual model explains the data of 17.8% of the participants in Experiment 1a and 26.1% of the participants in Experiment 1b. These results suggest that, rather than there being a mix of participants within our sample who employ different strategies, individual participants reason about both agents' actual and counterfactual contributions to make responsibility attributions.

3.3. Discussion

In Experiments 1a and 1b, we compared participants' responsibility judgments with the three actual-contribution models and the three counterfactual-contribution models. Among the three actual-contribution models, we found that the Effort model provided the best fit to empirical responsibility judgments. Further support for this claim came from regression results suggesting that effort is the major driving force behind reasoning about agents' actual properties, in contrast to past findings that support a force-based account.

Among the three counterfactual-contribution models, we found that the Both-agent counterfactual model was the best account of our data. Additional regression analyses revealed that effort, counterfactual dependency on the focal agent, and counterfactual dependency on the non-focal agent all contribute to responsibility judgments in some form.

In light of these findings, we combined the Effort model and the Both-agent counterfactual model and created an Ensemble model, which provided the overall best fit to the data, yielding the lowest BIC and the highest correlation coefficients. Our single-participant analysis further showed that most participants, individually, are best explained by the Ensemble model. We conclude from these findings that both actual-contribution and counterfactual-contribution reasoning are necessary to explain responsibility judgments for group effort tasks.

4. Experiments 2a and 2b

In Experiments 1a and 1b, we showed participants the strength and effort of each contestant, but not their force. In theory, participants could calculate the force of each contestant by multiplying strength and effort. However, this could potentially bias participants to use strength and effort information over force as the basis for their judgments. To rule out this possibility, we ran Experiments 2a and 2b, in which everything was kept the same as in Experiments 1a and 1b, except that we explicitly showed participants the level of force each contestant exerted, in addition to their strength and effort.

4.1. Materials and methods

4.1.1. Participants

We recruited 180 participants for Experiment 2a and 177 participants for Experiment 2b via Amazon's Mechanical Turk platform (MTurk). Participants' demographic information was not collected. As in Experiments 1a and 1b, participants completed a comprehension check following the instructions and two attention checks during the experiment, and participants who failed these attention checks were asked to leave the experiment early. Participants in Experiment 2a received $3.00 to complete 30 trials with an estimated completion time of 15 min. Participants in Experiment 2b were paid $2.50 to complete 25 trials with an estimated completion time of 12 min. The experiments were approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/PNZ_2HW.

4.1.2. Stimuli

The stimuli were identical to Experiments 1a and 1b.

4.1.3. Procedure

The procedure was identical to Experiments 1a and 1b, except that participants also saw how much force each contestant exerted, in addition to their strength and effort (see Figure S8 for an illustration of the experimental setup).

4.2. Results

Participants' judgments in Experiments 2a and 2b were similar (see the leftmost columns of Fig. 6 and Fig. 7, respectively), and they were both similar to Experiments 1a and 1b. We tested the same hypotheses by running the same regressions as for Experiments 1a and 1b. When the contestants' strengths were matched, we saw a significant interaction between the Contestant and Condition [$t(179.0) = 27.77, p < .0001$ in Experiment 2a and $t(176.0) = 26.66, p < .0001$ in Experiment

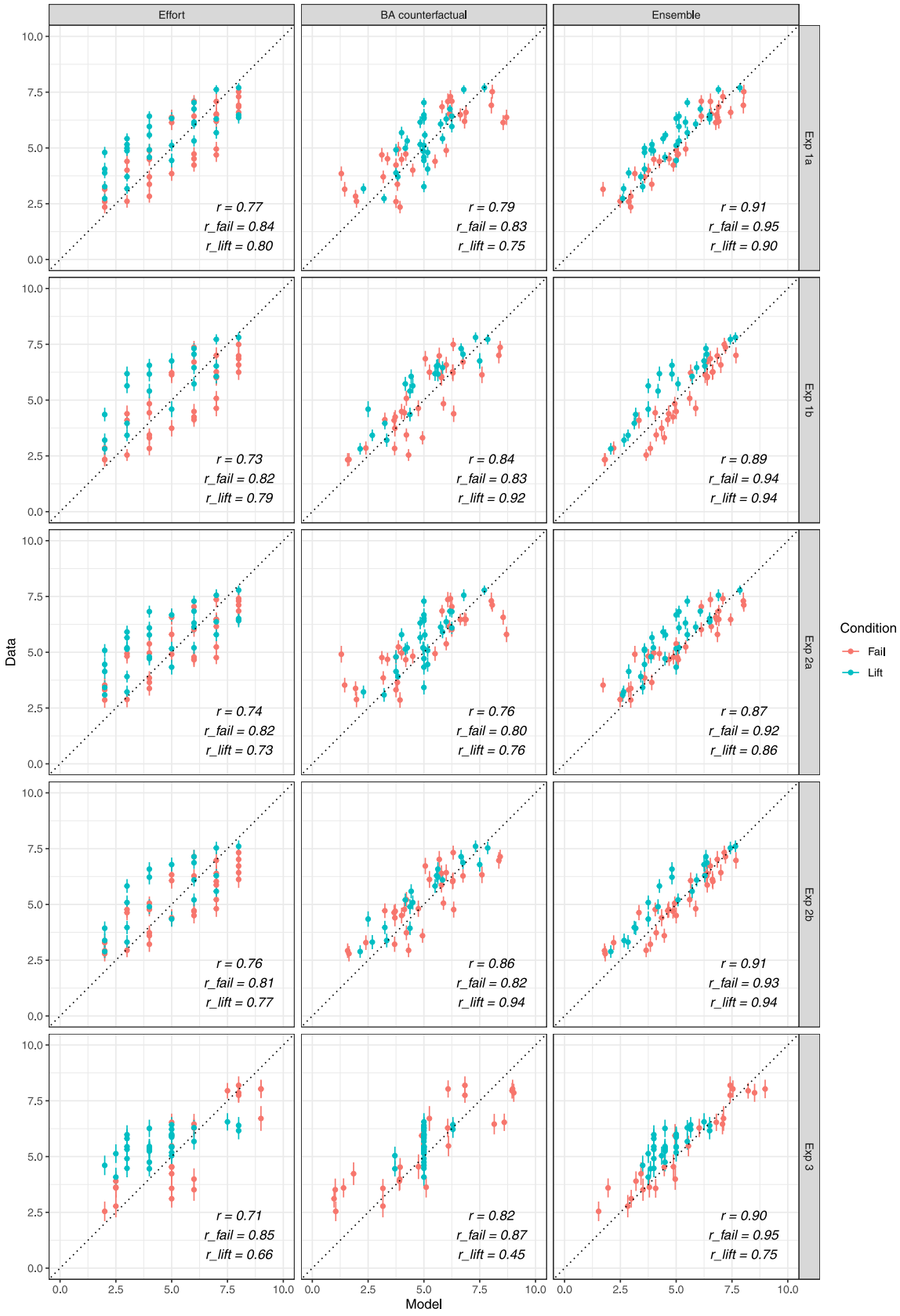

Fig. 5. Comparison between behavioral data and model predictions. Pearson correlation coefficients for data and model predictions pooled across both conditions and for each condition are shown at the bottom right of each subplot. Each dot denotes one agent in one scenario; error bars indicate 95% confidence intervals.

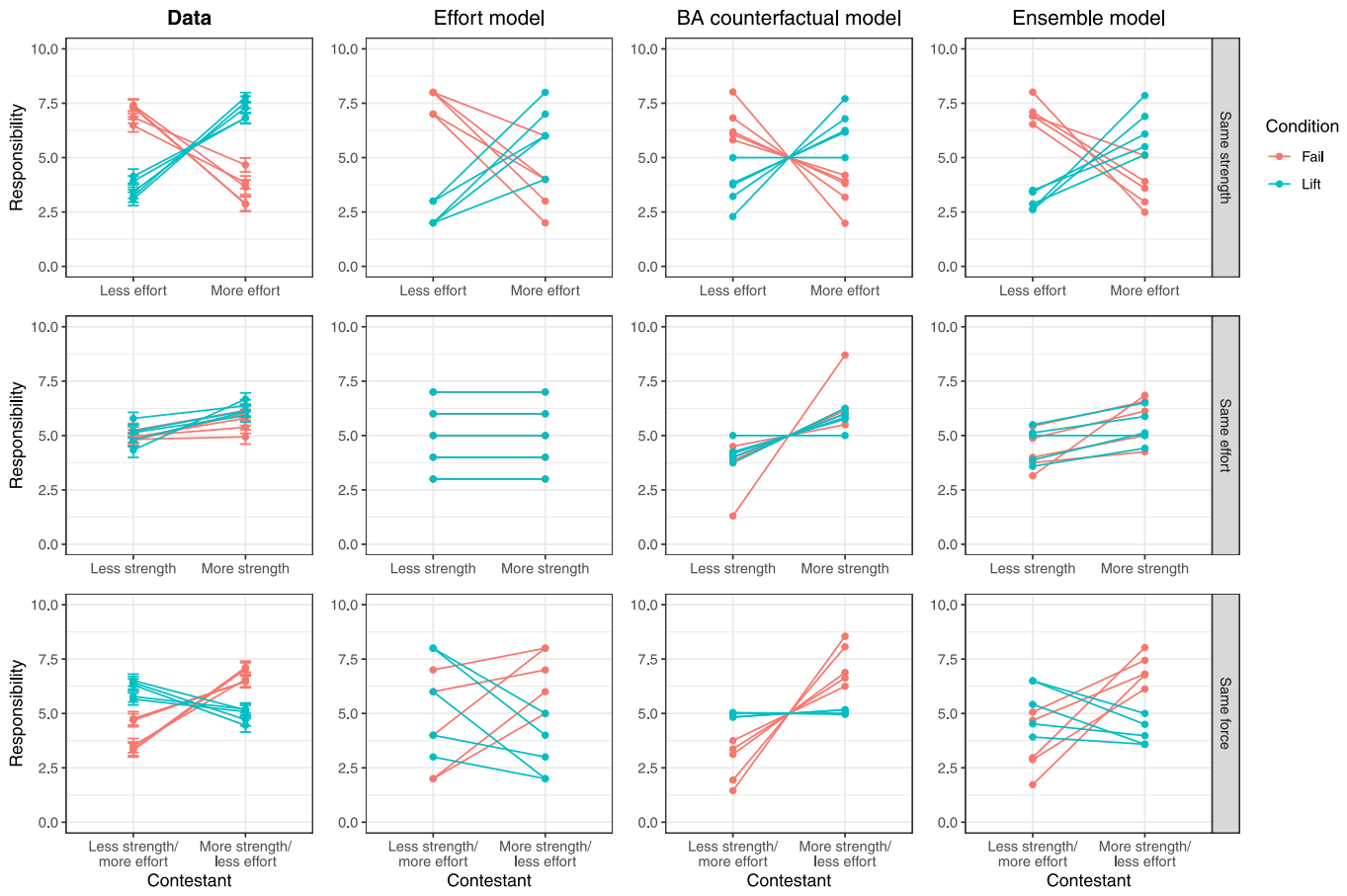

Fig. 6. Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 2a. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

2b]. As shown in Figs. 6 and 7, contestants who exerted more effort received less blame for failures and more credit for successes. When the contestants' efforts were matched, we saw a significant interaction between the Contestant and Condition in Experiment 2a [$t(179.0) = 2.58, p < .05$], but not in Experiment 2b [$t(176.0) = 1.80, p = .07$]. When both contestants exerted the same level of effort, the stronger of the two contestants received more blame for failures and more credit for successes, with successes having slightly larger differences in contestants' responsibility in Experiment 2a. When the contestants' force levels were matched, we again saw a significant interaction between Contestant and Condition in Experiment 2a [$t(179.0) = -19.38, p < .0001$]. Fig. 6 visualizes the effect: The stronger contestant who exerted less effort received more blame for failures and less credit for successes. Although we do not have "Same force" trials in the Lift condition of Experiment 2b, the trend of trials in the Fail condition again is similar to that in Experiment 2a, with the stronger but lazier contestant receiving more blame for failures [$t(176.0) = 19.91, p < .0001$].

Moving to the comparison between participants' responsibility judgments and each of the seven models, we again found that the Effort model has the lowest BIC among the three actual-contribution models and the Both-agent counterfactual model has the lowest BIC among the three counterfactual-contribution models (see Figures S9 and S10 in the Supplement for a comparison between the three actual-contribution models and data, and Figures S11 and S12 in the Supplement for a comparison between the three counterfactual-contribution models and data). However, the Ensemble model has the lowest BIC among all seven models (see Fig. 4A) and provides the best quantitative fit to participants' judgments (Pearson's $r = 0.87, p < .0001$ in Experiment 2a, $r = 0.91, p < .0001$ in Experiment 2b), compared to the Effort model ($r = 0.74, p < .0001$ in Experiment 2a, $r = 0.76, p < .0001$ in Experiment 2b) and Both-agent counterfactual model ($r = 0.76, p < .0001$ in Experiment 2a, $r = 0.86, p < .0001$ in Experiment 2b). See Fig. 5 for a comparison between participants' judgments and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model. Figures S6 and S7 in the Supplement show comparisons between data and the three actual-contribution models and between data and the three counterfactual-contribution models, respectively. As in Experiments 1a and 1b, our single-participant analysis showed that the largest number of participants were best described by the Ensemble model (57.8% of the participants in Experiment 2a, 46.3% of the participants in Experiment 2b), whereas 12.8% of the participants in Experiment 2a and 8.5% of the participants in Experiment 2b were best described by the Effort model, and 15.6% of the participants in Experiment 2a and 26.6% of the participants in Experiment 2b were best described by the Both-agent counterfactual model (see also Fig. 4B).

We also confirmed that, when Effort and Force were both used to predict responsibility judgments, the Effort model's predictions were positively correlated with participants' responsibility judgments [$t(181.8) = 26.45, p < .0001$ in Experiment 2a and $t(180.5) = 27.83, p < .0001$ in Experiment 2b], whereas predictions of the Force model were negatively correlated with participants' responsibility judgments [$t(181.8) = -7.91, p < .0001$ in Experiment 2a and $t(180.4) = -8.52, p <$

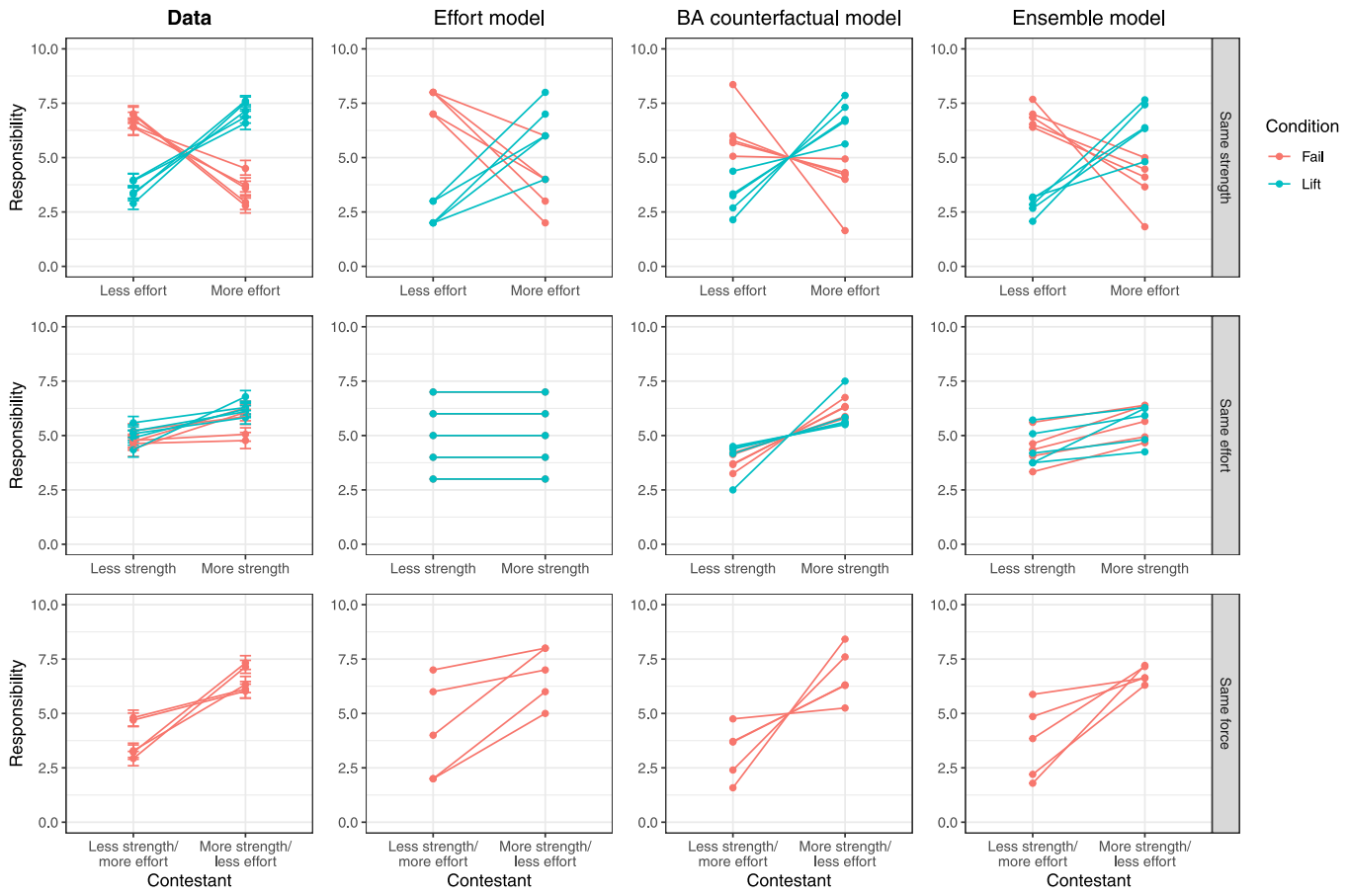

Fig. 7. Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 2b. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

.0001 in Experiment 2b]. In addition, as in Experiments 1a and 1b, we found that effort, counterfactuals of the focal agent, and counterfactuals of the non-focal agent all contributed to responsibility judgments: The Effort model's predictions positively correlated with participants' response [$t(205.5) = 18.69, p < .0001$ in Experiment 2a and $t(195.3) = 15.03, p < .0001$ in Experiment 2b], as well as the Focal-agent-only counterfactual model [$t(181.8) = 25.20, p < .0001$ in Experiment 2a and $t(188.7) = 16.02, p < .0001$ in Experiment 2b] and the Non-focal-agent-only counterfactual model [$t(207.0) = 15.82, p < .0001$ in Experiment 2a and $t(204.1) = 12.18, p < .0001$ in Experiment 2b].

4.3. Discussion

Experiments 2a and 2b were designed to level the playing field for the models by providing an explicit representation of force to participants. We successfully replicated the findings of Experiments 1a and 1b; the Effort model and the Both-agent counterfactual models were the actual- and counterfactual-contribution models, respectively, that best captured participants' judgments, and the Ensemble model – which combines the two – provided the best overall fit to the data. Moreover, analysis of individual participants confirmed that more people were best described by this combination than by any other model.

5. Experiment 3

The purpose of Experiments 1a, 1b, 2a, and 2b was to test participants' judgments on scenarios that were tailor-made to discriminate between the models. Experiment 3 was designed as a validation experiment, to test how well the models capture participants' judgments on a wider range of randomly-generated scenarios.

5.1. Materials and methods

5.1.1. Participants

We recruited 99 participants via Amazon's Mechanical Turk platform (MTurk). Participants' demographic information was not collected. As in all the experiments described above, participants completed a comprehension check following the instructions and two attention checks during the experiment, and participants who failed these attention checks were asked to leave the experiment early. Participants received $3.00 for completing the experiment with an estimated completion time of 15 min. The experiment was approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/ZC3_VG4; we explain deviations from the preregistered analysis plan in the Supplement (see Deviations from pre-registrations).

5.1.2. Stimuli

The stimuli were constructed in the same way as in the previous experiments, except that agents' strength, effort, or force were randomly sampled to generate 30 scenarios; using this method allowed us to test participants' intuitions on a wider variety of scenarios where agents were not guaranteed to be matched in their strength, effort, or force. The only restriction we had when creating the stimuli was that both contestants were difference-makers. Three out of 30 scenarios were excluded from our analysis due to errors (the box weight was greater than 10 or a contestant's strength was greater than 10, which were impossible scenarios given our setup). A list of the stimuli used in the experiment is shown in the Supplement (Table S3).

5.1.3. Procedure

The procedure was identical to Experiments 1a and 1b.

5.2. Results

We conducted the same analyses as above. Similarly, the Effort model has the lowest BIC among the three actual-contribution models and the Both-agent counterfactual model has the lowest BIC among the three counterfactual-contribution models. The Ensemble model has the lowest BIC among all seven models (see Fig. 4A) and provides the best quantitative fit to participants' judgments (Pearson's $r = 0.90, p < .0001$), compared to the Effort model ($r = 0.71, p < .0001$) and Both-agent counterfactual model ($r = 0.82, p < .0001$). See Fig. 5 for a visualization. Figures S6 and S7 in the Supplement show comparisons between data and the three actual-contribution models and between data and the three counterfactual-contribution models, respectively. Once again, the single-participant analysis revealed that more participants were best explained by the Ensemble model (48.5% of the participants) than any other model.

When Effort and Force were used to predict responsibility judgments at the same time, the Effort model's predictions were positively correlated with participants' responsibility judgments [$t(99.1) = 17.30, p < .0001$], while the Force model's predictions were negatively correlated with participants' responsibility judgments [$t(99.1) = -6.17, p < .0001$]. Also, we found that effort, counterfactuals of the focal agent, and counterfactuals of the non-focal agent all contributed to responsibility judgments: The Effort model's predictions positively correlated with participants' response [$t(101.4) = 12.45, p < .0001$], as well as the Focal-agent-only counterfactual model [$t(100.0) = 14.14, p < .0001$] and the Non-focal-agent-only counterfactual model [$t(101.7) = 11.63, p < .0001$].

5.3. Discussion

Experiment 3 was a generalization test of our theory. We did not select the models on the basis of data from Experiment 3, yet the best model from the previous experiments (the Ensemble model) predicts responsibility judgments in this experiment very accurately ($r = 0.90, p < .0001$). Thus, we find converging evidence that people combine actual-contribution and counterfactual-contribution reasoning when making responsibility judgments about group effort.

6. General discussion

Responsibility for the outcomes of collaborations is often distributed unevenly. For example, the lead author on a project may get the bulk of the credit for a scientific discovery, the head of a company may shoulder the blame for a failed product, and the lazier of two friends may get the greater share of blame for failing to lift a couch. However, past work has provided conflicting accounts of the computations that drive responsibility attributions in collaborative tasks. Here, we compared each of these accounts against human responsibility attributions in a simple collaborative task where two agents attempted to lift a box together. We contrasted seven models that predict responsibility judgments based on metrics proposed in past work, comprising three actual-contribution models (Force, Strength, Effort), three counterfactual-contribution models (Focal-agent-only, Non-focal-agent-only, Both-agent), and one Ensemble model that combines the best-fitting actual- and counterfactual-contribution models. Experiment 1a and Experiment 1b showed that the Effort model and the Both-agent counterfactual model capture the data best among the actual-contribution models and the counterfactual-contribution models, respectively. However, neither provided a fully adequate fit on their own. We then showed that predictions derived from the average of these two models (i.e., the Ensemble model) outperform all other models, suggesting that responsibility judgments are likely a combination of reasoning about agents' actual properties and counterfactual dependence. Further evidence came from analyses performed on individual participants, which revealed that the Ensemble model explained more participants' data than any other model. These findings were subsequently supported by Experiment 2a and Experiment 2b, which replicated the results when additional force information was shown to the participants, and by Experiment 3, which validated the model predictions with a broader range of stimuli.

Prior studies have largely examined how people assign responsibility for events with agent-specific outcomes, such as a darts game where the whole team wins if any player lands a bullseye (Gerstenberg et al., 2011), or a coordination game where fishers decide independently whether to catch fish or to clear the roads to go to market (Allen et al., 2015). Real-world collaborations, however, typically involve a group outcome – such as lifting a couch together – and it is usually hard to single out each individual's contribution. In our experiments, participants observe only a joint outcome – agents either succeeded or failed to lift a box together – as opposed to each agent attempting to lift the box by themselves. Our work thus complements and extends this line of research to tasks where only a single group outcome is observed and individual outcomes are not observed separately. Therefore, our work taps into the complexities common to many real-world collaborative tasks.

Our results have broader theoretical implications for theories that track agents' actual contributions. Production-style accounts of causal reasoning have often focused on *force* as the dominant metric of causation (e.g., Dowe, 2000; Talmy, 1988; Wolff, 2007). For example, if a traffic cop signals for a woman to cross the street, the cop's gesture can be understood as a social "force" that causes her to move, much as how the wind exerts a physical force on a ship's sails that causes it to glide across the water (Wolff, 2007). However, here we found ample evidence against a Force model: When effort is accounted for, the effect of force disappears, even when participants receive explicit information about how much force each agent exerted (Experiments 2a and 2b). That is to say, people consider how much effort, rather than force, agents contributed when assigning responsibility. More work needs to be done to understand the cause for this discrepancy. One possibility is that people may assign responsibility based on force in situations where the concept of "effort" is less applicable (e.g., in the case of inanimate objects), as in situations where someone is unaware of the act she is performing, or accidentally causes an outcome without the intention of doing so.

Another potential explanation for the discrepancy between our findings and prior evidence is that our stimuli differ from stimuli used to support production-style accounts in the past. We explicitly presented information about strength, effort, and force to participants. Making this information explicit enabled us to control participants' representations of agents' strength and effort and directly look at how those influence responsibility judgments. Our past work has also demonstrated that people are indeed capable of inferring agents' strength and effort based on additional information, such as observations of what they can or cannot lift by themselves given a certain incentive (Xiang et al., 2023). By contrast, in past work, these variables were either not presented or not precisely quantified. For example, if participants are shown a vignette where a police officer causes a woman to approach a man, that vignette is not guaranteed to contain information about variables that were relevant to our task, such as the cop's competence or effort. In the real world, this information is also often inferred rather than explicitly provided. Thus, it is possible that people might be more likely to use effort to make responsibility attributions when that information is available and assumed to be veridical. An open question for future work is how closely this resembles human judgments in naturalistic settings, where information about strength or effort may need to be inferred through other cues (e.g., facial expression, muscularity, past record of failures and successes) or may be less reliable (e.g., someone might lie about how hard they worked).

Beyond reasoning about agents' physical properties in lifting events, our modeling framework can be extended to capture judgments in a wide variety of non-physical tasks. Roughly, we can think of an agent's strength as the total amount of a resource available to an agent, force

as the amount of that resource an agent contributed, and effort as the amount contributed as a proportion of the agent's total resources. When viewed through this lens, many social judgments have a similar underlying structure to our task. For example, in ultimatum games (Güth et al., 1982) or instrumental learning tasks (Hackel et al., 2015), where participants' social partners donate some number of points from an endowment, the number of points donated is analogous to the agent's force, the size of the endowment is analogous to strength, and the number of points donated as a fraction of the endowment – that is, the agent's *generosity* (Hackel et al., 2015) – is analogous to effort. Similarly, in the context of elections, force is analogous to the number of votes a candidate received, strength is analogous to size of the voter population supporting a candidate, and effort is analogous to voter turnout. Moving beyond physical tasks, it remains to be seen whether our work might provide a framework to understand how agents' actual and counterfactual properties contribute to responsibility attributions in a wide range of domains.

Motivated by this analogy, we propose that one reason why effort may be particularly important in responsibility attribution is that the amount of effort exerted – rather than sheer force – indicates a person's desire to successfully complete a collaborative task. Indeed, even infants use the amount of effort expended by an agent to infer how much that agent values the outcome (Liu et al., 2017), and adults tend to judge more effortful prosocial actions as more praiseworthy (Bigman & Tamir, 2016), in part because effort signals the importance of the prosocial goal to the agent (Anderson et al., 2020). Understood this way, the combination of effort and counterfactual reasoning could be taken as support for a framework proposed in some past work that regards responsibility judgments as both about the causal role a person's action plays in bringing about an outcome and about the inferences we can make about the person's dispositions and mental states from her action (Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Uhlmann et al., 2015).

One simplification of our work is that we only entertained counterfactual models that compute the effect of counterfactual effort levels, but not the effect of counterfactual strength levels. This is a reasonable simplification under our task design, since effort is more mutable than strength, and is informed by prior evidence that effort plays a central role in moral judgments (Celniker et al., 2023). However, people may consider other variables when judging responsibility in other collaborative tasks. For example, over longer timescales, it is possible that people generate competence counterfactuals by imagining what the outcome could have been if agents had practiced or trained differently. If a slower runner causes a team to lose a relay race, people may give her a smaller share of the blame if she worked consistently to advance from a beginner runner to her current level of fitness—and a far greater share of the blame if she is a champion runner who has missed practice for months. It is also possible that people might simulate what would have happened if an agent had been replaced by someone else when replacements are available (Wu & Gerstenberg, 2023). This was unlikely in our paradigm since agents involved in each lift were not replaceable; however, one might readily imagine replacing a player with a substitute in a football game particularly if the team had lost by a small margin.

It remains an open question how people define the probability distribution over counterfactual effort levels when engaging in counterfactual reasoning. In the current work, we assumed that alternative effort allocations were drawn from discrete uniform distributions in increments of 0.01, and we only considered one-sided counterfactual effort allocations (either upward or downward, depending on the outcome). This is a simple yet plausible way to construct counterfactuals, although there certainly exist many alternatives. Past models simulate an agent as not-acting, for instance, when Billy trips over a tree root on his way rushing to stop Suzy from throwing a rock at a window, simulating whether Billy would have stopped Suzy if he had not tripped (Hall, 2004). More recent models instead sample counterfactual events from a continuous distribution; these models generate counterfactuals by simulating alternatives that are close to the actual world or are high probability under a generative model (Lucas & Kemp, 2015; Quillien & Lucas, 2023). For example, taking inspiration from noisy models of Newtonian physics (Gerstenberg et al., 2012), one could draw counterfactual effort allocations from a Gaussian distribution centered around the actual effort, such that counterfactuals closer to the actual effort are more likely to be imagined. It is worth noting that we are not making a strong claim about how counterfactual effort allocations are generated, but rather that even simple approximations such as drawing counterfactuals from a uniform distribution lend support to our theory.

More broadly, the question of how people judge responsibility in group effort tasks is critical to understanding the cognitive foundations of collaboration. Responsibility attributions may help diagnose problems when collaborations go awry and inform decisions about how to divide the spoils of collaboration and whom to recruit in the future. Responsibility attributions may also serve an important purpose in motivating collaborators to consistently apply effort in collaborative tasks, rather than loafing and benefiting from the efforts of their collaborators. Assigning responsibility for past collaborations paves a path for success in future collaborations.

CRediT authorship contribution statement

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

Data availability

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

Acknowledgments

We thank Tobias Gerstenberg, Sarah Wu, and David Rose for helpful discussions. This research was supported by the Center for Brains, Minds and Machines (CBMM) and funded by an NSF STC award (award number CCF-1231216), two Stimson research grant awards to Y.X. from the Harvard University Department of Psychology, and an NIMH K00 award (award K00MH125856) to N.V.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105609.

References

Allen, K. R., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. *Vol. 37*, In *Proceedings of the annual meeting of the cognitive science society* (pp. 84–89).

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

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

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), 60–79.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115.
Dowe, P. (2000). *Physical causation*. Cambridge University Press.
Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. *Vol. 33*, In *Proceedings of the annual meeting of the cognitive science society* (pp. 720–725).
Gerstenberg, T., Goodman, N., & Lagnado, D. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. *Vol. 34*, In *Proceedings of the annual meeting of the cognitive science society* (pp. 378–383).
Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.
Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122–141.
Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1–3), 111–133.
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.
Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388.
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.
Hall, N. (2004). Two concepts of causation. *Causation and Counterfactuals*, 225–276.
Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.
Jara-Ettinger, J., Kim, N., Muetener, P., & Schulz, L. (2014). Running to do evil: Costs incurred by perpetrators affect moral judgment. *Vol. 36*, In *Proceedings of the annual meeting of the cognitive science society* (pp. 684–688).
Kahneman, D., & Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psychol. Rev.*, *93*(2), 136–153.
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.
Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.
Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, *61*(4), 303–332.
Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700.
Nagel, J., & Waldman, M. (2012). Force dynamics as a basis for moral intuitions. *Vol. 34*, In *Proceedings of the annual meeting of the cognitive science society* (pp. 785–790).
Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.
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.
Sarin, A., & Cushman, F. (2022). One thought too few: Why we punish negligence. PsyArXiv.
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 (New York, NY)*, *228*, Article 105609.
Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, Article 104890.
Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.
Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.
Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253.
Weiner, B. (1993). On sin versus sickness: A theory of perceived responsibility and social motivation. *American Psychologist*, *48*(9), 957–965.
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. guilford Press.
Wolff, P. (2007). Representing causation.. *Journal of Experimental Psychology: General (Washington, DC)*, *136*(1), 82–111.
Wu, S. A., & Gerstenberg, T. (2023). If not me, then who? Responsibility and replacement. PsyArXiv.
Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565–1579.
Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: causality and counterfactuals in group attributions. *Cognition*, *125*(3), 429–440.