HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Psychology

have examined a dissertation entitled

Cognitive Foundations of Collaboration

presented by Yang Xiang

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Samuel Gershman

Signature _____

Typed name: Prof. Fiery Cushman

Signature _____

Typed name: Prof. Natalia Vélez

Date: August 18, 2025

Cognitive Foundations of Collaboration

A DISSERTATION PRESENTED
BY
YANG XIANG
TO
THE DEPARTMENT OF PSYCHOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
PSYCHOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
AUGUST 2025

Dissertation advisor: Samuel J. Gershman Yang Xiang

Cognitive Foundations of Collaboration

Abstract

Collaboration enables humans to achieve goals that are beyond the reach of a single person by integrating the efforts of people who differ in what they know, what they want, and what they can do. This integration is at the heart of both the greatest benefits of collaboration and its greatest challenges—collaborators can disagree, they can work on the wrong things, or they can misunderstand one another. What cognitive capacities enable humans to navigate these challenges? In this dissertation, I propose the Collaborative Utility Calculus as a cognitive framework for understanding human collaboration. According to this framework, humans have an intuitive theory of competence and effort, which they knit together with their intuitive theory of belief-desire reasoning to make joint inferences and plan joint actions.

Chapter 1 introduces the Collaborative Utility Calculus framework and shows that people make joint inferences about collaborators' competence and effort from minimal observations. The framework also captures how these inferences guide collaborative decisions, such as how much effort to invest in a task, how to incentivize collaborators to put in their fair share, and whom to recruit to a team. Chapter 2 shows that these inferences about competence and effort are also crucial for evaluating collaborators. In particular, people assign responsibility to collaborators based on a combination of their *actual* effort—how much effort they actually exerted—and *counterfactual* effort—how much effort they could have exerted to change the outcome, given their competence. Chapter 3 shows that people care not just about how much effort their collaborators expend, but

also about *why* they do so. People make deeper trait inferences about collaborators' willingness to contribute, and are more sensitive to effort when it is voluntarily chosen than when it is required. While Chapters 1–3 focus on static competence, Chapter 4 turns to people's reasoning about how competence changes with training. This chapter shows that people do not simply recruit the most competent collaborators for the job, nor invest in training collaborators who stand to improve the most. Rather, they plan ways to optimize competence training in the service of collaboration, balancing the benefits of successful collaboration against the costs of recruitment and training. These decisions, too, are captured by the Collaborative Utility Calculus. Together, these studies offer empirical support for the Collaborative Utility Calculus as a theoretical framework for understanding how commonsense psychological reasoning gives rise to collaborative achievements.

Contents

List of Tables

List of Figures

To the people I love.

Acknowledgments

I am deeply grateful to the many people who have supported me throughout this journey.

First and foremost, I thank my advisors—Sam Gershman, Natalia Vélez, and Fiery Cushman. My dissertation is about understanding collaboration, and as I argue in Chapter 4, collaboration is also a process through which skills and knowledge are passed on. As I worked with my advisors toward shared goals, they also taught me skills and helped me grow along the way. Each of my advisors has taught me an extraordinary amount and influenced me in distinct ways: Sam taught me how to think, Natalia taught me how to write, and Fiery taught me how to speak. Sam, with his unparalleled time management, breadth of knowledge, intellectual sharpness, and an eagerness to challenge and be challenged, embodies many of the qualities of the advisor I hope to be. He has seen me at my highest highs and lowest lows, and has celebrated my successes and encouraged me through setbacks. Natalia, who first inspired me to pursue a dissertation on collaboration, is one of the most thoughtful and gracious people I have worked with. She has shown me how to move through academia with both rigor and kindness—how to make strong, compelling arguments with a voice that remains calm, generous, and respectful. My first impression of Fiery came during my graduate school interview, when he asked me a question and then answered it himself. It turned out to be a perfect preview of his remarkable ability to spark ideas and guide conversations. I am very grateful for his guidance, encouragement, wisdom, and the example he has set.

I thank my wonderful collaborators—in particular, Tobi Gerstenberg, Rahul Bhui, Tomer Ullman, and Kevin Dorst during graduate school; and Ben Enke, Toto Graeber, Jiangping Kong, and Jiyou Jia, who inspired me to begin this journey. They have all taught me a lot, both academically and personally. Special thanks to Judy Fan for her thoughtful feedback on my presentations and for sharing stories that encouraged me when things did not go well. I also thank my undergraduate research assistants: Jenna Landy, Kira Tian, Ossimi Ziv-Petreanu, Stephanie Hsu, and many others. Here again, I saw how collaboration can become a space for mentoring the next generation of scientists.

Thank you to my friends. I thank Haoxue Fan for her company in navigating academia—it's hard to believe we've known each other for ten years and, after all these years, she still constantly outwits me in jokes. I am grateful to Shuchen Liu for always being there for me, whether for major life events or a broken bike. I thank Bright Ye for being endlessly supportive and for always offering thoughtful perspectives, including the brutal honesty I sometimes needed most. I also thank Weiyao Dong, who helped design my signature agent stimuli, for the many cherished memories through college and beyond. I deeply appreciate my friends outside the lab—Yiqiao Wang, Cindy Luo, Grace Deng, Arnav Verma, and Danielle Appel. I have been equally fortunate to find friendship within the lab—Shuze Liu, Ryan Truong, David Behery, and Arthur Prat-Carrabin—as well as among col-

leagues in partner labs (CoCoDev Lab, Cushman Lab) and in the labs I visited for extended periods (Stanford CiCL and Princeton CoLab).

Finally, a warm thank you to my family. I am deeply indebted to my parents for their infinite love and support. I thank my dad for always having faith in me and unconditionally supporting my decisions, even if it meant seeing me only once every few years, and my mom for inspiring me to become a scientist and for teaching me to be humble, open-minded, and, most importantly, kind. My parents have been—and will always be—a core part of whatever I achieve. Lastly, I want to thank my partner, Eric Bigelow, with whom I feel complete. We have grown so much alongside each other, and I am excited for all the growth and adventures still ahead.

It has been a long journey, but a fun one!

0

Introduction

Two friends are lifting a couch up a flight of stairs. It's a small moment—two people aligned on a fairly simple goal, no need for elaborate tools or complicated planning. And yet in this miniature, we see all the basic components of collaboration at play. First, the two partners are *responsive* to each other's actions: if one person pauses or adjusts their grip, the other instinctively slows down or shifts position. Second, they have a *joint commitment* to completing the task. If one person stops to take a break, the collaboration does not dissolve; together, they pause, wait, and start again. Finally, they *support* each other's actions. If one partner is carrying more weight, the other might take shorter steps or lift a bit higher to rebalance the load.

Philosophical treatments of collaboration emphasize these features—mutual responsiveness,

joint commitment, and mutual support—as the core criteria for shared cooperative activity[23,78,86]. These features allow teams to form shared beliefs[76,227] and act with shared agency. This capacity for joint action underlies many of humanity's greatest achievements—including building cities, curing diseases, and landing on the moon—and manifests across an enormous range of contexts and scales. Humans form an incredible variety of collaborative teams, from small project groups to sprawling scientific consortia, from informal family dinners to tightly organized kitchen brigades. And the tools we use to collaborate are just as varied, including email, video conferencing, physical whiteboards, and shared digital documents. Yet no matter how varied the form or how advanced the tools, human collaboration is fundamentally constrained by the limits and architecture of the human mind.

At the group level, a wide range of research has characterized how collaboration unfolds in groups and teams: how people coordinate joint actions, form shared task representations, anticipate one another's behavior, distribute roles, and collaborate in real time[29,195,127,239,176,196,33,197]. These approaches have been enormously successful at describing the structure and dynamics of collaborative systems. What remains less well understood, however, is how collaborative activity is supported by cognitive representations within individual minds. How do individuals evaluate the costs and benefits of different coordination plans? How do they infer how challenging an action will be for a partner, or decide how much effort to contribute themselves? How do these local decisions that happen within individual minds accumulate and manifest at the group level, in the form of shared plans and coordinated group behavior?

In response to this challenge, several theoretical traditions have proposed key ingredients that may support collaboration at the individual level. Cognitive and developmental research has mapped out how people represent others' beliefs and desires through an intuitive *Theory of Mind*[179,10,9]. Economic theories emphasize how cooperative behaviors are shaped by external incentives and tradeoffs[242,167]. Philosophical accounts examine the structure of shared intentions and the nor-

mative commitments they entail [177,194,86,23]. However, these accounts often stop short of explaining how the mind solves the problem of collaboration—for example, how people determine who should do what, how hard it will be for them, and whether someone is a worthwhile partner in the first place. What's missing is a framework that connects these accounts, linking theory of mind, decision-theoretic reasoning, and shared intention into a unified account of the cognitive mechanisms that support collaboration.

In the following sections, I introduce the Collaborative Utility Calculus, a cognitive framework for understanding human collaboration that operates at the level of individual cognition. I first review relevant theoretical precursors, focusing on classical game theory and theory of mind. I then describe the new framework, explain how it advances existing accounts, and review recent empirical evidence for the framework.

0.1 Theoretical Precursors

To situate the Collaborative Utility Calculus within existing computational frameworks, I review two prominent traditions: classical game theory and classical Theory of Mind. These frameworks formalize the inferences and reasoning that support cooperation and collaboration. Understanding the contribution of these frameworks—and their limitations—sets the stage for my current proposal.

0.1.1 Classical Game Theory

Classical game theory is one of the most influential frameworks for modeling strategic and collective behavior [242,167,31,207]. Game theory models how agents act optimally when the payoffs of their actions depend on others' actions. In doing so, it explains what conditions motivate people to cooperate—that is, to act together for a collective benefit rather than each selfishly pursuing their

own interests. Cooperation is an umbrella term that includes not just collaboration, but also activities that do not satisfy the criteria for shared cooperative activity described above, such as donating to charity or following traffic laws. Collaboration, by contrast, is a special kind of cooperative activity in which agents work together to achieve a shared goal[86]. Understanding the basic conditions that give rise to cooperation can therefore help us understand the preconditions for human collaboration.

Game-theoretic models often assume that individuals act solely to maximize their own payoffs. In a standard coordination game, there are often multiple Nash equilibria—situations where no player can obtain a higher payoff by unilaterally changing their strategy. Contrary to this assumption, however, people tend to prefer some equilibria over others, especially those that maximize the combined payoffs for everyone[6,44]. In fact, people often seek to maximize this joint utility even when doing so reduces their own payoff[58,57,22] or increases their own costs[225,226]. These preferences and behaviors point to an underlying representation of joint utility that transcends individual payoffs. One way to account for these behaviors is to modify the rewards that players receive by adding social preferences or other-regarding concerns[59,20,62,36]. But this addition raises a deeper question: where do these preferences come from? What internal representations support them? Although changing the payoff structure can allow cooperative strategies to emerge as equilibria, this does not by itself constitute a psychologically realistic explanation[43]. Without a cognitive theory of how individuals construct and update these utilities, simply adding social preferences to game-theoretic models cannot provide a full explanation of cooperation or collaboration.

Theories of team reasoning take a different approach, by altering the unit of agency from the individual ("What do I want?") to the team ("What do we want as a team?")[213,77]. This approach can explain why people converge on collectively beneficial equilibria, but it ignores agent-specific costs. For example, suppose two equally strong people are lifting a couch together. The couch can be lifted either by one person exerting 100% effort or by both exerting 50% effort. From a team-reasoning

perspective, both solutions are equally good. However, people tend to prefer the latter[59], which reflects sensitivity not only to the outcome of the collaboration, but also to how individual costs are distributed. A related challenge concerns the formation of the team itself. Team-reasoning models typically assume that agents already reason as a team, but do not specify *ex ante* how individuals come to adopt a collective frame, or under what conditions they shift from individual to collective reasoning.

Moreover, game-theoretic models capture equilibria in stylized scenarios where the costs and payoffs to each agent are fixed. But this simple scenario underestimates how creative humans can be when they collaborate[103,173]—for example, bringing in a more capable partner can lower the effort required from everyone. Such flexibility relies on humans' ability to represent others' goals, abilities, and effort, and use that to divide labor, train novices, assign responsibility, and respond to breakdowns in coordination. To explain how humans actually collaborate, we need a framework that captures that behavioral flexibility.

0.1.2 Classical Theory of Mind

A key part of what makes humans such good collaborators is their ability to reason about other people's minds, a capacity known as *mentalizing*. This ability allows humans to infer what others want, what they believe, and what they are likely to do. For example, if you see a friend straining to lift one end of a couch, you might naturally infer that their goal is to move it. This everyday inference reflects a *Theory of Mind*[179]—an intuitive understanding of how people's observable behaviors are shaped by their unobservable beliefs and desires. The ability to engage in belief-desire reasoning is core to collaboration; to coordinate with a partner, we need to understand what they are trying to achieve and what they believe about the situation.

The computations behind Theory of Mind have been formalized in Bayesian models of commonsense psychological reasoning[10,9], which frame action understanding as a process of probabilis-

tic inference over unobservable mental states. In these models, people infer what others believe and desire by observing their behavior, under the assumption that agents take efficient actions to achieve their goals. Thus, deviations from the most efficient action—such as taking a detour to check out an option hidden from view[9]—reveal the agent's goals and preferences. Extending this framework to multi-agent settings has proven fruitful for explaining group behaviors, such as how agents infer collaborators' intentions and what they should do without centralized planning[260], and how agents use other people's behavioral trajectories to learn about the environment when reward information is private[95]. These models illustrate how probabilistic mental-state reasoning can scale to more complex social environments. However, these models assume a universal notion of efficiency by treating action costs as fixed. In reality, what counts as an "efficient" action depends on individual differences in ability and effort.

The Naive Utility Calculus framework extends Bayesian Theory of Mind models by incorporating individual variation in costs and rewards[114]; for example, climbing a tall box may be easy for one agent but impossible for another[115]. However, in collaborative settings, the cost to individuals depends not only on their own capabilities, but also on what their collaborator contributes. Climbing is easier if your partner gives you a boost from below, compared to if your partner is barely trying to help. This gap motivates the need for a framework that captures how collaborators redistribute effort to achieve shared goals.

0.2 Collaborative Utility Calculus

In this section, I propose the *Collaborative Utility Calculus* as a cognitive framework for understanding human collaboration. The framework makes unique contributions to existing approaches. It explains collaboration from the perspective of basic mechanisms of human social cognition, which includes concepts such as mental representations of joint utility that are not typically cap-

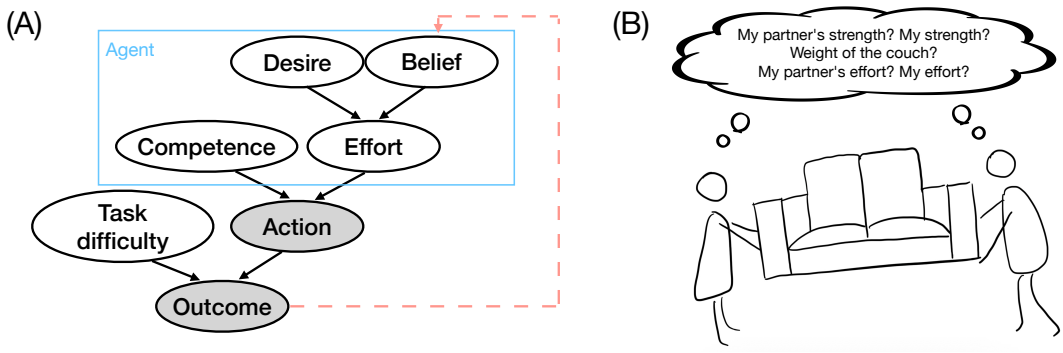


Figure 1: Overview of the Collaborative Utility Calculus. (A) Schematic of the Collaborative Utility Calculus, (B) illustrated in an example of two friends lifting a couch together. Each person considers their own mental states and competence, as well as their partner's mental states and competence, and chooses actions that maximize the joint utility of the team.

tured in classical game theory. The framework also extends existing models of commonsense psychological reasoning into the domain of collaboration to explain how people reason about joint utility and plan shared actions. It makes explicit theoretical commitments about how *competence* and *effort* jointly determine an agent's contribution to a collaborative task. It also formalizes the idea that people anticipate how work will be distributed in joint settings, and that how much effort one exerts depends in part on how much their partner is expected to contribute.

Figure 1 shows an illustration of the framework. The framework includes a number of agent-specific properties: (a) the reward value of an action or outcome (*desire*), (b) what the agent knows about themselves and the world (*beliefs*), (c) the agent's ability to carry out the task (*competence*), and (d) how hard the agent tries at the task, i.e., the proportion of their competence applied to the task (*effort*). These properties produce an action, which is compared with the task difficulty to determine whether a collaboration succeeds or fails (*outcome*). The outcome then feeds back into the agent's beliefs: it becomes part of what the agent knows and is used to adjust their future actions.

Note that the framework applies more broadly to agent properties that are conceptually similar to competence and effort, beyond the literal sense. For example, a person's available resources or

wealth can be thought of as analogous to competence—that is, the maximum they can contribute, whereas their generosity or fairness maps onto effort, as it reflects the proportion of their capacity they choose to allocate to their partner[89,181]. The framework can also capture situations where factors other than effort might reduce how much competence an agent brings to a task, such as temporary limitations (e.g., fatigue or illness), environmental constraints (e.g., bad weather or a noisy room), or even strategic self-handicapping[264].

All these nodes interact with one another in intricate ways. First, competence and effort work together to determine the outcome of a collaboration. Even the most competent agents cannot succeed without investing the required effort, nor can the most hardworking agents if they lack the required competence. Second, agents choose effort allocations that maximize their joint utility—the expected reward for the team minus the costs of exerting effort. This ensures that agents prioritize the team's shared goal while also being sensitive to the individual costs they each incur. Third, the expected reward calculation hinges on joint inferences about one's own and their collaborators' competence and mental states. For example, if you are lifting a heavy couch with a friend, it is important to understand how strong your friend is—i.e., their competence—and how much effort they will put in, so that you can calibrate how much effort to apply to lift your end of the couch. The framework predicts that agents recursively reason about how much effort they and their collaborators will allocate to a task. With these elements, the framework can capture a wide range of phenomena in human collaboration, as I review in the next section.

0.3 Evidence for the Collaborative Utility Calculus

In this section, I review recent empirical evidence for the Collaborative Utility Calculus as a cognitive framework for understanding human collaboration. These studies collectively show how reasoning about other minds enables humans to navigate basic problems in collaboration, including

deciding how much effort to invest in a collaborative task, how to incentivize collaborators to put in their fair share, whom to recruit to a team, how to invest in their training, and how to assign responsibility and rewards for the outcome of a collaboration. In many of the studies below, the Collaborative Utility Calculus was compared against simple alternative models to demonstrate the need to reason recursively about collaborators' mental states and competence. In addition, I highlight the potential for this framework to shed light on a wide range of phenomena, such as how people reason about external and internal constraints, how collaborators achieve efficient division of labor, and how collaborations evolve over time.

0.3.1 Inferring competence and effort

In an episode of *30 Rock*, Kenneth Parcell created a game show where models held up identical-looking cases and contestants guessed which one contained $1 million worth of gold. The show was quickly shut down after it became clear that the design had a fatal flaw: since gold was heavy, contestants could easily identify the correct case by noticing which model was visibly struggling to hold it up. Underlying this funny scene is a basic principle of social reasoning: people can infer others' effort and competence—in this case, because the struggling model is not obviously weaker than the other models, one can infer that her case is unusually heavy. This ability is also core to collaboration: to work together effectively, we need to understand what our collaborators can do and how much effort they are willing to contribute.

Even children are capable of inferring other people's competence based on the difficulty of the tasks they are willing to take on. In one study, 5–6 year olds watched two puppets—Grover and Cookie Monster—take a cracker that was placed on a short box rather than a cookie that was placed on a tall box. Children were told that Cookie Monster likes cookies more than crackers, while Grover likes both equally. From this scene, children inferred that Cookie Monster was *unable* to climb the box—if he could, he would have taken the cookie [115]. Even in cases where task difficulty

is unknown, adults can jointly infer task difficulty and other people's competence by observing which agents succeed at more tasks and which tasks have higher success rates[113]. The Naive Utility Calculus explains how people make these inferences by reasoning about subjective rewards and costs—that is, how much an agent values each goal and how hard it is for that agent to reach it.

The Collaborative Utility Calculus builds on this foundation in two ways. First, it accounts for *effort* as a component of subjective cost that is distinct from ability—recognizing that people may fail to reach a difficult goal not because they lack the ability, but because they are unwilling to expend effort or because their collaborator did not put in their fair share. In one study, I showed participants two game show contestants who attempted to lift boxes of different weights for different reward amounts[267]. People inferred which contestant was stronger not just from their successes or failures, but also from how these outcomes related to the *incentives* offered to contestants. For example, when both contestants successfully lifted a heavy box, participants inferred that the contestant who succeeded with a smaller reward was stronger—intuitively, stronger contestants require smaller rewards to offset the costs of lifting. Consistent with this reasoning, participants also inferred that stronger contestants exerted less effort to lift the same weight (see Figure 2A-B).

Second, the Collaborative Utility Calculus goes beyond inferring utilities from *individual* actions to disentangling individual utilities from *joint* actions, where multiple people's contributions are combined into a single outcome. In the same study as above[267], participants then saw pairs of contestants attempt to lift a box together. They inferred that each contestant exerted less effort when sharing the weight with a partner, compared to lifting it alone—suggesting that participants understood that collaborations distribute effort between partners (see Figure 2A-B). Participants' judgments were captured by the Collaborative Utility Calculus, which models how each contestant adjusts their effort by reasoning recursively about how much their partner will contribute, but not by simpler heuristic models that lacked recursive reasoning.

Thus, the Collaborative Utility Calculus captures human inferences about competence and ef-

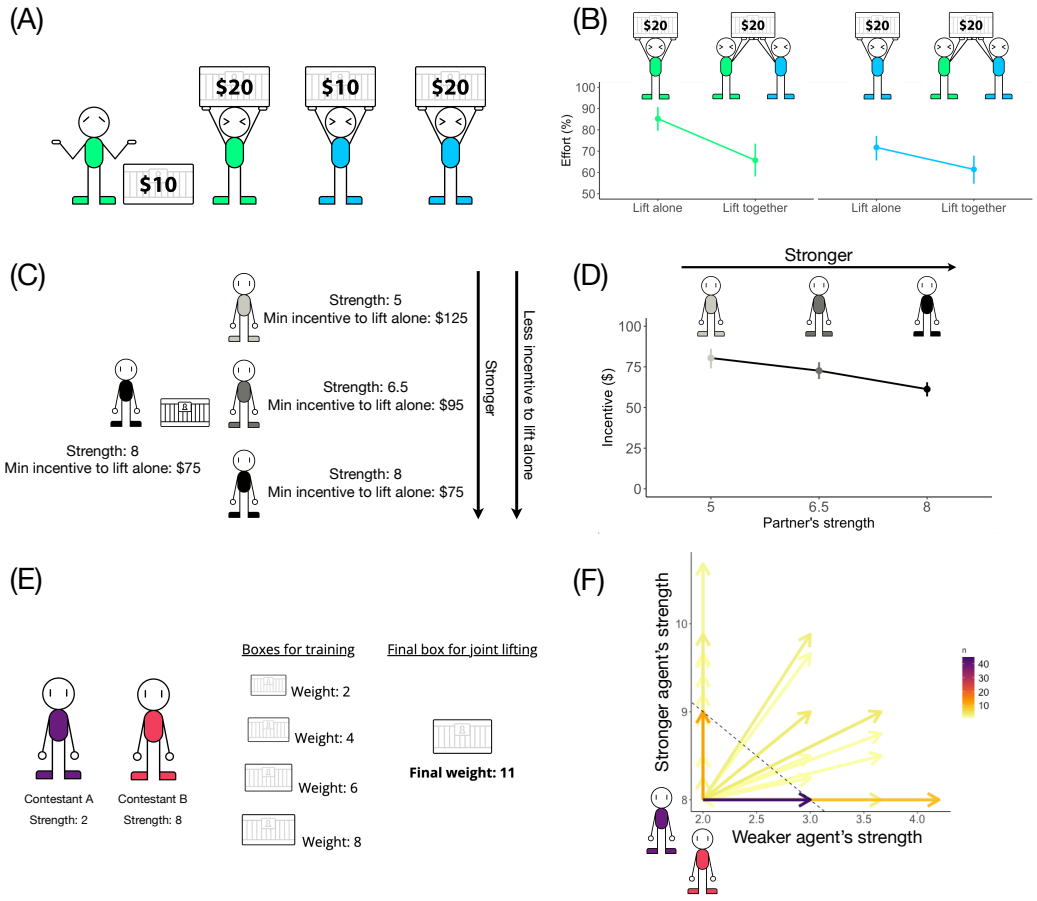


Figure 2: Empirical evidence for the Collaborative Utility Calculus. (A-B) Effort allocation[267]: People can infer agent-specific costs—namely, competence and effort—based on the rewards that agents are willing to accept to do a difficult task. When agents lifted a heavy box alone, participants inferred that agents who lifted the box for a small, $10 reward (blue agent) were stronger and exerted less effort than agents who only lifted the box for a large, $20 reward. When the agents lifted the box together, participants inferred that each agent would exert less effort than when they lifted on their own. **(C-D) Incentive allocation**[267]: The Collaborative Utility Calculus enables inferences about how one person's effort may change based on their partner's contributions. Participants offered smaller incentives when the agent on the left was paired with a stronger partner, reflecting the expectation that stronger teams expend less effort to complete the task. **(E-F) Optimizing competence**[268]: The Collaborative Utility Calculus also supports more sophisticated planning based on how collaborators' competence may change with training. Participants trained agents to be just strong enough to lift a heavy box together through their combined strength, balancing the costs and benefits of training.

fort in collaborative tasks. A key question for future work is whether this same reasoning also governs how people calibrate their *own* effort in collaborations. This possibility aligns with recent developmental work suggesting that young children use social information to estimate the difficulty of a task and calibrate their effort accordingly. For example, infants pull harder with each subsequent attempt to retrieve a toy that is difficult for an adult to move, but do not increase their effort if the task was impossible for the adult [151]. Infants also try harder after observing an adult try repeatedly at a different task, rather than when the adult succeeds effortlessly [138]. Critically, infants do not indiscriminately copy adults' effort; they selectively increase their own efforts when they see that adults' struggles lead to success, rather than failure [137]. These social cues provide infants with information about how effortful a task will be *for them*, and whether their effort is likely to pay off. Applying the Collaborative Utility Calculus to this domain could shed light on the psychological mechanisms that enable people to translate social information about other people's capabilities and outcomes into principled decisions about when and how much to contribute.

0.3.2 Using competence and effort for collaborative decision-making

A key implication of the Collaborative Utility Calculus is that reasoning about other people's competence and effort guides practical decisions about how to build a team, by revealing what team members can do and what costs they are expected to incur. This information can then be used to incentivize collaborators to put in their fair share. In one study, I asked participants to choose a monetary incentive that both team members would receive if they successfully lifted a heavy box together [267]. Across conditions, participants saw the same agent paired with partners of different strengths. Participants offered lower incentives when this agent was paired with a stronger teammate, and higher incentives when paired with a weaker teammate (Figure 2C-D). These results suggest that participants set incentives to compensate for the *effort* that teams would have to put in to complete the task. Intuitively, stronger teams expend less effort to produce the same physical force

as weaker teams, so they require smaller incentives to make the task worthwhile. Similarly, when I increased one agent's strength, participants reduced the team's incentive accordingly. These patterns were captured by the Collaborative Utility Calculus, but not by alternative models that use simple heuristics in place of recursive reasoning.

The cost of effort can also vary depending on individual differences in each collaborator's *willingness* to contribute. In another study, participants recruited pairs of agents out of a lineup [267]. Each agent in the lineup differed not only in their strength, but also in their *effort cost*—that is, the lowest incentive that they would accept to put in an all-out effort. Intuitively, agents with lower effort costs are "hard-working", or more willing to exert effort. After picking a team, participants then offered each team member an incentive to lift boxes of varying weights. This design enables me to test how participants build teams by weighing each agent's strength against their willingness to expend effort. When the box was light, participants recruited weaker but more hardworking agents who required a smaller incentive to expend the necessary effort. In contrast, when the box was too heavy for the weaker agents, participants recruited stronger but less hardworking agents and were willing to offer larger incentives.

Willingness to exert effort matters not only for the success of a single collaboration, but also for identifying people who will be good collaborators in the future. In another study, participants selected how much each agent in a pair should be rewarded for successfully lifting a box together [265]. Participants were told that one agent was *instructed* to exert a fixed amount of effort, while the other chose how much effort to invest on their own. Participants apportioned rewards based not only on how much effort each agent had invested, but also on their reasons for doing so. Overall, participants gave agents rewards that were in line with their expended effort, but this effect was moderated by whether they were instructed or self-directed. They gave the self-directed agent rewards that were more proportional to actual effort expended, compared to the instructed agent. These results suggest that people evaluate collaborators based not only on effort, but also on what that effort reveals

about their willingness to contribute in future collaborations.

Finally, reasoning about competence and effort is crucial for evaluating the outcome of a collaboration—that is, to decide who deserves a larger share of blame for a failed collaboration or credit for a successful one. Past work has shown that people are sensitive to both effort [17,116] and competence [68] in responsibility judgments. However, these factors have largely been studied in isolation, which leaves open the question of how people integrate competence and effort to assign responsibility. In a set of studies, participants observed pairs of agents lifting a box together and evaluated how responsible each agent was for the outcome. I systematically varied box weight, agent strength, and effort to test how each factor contributed to participants' responsibility judgments. Overall, participants assigned responsibility based on effort, rather than competence or the amount of force exerted on the box. Moreover, participants' judgments were shaped by how much effort each agent *actually* contributed and by a counterfactual judgment of how much the agent *could have* changed the outcome by exerting more or less effort, given their competence. Intuitively, the agents who received the most credit for successes exerted a large amount of effort *and* were pivotal to their team's success [266].

These studies collectively show how reasoning about other minds enables humans to navigate basic problems in collaboration, including deciding how much effort to invest in a collaborative task, how to incentivize collaborators to put in their fair share, whom to recruit to a team, and how to assign responsibility and rewards for the outcome of a collaboration. A natural next step is to consider how the Collaborative Utility Calculus may inform decision-making as collaborations unfold. One intriguing possibility for future work is that the framework may provide a psychological mechanism to explain how human collaborations achieve efficient divisions of labor. As collaborators reason recursively about each other's competence and effort, they may converge on efficient task allocations through mutual adaptation. This proposal aligns with recent empirical work on "co-efficiency", showing that collaborators often act to minimize joint costs [225,226,195,127,24].

0.3.3 Optimizing competence

The sections above have established that, in order to efficiently divide labor with others, we need to represent our collaborators' competence. However, competence is not static; people get better at particular tasks the more they perform them. Indeed, this opportunity to improve has long been touted as a key benefit of division of labor: splitting up tasks affords individuals more chances to practice and refine their area of specialization[205]. However, it also raises an important question in collaborative decision making: Is it better to recruit a superstar who is the most capable right now, or a rising star whose competence can grow with training and investment? And how do decisions about whom to train affect whom people choose to recruit in the first place?

According to the Collaborative Utility Calculus, people make these decisions by weighing the costs of recruiting and training a collaborator against the benefits that they will bring to the collaboration. In one study, participants were shown a heavy box that required two agents working together to lift successfully. They then had to recruit agents of varying strengths from a lineup and train them to reach the combined strength needed for this collaborative goal[268]. I found that participants did not simply recruit the strongest agents for the job, nor invest in those who would improve the most. Instead, they strategically selected and trained agents to be *just strong enough* to lift the target weight, balancing the costs of training against the benefits (see Figure 2E-F). These results generalized to a separate experiment where participants trained students for a math Olympiad, which suggests that these training decisions generalize beyond increasing physical strength. In another study, participants trained two military defense teams to counter two types of attacks (land and air), where the long-term goals and the teams' relative competences varied[221]. Overall, participants trained collaborators by taking the long view—considering how collaborators' competence would develop, and whether they could rise to meet task demands—rather than making myopic decisions based on simpler heuristics such as current competence, learning potential, fairness, or versatility. Together, these

findings illustrate how the Collaborative Utility Calculus captures not just static inferences about others' abilities, but also how people cultivate individual competences to support collaboration over time and ultimately benefit from each other's evolving competences.

1

Collaborative decision making is grounded in representations of other people's competence and effort

1.1 Abstract

By collaborating with others, humans can pool their limited knowledge, skills, and resources to achieve goals that outstrip the abilities of any one person. What cognitive capacities make human collaboration possible? Here, we propose that collaboration is grounded in an intuitive understand-

ing of how others think and of what they can do—in other words, of their mental states and competence. We present a Collaborative Utility Calculus framework that formalizes this proposal by extending existing models of commonsense psychological reasoning. Our framework predicts that agents recursively reason how much effort they and their partner will allocate to a task, based on the rewards at stake and on their own and their collaborator's competence. Across three experiments (N = 249), we show that the Collaborative Utility Calculus framework captures human judgments in a variety of contexts that are critical to collaboration, including predicting whether a joint activity will succeed (Experiment 1), selecting incentives for collaborators (Experiment 2), and choosing which individuals to recruit for a collaborative task (Experiment 3). Our work provides a theoretical framework for understanding how commonsense psychological reasoning contributes to collaborative achievements.

1.2 Introduction

Collaboration enables humans to achieve goals beyond the capabilities of any one individual: No one can build a city, land on the moon, or even carry a couch up a flight of stairs entirely on their own. Humans are unique in their ability to form collaborative arrangements that are maintained over generations, as in the case of institutions, and that span continents, as in the case of large scientific collaborations. Compared to other animals, we've made it a key part of our niche [223], and the roots of this ability arise early in development (see Tomasello and Hamann [224], for a review). How do we do it? What cognitive capacities give rise to these collaborative achievements?

A key part of the answer is that, in order to work with others, we have to understand how our collaborators think and act. Philosophical treatments of collaboration provide an inventory of basic mental states and commitments required for collaborative action. To fulfill these criteria, agents must hold certain representations in mind, including a representation of shared goals, knowledge

and beliefs about which actions will lead to the fulfillment of the goal, and intentions to carry out said actions [177,194,86,23]. Recent work suggests that agents plan their actions by anticipating those of their collaborative partners, revealing that people minimize joint action costs [225,226], compute a "mind of the group" i.e., an average group member's mind; [123], as well as represent, monitor, and predict each other's actions to achieve a joint goal [196,240,261]. Thus, fundamentally, successful collaboration requires that we understand how others think and act.

Bayesian models of commonsense psychological reasoning formalize these computations as theory of mind [179], which takes the form of belief-desire reasoning: People act based on their beliefs to achieve their desires [10,9]. However, even state-of-the-art psychological models are missing two ingredients that are particularly important to successfully work together. The first missing ingredient is a representation of one's own and of others' competence. A person's competence entails all kinds of physical and mental abilities, such as strength, speed, memory, linguistic skill, etc. Traditional models of belief-desire reasoning assume that agents are capable of fulfilling their desires as long as the constraints of the environment allow it. But this is not always so—an individual might have the necessary beliefs and desires to complete a task, but lack the competence to carry it out. For example, imagine a child who knows where a box is, has the desire to lift the box for a cookie as reward, but is not strong enough do so. The child would not be able to lift the box and get the cookie reward herself. However, she could choose to collaborate with another child who is strong enough to lift the box but does not know where it is. This example illustrates the necessity of reasoning about one's own and other people's competence in collaborative activities. Past work suggests that even young children are capable of inferring other people's competence based on the difficulty of the tasks they are willing to take on [115,114]. However, models of commonsense psychological reasoning have yet to incorporate a representation of other people's competence that is separate from other costs, such as the inherent difficulty of a task.

The second missing ingredient is a formal theory of how individuals allocate effort to joint tasks.

Recent work suggests that even young children use social information to decide how to allocate effort to a task; for example, young children might observe others to infer the difficulty of a task [151] or use their example to determine whether expending effort will pay off [138,137]. Children are also capable of dividing physical and cognitive labor in collaborative tasks based on relative ability and task difficulty [153,7]. More recent models have enriched theories of psychological reasoning by demonstrating that people do not merely expect others to fulfill desires—thus solely maximizing rewards—but rather expect others to act as utility maximizers—balancing the rewards of an outcome against the costs of achieving it [114]. However, existing models do not capture the fact that engaging in collaborative action reduces the effort that any one individual needs to expend on a task: lifting a couch alone may take an intense all-out effort, but feel significantly easier when the burden is shared between two friends. It remains an open question how agents figure out how to adjust their effort in collaborative settings—for example, how we anticipate the effort of others and calibrate our own in order to lift the couch successfully without tiring ourselves completely.

To fill these gaps, we present a Collaborative Utility Calculus framework that extends existing models of commonsense psychological reasoning to incorporate representations of competence and effort. Our model predicts that agents recursively reason how much effort they and their partner will allocate to a task, based on the rewards at stake and on their own and their collaborator's competence. In addition, we contrast our model with three alternative models that use simpler heuristics (solitary, compensatory, and maximum effort models) and show that joint reasoning is necessary for explaining collaborative behavior. The alternative models might seem impossible at first blush, but they are not a priori impossible for all the scenarios. Moreover, the alternative models are motivated by phenomena such as social compensation and social loafing (see Theoretical Framework, below). Therefore, demonstrating that these models are not sufficient to explain behavior in our studies provides support for our claim that collaborative decision making is grounded in joint reasoning about competence and effort.

Across three experiments ($N = 249$), we demonstrate that the Collaborative Utility Calculus framework captures human judgments in a variety of contexts that are critical for collaboration. In each experiment, participants watched contestants in a game show attempt to lift a box to win cash prizes. In Experiment 1, each contest was divided into three rounds: In the first two rounds, contestants attempted to lift the box on their own, in order to provide information about their strength and about how they respond to incentives. In the third, critical round, two contestants then attempted to lift the box together. Participants inferred how strong each contestant is, how much effort they would put into lifting the box in the third round, and how likely they are to succeed. In Experiments 2–3, participants saw contestants with different strengths who required different incentives to lift a heavy weight, and they decided how much incentive to provide (Experiment 2) and which contestants to recruit (Experiment 3) to lift boxes of different weights. Across all models, we find that a model that recursively infers how much effort to allocate best matches human judgments, compared to various alternative models that implement simpler heuristics rather than reasoning recursively about effort. * Together, our results provide evidence that collaborative decisions are grounded in an intuitive understanding of how others think and of what they can do.

1.3 Theoretical Framework

We propose a Collaborative Utility Calculus framework (Figure 1.1) to illustrate how people's reasoning about beliefs, desires, and competence determine the action they take. After laying out the framework, we formalize the idea in a simple context of box-lifting.

*All data and code are publicly available at https://github.com/yyyxiang/competence_effort.

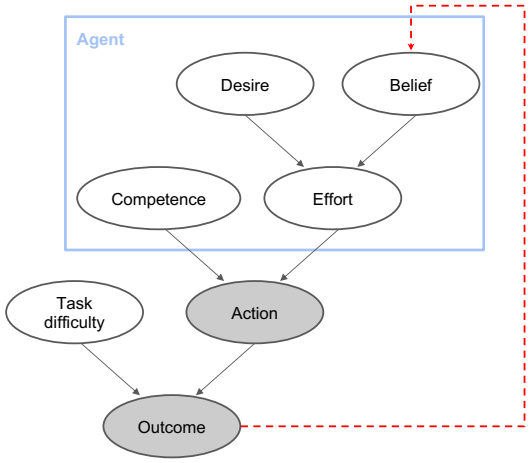


Figure 1.1: An illustration of the Collaborative Utility Calculus framework. Shaded nodes represent observed variables; unshaded nodes represent latent variables to be inferred. The agent's desires and beliefs determine their effort. Their competence and effort together determine the action to take. The outcome is a function of their action and the task difficulty. Observing the outcome updates the agent's beliefs.

1.3.1 Overview of the Collaborative Utility Calculus framework

We assume that the agent's desires and beliefs about the internal state and external environment determine the amount of effort the agent would exert. The action to take depends on both the agent's effort and competence, which encompasses all kinds of physical and mental abilities. The final outcome is a function of the difficulty of the task and the action the agent takes. The agent updates their beliefs after observing the outcome.

We simplified the framework and formalized it in a setting where agents of different strengths attempt to lift boxes of different weights. This allowed us to generate hypotheses to test the key predictions of this framework. Specifically, the agent's desire is the reward they get from lifting the box, the agent's competence is their physical strength, and the task difficulty is the box weight. Observing the outcome (either Lift or Fail) allows the agent to update their belief about their strength, the box weight, etc. One advantage of testing the model in a physical domain like box lifting, compared to cognitive domains such as numerosity judgments, is that the task difficulty and agents' competence

and effort have concrete units. Thus, we can express the relationship between effort and action outcomes in a simple, deterministic model: Agents' competence and effort translate into the force that they exert, and they succeed if that force exceeds the weight of the box. [†] Below we provide a formal description of our model predicated on this simplified framework.

1.3.2 Formal description

We first consider a single-agent case, then generalize it to a multi-agent setting which involves recursive reasoning between agents, followed by three alternative models without recursive reasoning.

Single-agent model

The single-agent model describes how an agent decides the amount of effort to exert in an attempt to lift a box. It consists of the following components:

- $S \geq 0$ is the strength of the agent, expressed in units of weight (i.e., the maximum amount of weight they are able to lift). We assume strength is a static property of each agent.
- $W \geq 0$ is the weight of the box.
- $E \in [0, 1]$ is the effort the agent puts into lifting the box. It defines the proportion of an agent's strength that is applied to lifting.[‡]
- $L = 1$ indicates that the agent lifts the box ($L = 0$ otherwise). This is a deterministic function of effort, strength and weight: $L = \mathbb{I}[E \cdot S \geq W]$, where the indicator function $\mathbb{I}[\cdot] = 1$ if its argument is true, 0 otherwise.

[†] In real-world lifting events, it is possible for a lifter to fail to lift a weight, even if they are strong enough to lift it and exert the needed effort, due to slight variations in body posture, hand positioning, etc. However, we think that a deterministic function provides a reasonable approximation within this simplified task.

[‡] For computational tractability, we constrained the effort space to discrete values, ranging from 0 to 1 with increments of 0.05.

- $R \geq 0$ is the reward for successfully lifting the box.
- $C(E)$ is the agent's effort cost. For simplicity, we adopt a linear cost function $C(E) = \alpha E$, where $\alpha \geq 0$ is a scaling parameter that denotes the agent's laziness (larger α means lazier) [§].
- $U(E)$ is the utility function that measures the net reward minus cost:

$$U(E) = R \cdot P(L = 1|E, W) - C(E), \tag{1.1}$$

  where the probability of lifting the box is given by marginalizing over all possible strengths:

$$P(L = 1|E, W) = \int_S P(S)\mathbb{I}[E \cdot S \geq W]dS. \tag{1.2}$$

  This model can be straightforwardly extended to account for uncertainty about W and α.

We model agents as utility maximizers, such that the optimal effort is:

$$E^* = \underset{E}{\text{argmax}}\, U(E). \tag{1.3}$$

Multi-agent joint effort model

The multi-agent joint effort model generalizes the single-agent model to a multi-agent setting. The model is termed "joint effort" because agents reason about others' effort recursively when deciding how much effort to allocate. When multiple agents attempt to lift a box together, the lift outcome

[§]While we refer to the α parameter as *laziness*, it can also be understood as a form of competence-dependent reward scaling. For example, an agent might deserve to be paid more due to their superiority over others in terms of competence.

depends on every agent's effort, strength, and the box weight:

$$L = \mathbb{I}[\textstyle\sum_a E_a S_a \geq W], \tag{1.4}$$

where a indexes agents. The optimal effort is given by:

$$\mathbf{E}^* = \underset{\mathbf{E}}{\operatorname{argmax}}\, R \cdot P(L = 1|\mathbf{E}, W) - C(\mathbf{E}), \tag{1.5}$$

where $\mathbf{E} = [e_1, \ldots, e_n]$ is the vector of efforts for n agents. In practice, we solve the joint optimization by fixed point iteration: holding the efforts of all but one agent fixed, we maximize the utility with respect to the focal agent, iterating over agents until convergence (convergence threshold set to 0.06).

People have a general preference for fairness [59,215], and these fairness preferences play a role in human collaboration [19,94]. In light of this, we add to the cost function a Gini coefficient that penalizes unequal effort allocations. ¶ The Gini coefficient is a very widely-used measure of income inequality in economics [5,198,32], and has been applied to a myriad of fields to measure education inequality [220], health inequality [184], and yield inequality [188]. The Gini coefficient is defined as half of the relative mean absolute difference in effort [198]:

$$G(\mathbf{E}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |E_i - E_j|}{2n \sum_{j=1}^{n} E_j}, \tag{1.6}$$

where $|E_i - E_j|$ denotes the absolute difference in effort between each pair of agents (i, j). To take

¶ We contrast the joint effort model with and without the Gini coefficient in the supplemental material (Figures S1-S5). From Figure S1B, we can clearly see that model predictions of effort change drastically with and without the Gini coefficient, thereby demonstrating the importance of penalizing unequal effort allocations.

fairness preferences into account, we add a term to the cost function:

$$C(\mathbf{E}) = \sum_a \alpha_a E_a + \beta G(\mathbf{E}), \tag{1.7}$$

where $\beta \geq 0$ is a scaling parameter. Note that we allow different agents to have different effort cost coefficients (α_a), an assumption that we explore further in Experiment 3.

Alternative models

In real life, not all teamwork operates as the joint effort model suggests. It is not uncommon that in some group projects, one person does almost all the work while the others do very minimal work, if at all. Indeed, similar types of behaviors have been documented in the literature and explained as *social compensation* [253], where individuals work hard collectively when they expect their teammates to be less invested or less capable, and *social loafing* [121,134], where free riders take advantage of their teammates and expect them to pick up the slack. Inspired by these phenomena, we introduce two alternative models which do not invoke recursive reasoning of effort: The solitary effort model and the compensatory effort model.

The **solitary effort** model instantiates an extreme case of social compensation. Agents assume that their partners will expend 0 effort on the task, and allocate effort as though they were performing the task alone:

$$\mathbf{E}_{/a} = 0, \tag{1.8}$$

where $/a$ indexes all the agents except agent a. Each agent's optimal effort only depends on their

own strength, effort cost, and the box weight:

$$E_a^* = \underset{E_a}{\operatorname{argmax}}\, R \cdot \mathbb{I}[E_a S_a \geq W] - \alpha_a E_a. \tag{1.9}$$

Therefore, we have:

$$\begin{aligned} E_a^* &= \underset{E_a}{\operatorname{argmax}}\, R \cdot P(L = 1 | E_a, \mathbf{E}_{/a} = 0, W) - \alpha_a E_a \\ &= \underset{E_a}{\operatorname{argmax}}\, U(E_a), \end{aligned} \tag{1.10}$$

where $U(E_a)$ is the single agent utility function described above (Equation 1.1). Note that the solitary effort model does not include a Gini coefficient. This is because a model that instantiates an extreme case of social compensation likely does not care about (un)fairness. The same reasoning applies to the compensatory effort model. ‖

The **compensatory effort** model instantiates the other extreme (social loafing): Agents assume that their partners will exert maximal effort, and therefore the agents will exert only the minimum effort needed to accomplish the task:

$$\mathbf{E}_{/a} = 1. \tag{1.11}$$

The optimal effort is given by:

$$E_a^* = \underset{E_a}{\operatorname{argmax}}\, R \cdot P(L = 1 | E_a, \mathbf{E}_{/a} = 1, W) - \alpha_a E_a. \tag{1.12}$$

The solitary effort and compensatory effort models make the same predictions as the joint effort

‖For completeness, we show the predictions of all models with the Gini coefficient in the supplemental material (Figures S6-S10).

model for single-agent events, but diverge for multi-agent events. The joint effort model assumes that agents reason recursively about each other's effort, while the solitary effort and compensatory effort models make fixed assumptions about the efforts of other agents.

In addition to the two alternative models described above, we include a third alternative: the **maximum effort** model. This model is motivated by an implicit assumption in some previous research e.g., [113] that people expect agents to exhibit their full capacity when they take actions. The maximum effort model simply assumes that agents are always exerting all their effort regardless of their strength, effort cost, partners, etc., which means $\mathbf{E} = 1$. This is similar to the compensatory effort model, except that here agents do not modulate their own effort allocation in response to the maximal effort of their partners.

1.3.3 Model implementation and assessment

We implemented the model in WebPPL, a probabilistic programming language embedded in Javascript [83].

The strength samples were drawn from a uniform distribution with a lower bound of 1 and upper bound of 10. We used Markov Chain Monte Carlo sampling with a Metropolis-Hastings kernel, 10000 samples, and 1000 burn-in samples (i.e., additional sampling iterations before collecting samples), conditioning on the observations. Note that these only apply to Experiment 1 where agents have uncertainty about their own strength and their teammate's strength. In Experiments 2 and 3, subjects observe strength information and therefore those variables do not need to be sampled.

Our joint effort model includes two scaling parameters in the cost function: α, which denotes an agent's effort cost (laziness), and β, which determines the degree of penalization for unequal effort allocations. These were the only two free parameters in our model. In Experiment 1, both α and β are free parameters; we tuned them by hand and optimized them to maximize the correlation between model-predicted and empirically measured effort judgments pooled across participants. In Experiment 2 and Experiment 3, the α values were constrained by our experimental design (they

were made explicit to the participants via the task description) and were not free parameters. To make model implementation consistent across experiments, we reused the β value from Experiment 1 for Experiment 2 and Experiment 3. See Table 1.1 for a complete list of α and β values used in the joint effort model. The three alternative models use the same α values as the joint effort model, but do not include the β parameter.

To assess how well the models match behavioral data, we calculated the Pearson correlation coefficient between model predictions and participants' judgments and plotted them against each other. Even though theoretically we could do likelihood-based model fitting, we are not making strong claims about the parametric details of the models. Rather, our claims concern the qualitative patterns of the model predictions.

1.4 Experiment 1

Our first experiment aims to test a few basic judgments about collaboration, including how people reason about competence and effort and how they predict how likely a collaborative event is to succeed. To be directly compatible with the model setup, we designed the experiment to be a game show where contestants attempt to lift a box by themselves or together with another contestant. We manipulated the individual lift outcome of different contestants shown to the participants and asked them to report judgments regarding what they predict would happen in a group lifting. We deliberately chose to not involve participants as active agents, but rather have them observe agents'

Table 1.1: Scaling parameter values.

	α	β
Experiment 1	13.5	24.5
Experiment 2	125	24.5
Experiment 3	Agent-specific (40, 20, or 2)	24.5

behavior. This is because our work focuses on people's theory of mind, which is how people think *others* behave: Participants observe two agents' behavior and make judgments regarding how they think the agents collaborate.

Across different scenarios, we held the weight of the box constant during the group lift events, and we manipulated the reward for a successful lift and outcomes of agents' lifts in prior events to affect observers' beliefs about each agent's competence. Participants observed that some contestants refused to lift a box for a low reward, but readily lifted it for a higher reward. Beyond simply tracking agents' competence, these covariation data also provide information about how different agents respond to incentives. The different models make the following predictions:

- Strength: We predict that people are able to estimate contestants' strength based on observations of individual lift events. If a contestant lifts a box successfully, then they should be stronger than another contestant who fails to lift the same box. Furthermore, if they could lift the box with lower reward, then they should be stronger than contestants who require a higher reward, because accepting a lower reward indicates that less effort is needed to do the lifting. The joint effort model and the alternative models make similar predictions.

- Effort: When both contestants are strong enough to lift the box by themselves, we predict that people should expect them to exert less effort when they attempt to lift the box together. We also expect the models to make qualitatively different predictions. In particular, the joint effort model predicts this decrease in effort. The solitary effort model predicts no change in effort, since it assumes no difference between lifting the box themselves and lifting with another contestant. The compensatory effort model assumes that every contestant should expect the other one, who is strong enough to lift the box themselves, to be solely responsible for the lifting. As a result, it predicts that neither contestant will exert any effort. Finally, the maximum effort model predicts that the effort does not change from individual lifting to

group lifting, since it assumes that contestants are always exerting all their effort.

- Lift probability: We predict that people should judge a group lifting to be more likely to succeed when the contestants involved are stronger. And when at least one contestant is strong enough to lift the box themselves, the group lift probability should be high. This pattern is predicted by the joint effort model, solitary effort model, and maximum effort model. In contrast, the compensatory effort model predicts that group lifting is impossible when at least one contestant is strong enough to lift the box by themselves. This is because the other contestant would expect the contestant who is strong enough to do the lifting and conversely, that contestant would expect the other one to exert their full effort, so they would reduce their effort accordingly, resulting in failure.

  We also predict that even if neither of the contestants is strong enough to lift the box themselves, it would still be possible for them to lift the box together. This prediction is consistent with the joint effort model and maximum effort model, which argue that "two are better than one." The compensatory effort model predicts something similar in this case, because it is possible that the sum of the contestants' strength equal the box weight, and if each contestant considers the remaining part after subtracting the other contestant's strength requiring their full effort, that would mean both contestants exerting all their effort and possibly lifting the box successfully. The solitary effort model in this case predicts that group lifting is impossible, because neither contestant can lift the box themselves.

1.4.1 Materials and Methods

Participants

We recruited 50 participants from Amazon Mechanical Turk. Participants' demographic information was not collected. Participants completed a comprehension check before they moved on

to the experiment. They were not allowed to proceed until they answered all the comprehension check questions correctly. Participants received a base pay of $2 and a potential bonus payment up to $1. The amount of bonus they received was equal to the probability they put on the realized lift outcome on a randomly picked round. We chose this bonus scheme because it is incentive compatible: Expected bonus is maximized by reporting their best estimate of the lift probability. To ensure data quality, we included two attention check questions in the experiment. Participants who failed one attention check were warned immediately to pay closer attention. Participants who failed both attention checks were asked to leave the experiment and they were not counted towards the 50 participants we recruited. A total of 10 participants failed one attention check, and we did not exclude their data in our analysis. The Harvard Institutional Review Board approved the experimental procedures and participants provided informed consent prior to the experiment.

Procedure

Participants observed six contests between different pairs of contestants (see Table 1.2 for a description of the contests; the order was randomized). In each contest, the contestants were given three attempts to lift a box, corresponding to three rounds. In the first two rounds, the contestants tried lifting the box themselves. The reward for lifting the box was $10 in Round 1 and $20 in Round 2. In the third round of each contest, the two contestants tried to lift the box together for a reward of $20 for each. Participants first saw the lift outcome of Round 1 and made strength judgments (1-10; 1 means extremely weak and 10 means extremely strong) and effort judgments (0-100%) for each contestant. For Round 2 and Round 3, they predicted the probability of the contestants lifting the box (0-100%) before seeing the outcome, then observed the actual outcome and made strength and effort judgments. Note that participants made effort judgments only when the outcome was Lift. Participants were informed that the weight of the box was always the same and equivalent to a strength of 5 (i.e., an average contestant with strength 5 exerting all of their effort would be able to

lift the box). Participants also saw a table of all the previous outcomes when making their guesses. Figure 1.2 shows an illustration of the task.

Table 1.2: Lift outcome Experiment 1 scenarios.

Scenario	Agent	Round 1	Round 2	Round 3
F,F;F,F	A	Fail	Fail	Fail
	B	Fail	Fail	
F,F;F,L	A	Fail	Fail	Fail
	B	Fail	Lift	
F,L;F,L	A	Fail	Fail	Lift
	B	Lift	Lift	
F,F;L,L	A	Fail	Lift	Lift
	B	Fail	Lift	
F,L;L,L	A	Fail	Lift	Lift
	B	Lift	Lift	
L,L;L,L	A	Lift	Lift	Lift
	B	Lift	Lift	

Note: We re-organized the data such that Agent A is the weaker contestant and Agent B is the stronger contestant in each contest when they had different outcomes. The side was randomized in the experiment.

1.4.2 Results

As a preliminary check, we plotted participants' strength judgments on top of the model predictions (Figure 1.3A). The model predictions were similar to the behavioral data overall, except that the compensatory effort model and maximum effort model could not generate predictions in a few scenarios and rounds. The compensatory effort model could not predict contestants' strength in Round 3 of 'F,L;F,L', 'F,F;L,L', 'F,L;L,L', and 'L,L;L,L', since it predicts that the Round 3 outcome of these scenarios should always be Fail, but the observation is Lift. The maximum effort model could not make predictions for 'F,F;F,L', 'F,F;L,L', and 'F,L;L,L', simply because if contestants always exert 100% of their effort, then they would not fail in Round 1 and lift the box in Round 2.

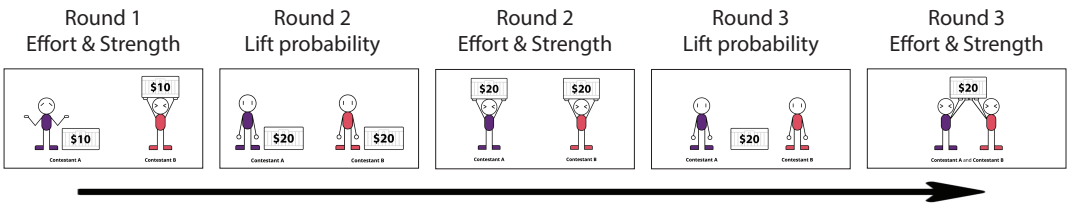


Figure 1.2: Example contest in Experiment 1. Each contest consists of three rounds. In Round 1, participants observed individual lift outcomes and reported their judgments of contestants' effort and strength. In Round 2, participants first guessed the lift probability when reward is increased from $10 to $20, then reported effort and strength judgments after observing Round 2 outcomes. In Round 3, participants guessed the probability of the two contestants lifting the box together, and reported effort and strength judgments of each contestant after observing the Round 3 outcome. At all times, participants saw a table of previous lift outcomes in that contest.

Note that the solitary effort model's predictions in all the scenarios except 'F,F;F,F' are disguised by that of joint effort model. Both the joint effort model and solitary effort model fail to make predictions for strength in Round 3 of 'F,F;F,L', as they both predict that the outcome should be Lift, conflicting the observation.

One major hypothesis concerns how effort judgments change from Round 2 to Round 3. If participants believed that contestants put in less effort when they worked together to lift the box in Round 3, compared to trying to lift the box by themselves in Round 2, then we should see a decrease in effort from Round 2 to Round 3. We selected trials from scenarios 'F,F;L,L', 'F,L;L,L', and 'L,L;L,L' where participants reported Round 2 effort judgments for both contestants, excluded Round 1 trials, and ran a linear mixed-effects regression regressing Effort on Round and Agent, with random effects for the intercept, Round, and Agent grouped by participants. Indeed, the regression results revealed that effort decreased from Round 2 to Round 3 [$t(49.0) = -5.160, p < .0001$]. Figure 1.3B visualizes this effect and shows that only the joint effort model makes this prediction. The solitary effort model predicts that effort should not change from Round 2 to Round 3, when contestants switched from lifting the box themselves to lifting together. The compensatory effort model could not predict contestants' effort in Round 3 and the maximum effort model could not

make predictions for 'F,F;L,L', and 'F,L;L,L', as explained above.

Another hypothesis concerns the lift probability in Round 3. The lift probability should increase as contestants get stronger (moving from left to right along the x-axis). The lift probability should also be pretty high when at least one contestant had a successful lift, that is, all the scenarios except for 'F,F;F,F'. However, in 'F,F;F,F' when both contestants failed in both rounds, the lift probability should be non-zero, given that the two contestants were attempting to lift the box together. One-sample t-test showed that participants believed that the lift probability for 'F,F;F,F' was different from zero [$t(49) = 10.42, p < .0001$]. Figure 1.3C confirms these hypotheses and shows that the joint effort model is the only model that exhibits these patterns. The solitary effort model predicts that the lift probability is zero in 'F,F;F,F'. Again, the maximum effort model could not make predictions for 'F,F;F,L', 'F,F;L,L', and 'F,L;L,L'.

Figure 1.4 compares model predictions to participants' judgments regarding the lift probability, effort, and strength. Overall, the joint effort model provides the best fit to the behavioral data. Aside from missing predictions in certain scenarios and rounds (the compensatory effort model could not predict effort and strength in Round 3 for any scenarios except 'F,F;F,F', and the maximum effort model could not make predictions for scenarios 'F,F;F,L', 'F,F;L,L', and 'F,L;L,L'), the compensatory effort model failed to predict the lift probability. The solitary effort model and the maximum effort model failed for effort judgments.

1.4.3 Discussion

In Experiment 1, we studied participants' judgments of strength, effort, and lift probability in both individual lifting and collaborative lifting, based on observations of previous lift events. The joint effort model provides predictions that are qualitatively similar to the behavioral data, including the decrease in effort from Round 2 to Round 3, the non-zero lift probability in Round 3 of scenario 'F,F;F,F', and the high lift probability in Round 3 of all the other scenarios. The joint effort model

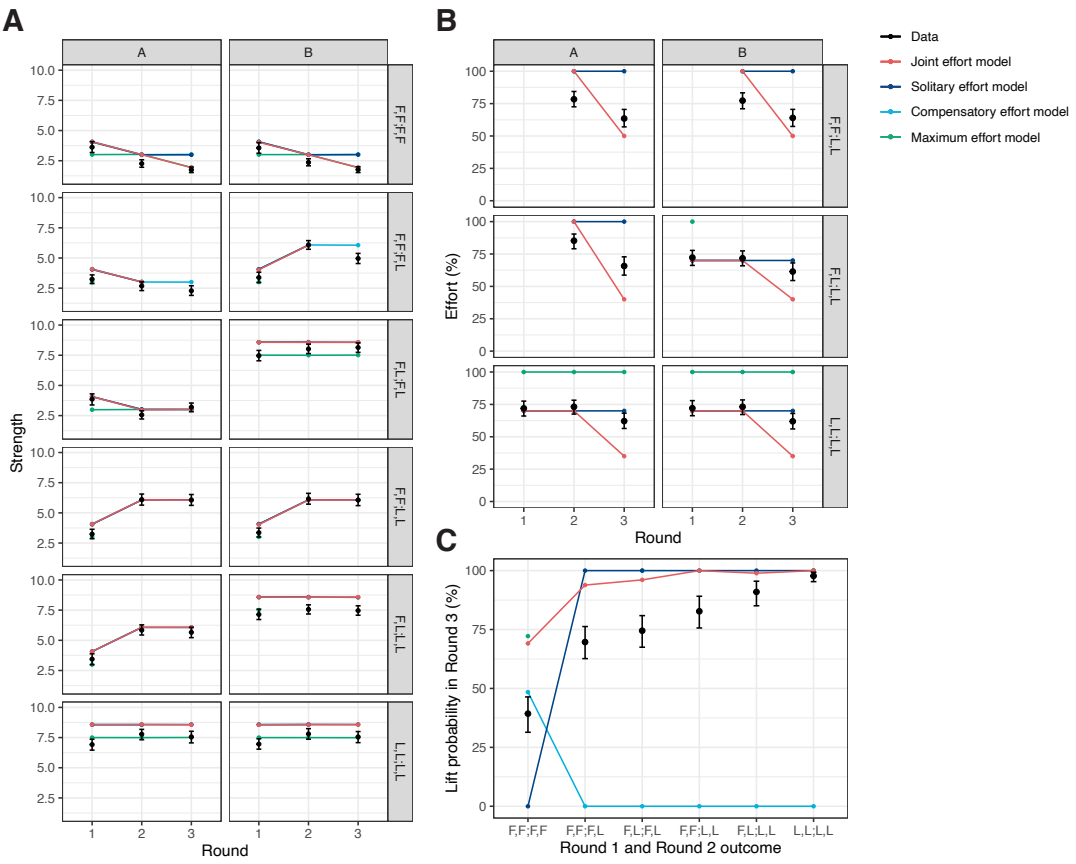


Figure 1.3: (A) Predictions of contestants' strength in Experiment 1. (B) Predictions of contestants' effort in Experiment 1. Effort judgments were not elicited when the lift outcome was Fail. (C) Predictions of the lift probability in Round 3 of Experiment 1. Model simulations averaged over 10 runs. Error bars indicate bootstrapped 95% confidence intervals.

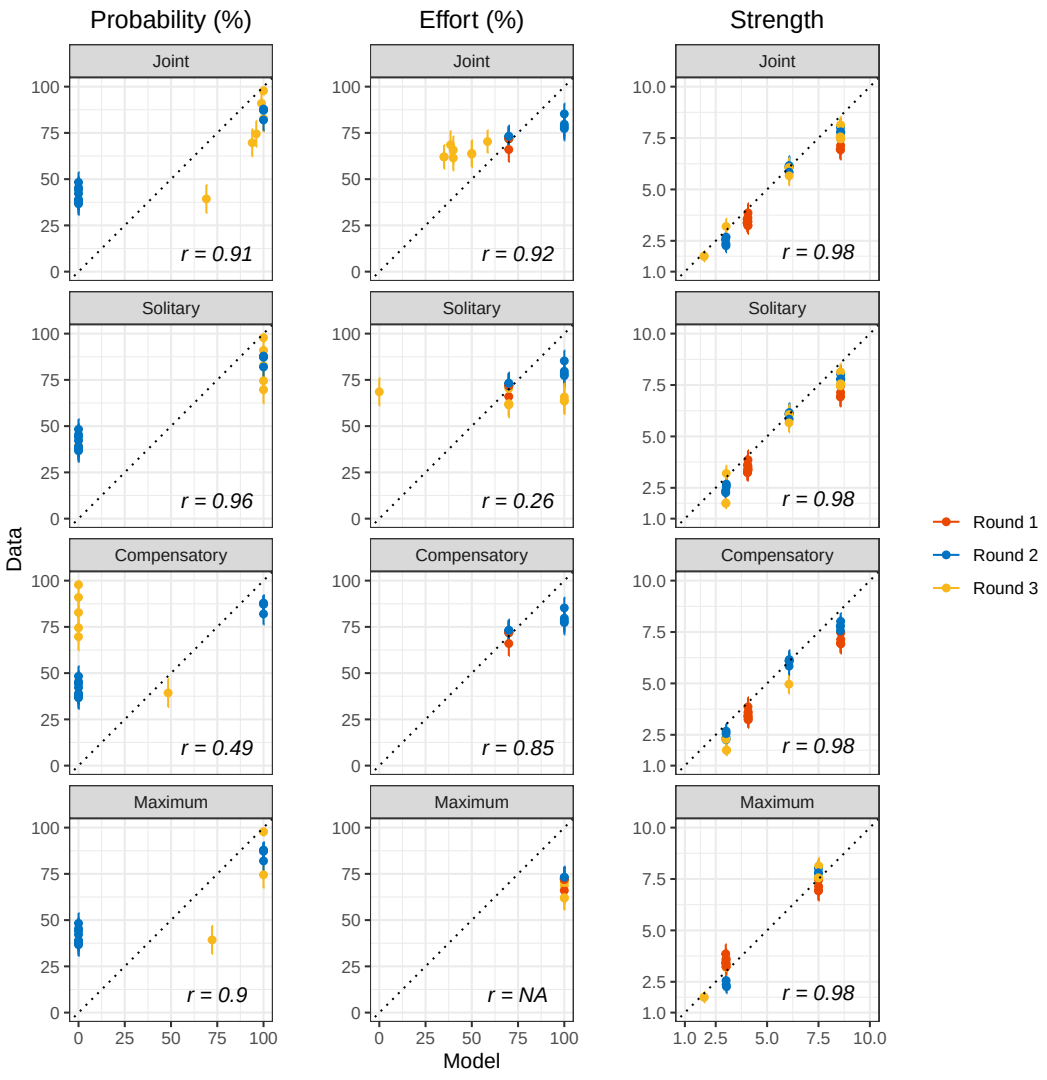


Figure 1.4: Comparing model predictions to behavioral data across predictions of lift probability, effort, and strength in Experiment 1. Model simulations averaged over 10 runs. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

also provides the overall best fit, compared to the alternative models. We conclude that people employ an intuitive theory of beliefs, desires, and competences, to reason recursively about joint effort.

1.5 Experiment 2

To test our theory's flexibility, in Experiment 2, we ask how people assign incentives to teams, based on knowledge about their strength. To differentiate between the models, we designed the experiment such that across multiple rounds in each contest, the strongest contestant is always involved in the group lifting. We also made sure that all the contestants in this experiment are strong enough to lift the box by themselves.

We expect qualitatively different predictions from the models. The joint effort model predicts a decrease in incentive when the contestants are stronger. The solitary effort model predicts no change in incentive across different rounds, because it assumes that the stronger contestant is effectively the one who determines the lower bound of the incentive required, and since that contestant is not changing across rounds in a contest, the incentive should not change either. The compensatory effort model and maximum effort model both predict zero incentive, though for different reasons. The compensatory effort model predicts that since every contestant is strong enough to lift the box by themselves, two contestants would never lift the box together successfully, thus no incentive should be wasted on this impossible mission. The maximum effort model, in contrast, predicts that the contestants would lift the box regardless of the incentive, assuming that everyone would always be exerting 100% of their effort. Therefore, $0 incentive would be the best choice.

1.5.1 Materials and Methods

Participants

We recruited 98 participants from Amazon Mechanical Turk. Participants' demographic information was not collected. Same as in Experiment 1, participants completed a comprehension check before starting the main experiment. Participants received a base pay of $0.5 and a potential bonus payment up to $6. The bonus payment depended on the incentive participants provided to the contestants and whether the lifting turned out successful given the incentive. Specifically, for every successful lift, participants received $0.4; for every dollar of incentive they gave out, $0.002 was deducted from their bonus payment, regardless of the lift outcome. ** The Harvard Institutional Review Board approved the experimental procedures and participants provided informed consent prior to the experiment.

Procedure

Participants observed five different contests between different teams of contestants. In each contest, participants saw four different contestants and the minimum incentive each would accept to lift the box alone as a reference point. To give participants a better sense of what the minimum incentives meant, we converted the minimum incentive each contestant required to an estimate of their strength (see Figure 1.5 for an illustration of the task). In each round of the contest, the strongest contestant was always among the pair of contestants attempting to lift the box, while the other contestant was one of the remaining contestants. We intentionally included a contestant (Contestant D) who is as strong as the strongest contestant (Contestant A). This is when the models (specifi-

**Note that we did not use attention checks in Experiments 2 and 3. Our results from Experiment 1 suggest that participants' overall performance on the attention checks was high (3 participants failed the first attention check, 7 participants failed the second attention check), and that the interpretation of the results does not change regardless of whether we exclude participants who failed one attention check.

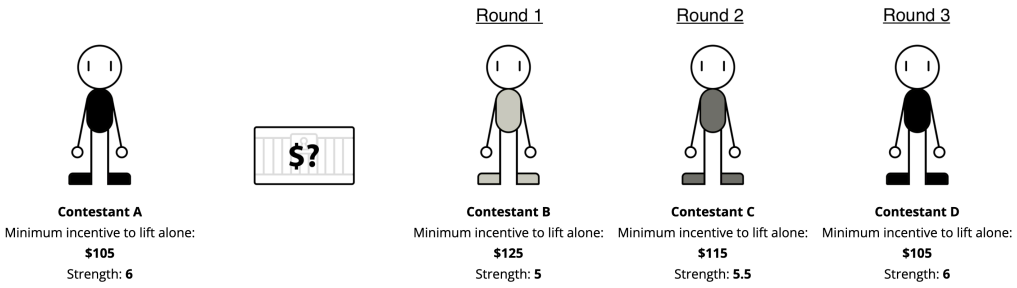


Figure 1.5: An example contest in Experiment 2. Each contest contains three rounds, and in each round, the strongest contestant (to the left) attempts to lift a box with one of the three contestants to the right. Participants reported how much incentive they were willing to provide to each pair of contestants.

cally, the joint effort model and the solitary effort model) make the most distinct incentive predictions quantitatively: The solitary effort model's predictions do not change as long as Contestant D is not stronger than Contestant A, whereas the joint effort model's predictions decrease when the teammate gets stronger. This difference in prediction is the largest when the teammate is as strong as the strongest contestant. Participants decided how much incentive to provide each pair of contestants to lift the box together. They were told that the same incentive would go to both contestants, i.e., if participants provided $50 to a pair of contestants, then both contestants would get $50 if they lifted the box together successfully. The box weight was fixed at 5, as in Experiment 1. Participants also completed three training trials where they tested how a random contestant would respond to different incentives.

1.5.2 Results

To test the hypothesis that the incentive will decrease when one teammate becomes stronger, we constructed a linear mixed-effects model, in which we regressed Incentive on the strength of the two contestants involved in the lifting, with participant-level random effects for all the regressors. We found that controlling for the strength of the stronger contestant who remained the same through-

out each contest, the strength of the weaker contestant had a statistically significant negative effect on the incentive provided by the participants [$t(1167.0) = -12.595, p < .0001$]. This result is visualized in Figure 1.6A.

Both the compensatory effort and maximum effort models predict that the optimal incentive should be $0 across all rounds and contests. The solitary effort model predicts that the incentive provided to the contestants should remain the same when the stronger contestant was unchanged, as shown in Figure 1.6A. Only the joint effort model showed the trend of incentive decreasing within each contest. This is also validated by the high Pearson correlation coefficient ($r = 0.95$; Figure 1.6B).

1.5.3 Discussion

Experiment 2 is the application of our theory to the problem of incentive selection. With a task that controlled for the stronger contestant's strength, the joint effort model showed qualitatively similar predictions to the behavioral data, while all the other models made distinct predictions regarding how the incentive should change as a function of teammate strength.

1.6 Experiment 3

In Experiment 3, we further extend the model to the problem of team selection. Besides manipulating contestants' strength, we manipulated their effort cost (the α parameter in our model). We also designed different box weights such that weaker but more hard-working contestants and stronger but lazier contestants are favored differently: Weaker but more hard-working contestants should be preferred when the box is light, due to their lower effort cost. In contrast, stronger but lazier contestants should be preferred when the box is heavy because weaker contestant would not be able to lift a box that is too heavy for them no matter how hard they try. In addition, if we allow contestant-

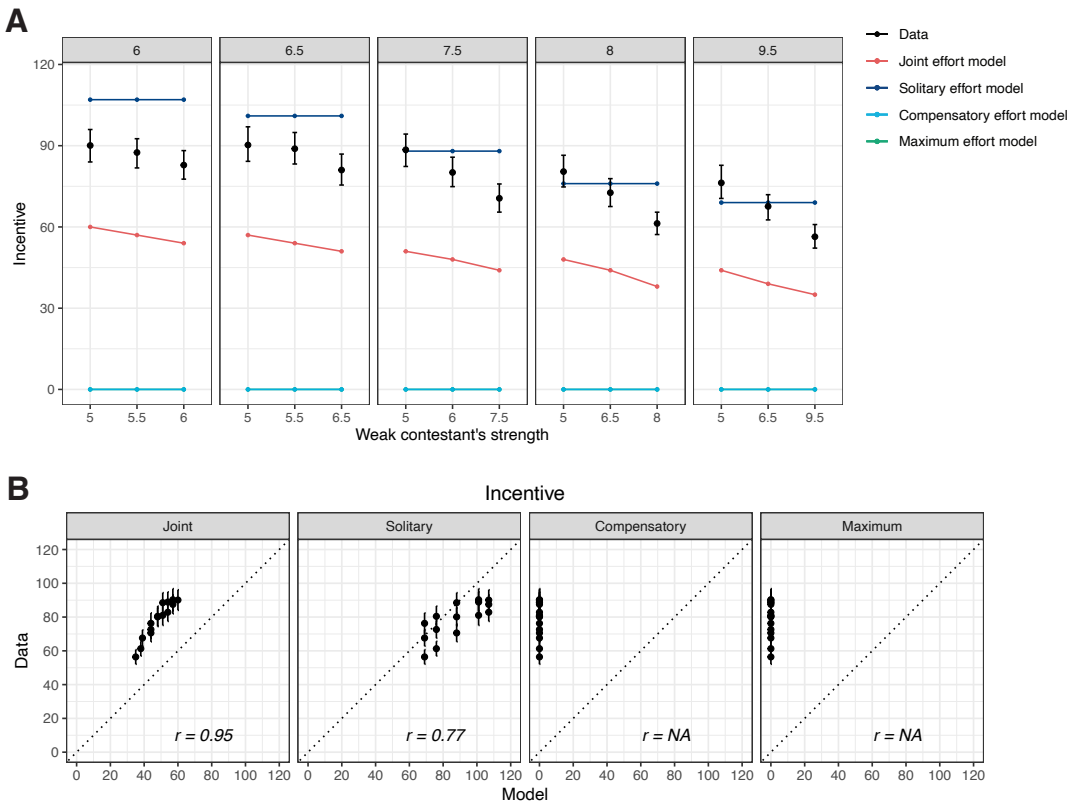


Figure 1.6: (A) Incentive provided to pairs of contestants in Experiment 2. Each panel corresponds to the strongest contestant's strength in each contest. (B) Comparison between behavioral data and model predictions of incentive in Experiment 2. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars in (A) indicate bootstrapped 95% confidence intervals; error bars in (B) indicate 95% normal confidence intervals.

specific rewards (i.e., different contestants on the team can receive different rewards) then we should see that the incentive allocated to each contestant correlates with how often they are selected.

The models generate distinct predictions regarding who to select. The joint effort model follows a similar logic as above. The solitary effort model will always pick contestants who are able to lift the box themselves, and if forced to include another contestant on the team, it will not expect the other contestant to exert any effort. The compensatory effort model only considers a group lifting possible when the sum of the contestants' strength equal the box weight (i.e., when both contestants exert all their effort). Otherwise, it assumes that the group lifting is impossible anyway and therefore values

every contestant equally. The maximum effort model makes similar predictions as the joint effort model in terms of who to select, but when more than one group of contestants are able to succeed in lifting, it values them all equally, since it doesn't take into account their effort costs; all it cares is each contestant's strength. These differences should also be reflected in the incentives assigned to each contestant.

1.6.1 Materials and Methods

Participants

We recruited 101 participants from Amazon Mechanical Turk. Participants' demographic information was not collected. Same as in Experiment 1 and Experiment 2, participants completed a comprehension check before starting the main experiment. Participants received a base pay of $0.5 and a potential bonus payment up to $3. The bonus payment depended on the incentive participants provided to the contestants and whether the lifting was successful given the incentive: For every successful lift, participants received $1; for every dollar of incentive they gave out, $0.01 was deducted from their bonus payment, regardless of the lift outcome. The Harvard Institutional Review Board approved the experimental procedures and participants provided informed consent prior to the experiment.

Procedure

Participants observed three different contests between three contestants who had different strength and laziness. Participants knew each contestant's strength and the minimum incentive each would accept to lift the heaviest box they could (box weight equivalent to their strength), as shown in Figure 1.7A. This provides participants information about each agent's effort cost, which equals the minimum incentive they would accept to lift the heaviest box they could (i.e., minimum incentive

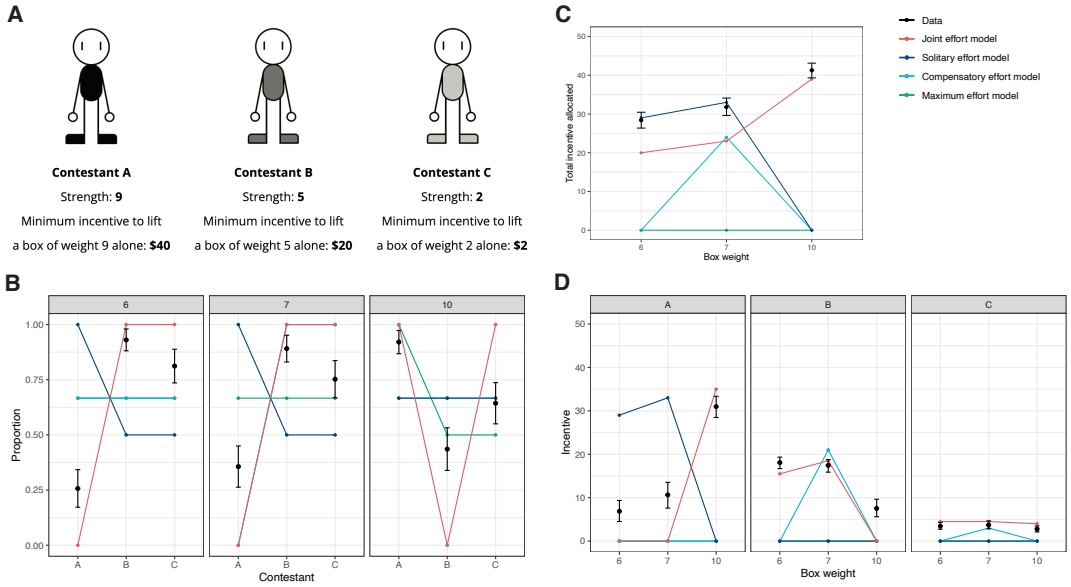


Figure 1.7: (A) Contestants in Experiment 3. (B) Proportion of times each contestant was selected for each contest in Experiment 3. Contests are identified by their box weights, which appear above each plot. (C) Total incentive provided to contestants in Experiment 3. (D) Incentive provided to each contestant in Experiment 3. Each plot corresponds to an individual contestant. Error bars in (B) indicate 95% confidence intervals of proportions; error bars in (C) and (D) indicate bootstrapped 95% confidence intervals.

is equivalent to $\alpha \cdot 100\%$). The box weight varied from contest to contest, ranging from 10 to 7 to 6. In each contest, participants first selected two contestants to do the lifting together, then decided the incentive they would provide to each contestant. Note that in this experiment, each contestant's incentive was contestant-specific. In other words, instead of both contestants receiving the same incentive, participants chose a separate incentive for each contestants. We constrained participants' budget in each contest to $\$50$ (i.e., the total incentive they gave out in each contest could not exceed $\$50$). Participants also completed six training trials where they tested how a random contestant would respond to different incentives given two different box weights.

1.6.2 Results

When the box weight was 10, the joint effort model predicts that Contestant A will always be selected, because without Contestant A, Contestant B and Contestant C's total strength was smaller than the box weight. As for the other teammate, the joint effort model predicts that Contestant C will be preferred over Contestant B, because selecting Contestant B would entail a higher effort cost. When the box weight was 6 or 7, however, the joint effort model predicts that Contestant B and Contestant C will be selected because their total strength was equal to or exceeded the box weight, and their effort would cost less than Contestant A. This was indeed what we saw from the behavioral data (Figure 1.7B): When the box weight was 10, 92.08% of the participants ($n = 93$) selected Contestant A. More participants selected Contestant C (64.36% of the participants, $n = 65$) over Contestant B (43.56% of the participants, $n = 44$). When the box weight was 7, only 35.64% of the participants ($n = 36$) selected Contestant A. 89.11% of the participants ($n = 90$) selected Contestant B, and 75.25% of the participants ($n = 76$) selected Contestant C. When the box weight was 6, only 25.74% of the participants ($n = 26$) selected Contestant A. 93.07% of the participants ($n = 94$) selected Contestant B, and 81.19% of the participants ($n = 82$) selected Contestant C. As shown in Figure 1.7B, the joint effort model is the only model that shows a pattern qualitatively similar to the behavioral data. Figure 1.8 validates this conclusion with a Pearson correlation coefficient of 0.91 between the behavioral data and the joint effort model's predictions.

We also looked at how the total incentive and individual incentive changed with box weight. If the participants believed that the box was liftable in all three contests, then the total incentive should increase with box weight. As expected, we saw a monotonic increase in the total incentive allocated to contestants as the box weight increased (Figure 1.7C), confirmed by a linear mixed-effects regression which regressed the total incentive on the box weight, with random effects of the intercept and the box weight grouped by participants [$t(201.0) = 12.232, p < .0001$].

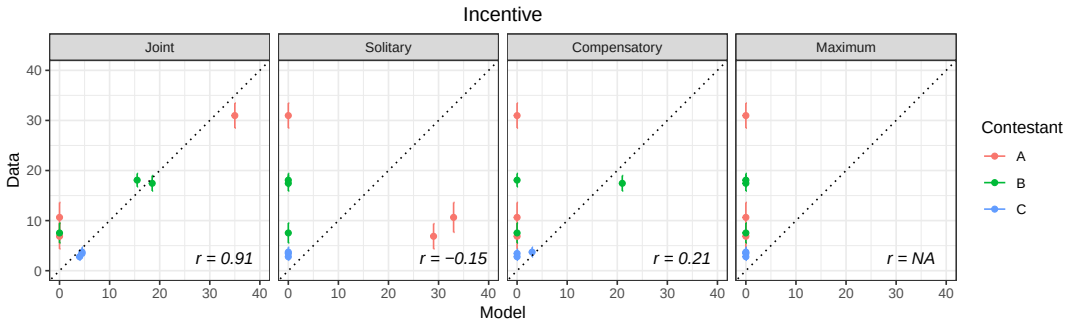


Figure 1.8: Comparison between behavioral data and model predictions of incentive in Experiment 3. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

The joint effort model predicts that the incentive allocated to each contestant should change according to how likely they are to be selected. Across the three contests, we should see an overall trend of Contestant A's incentive increasing with box weight, and Contestant B and Contestant C's incentive decreasing with box weight. There should not be a big difference between box weight of 6 and box weight of 7, so the transition from box weight of 7 to box weight of 10 should primarily determine the trend. From three separate linear mixed-effects regressions constructed for the three contestants, each of which regressed the incentive allocated to the contestant on the box weight, with random effects of the intercept and box weight grouped by participants, we found that the incentive for Contestant A increased with box weight [$t(201.0) = 14.801, p < .0001$], incentive for Contestant B decreased with box weight [$t(229.2) = -10.56, p < .0001$], and incentive for Contestant C decreased with box weight [$t(201.0) = -1.910, p = .058$]. The incentive for Contestant C showed a non-significant trend towards decreasing with box weight, likely because Contestant C's minimum incentive is too low to allow for sufficient variation. These effects are visualized in Figure 1.7D.

To confirm these results, we directly tested the correlation between the incentive allocated to each contestant and whether they were selected. From three separate linear mixed-effects regressions (one for each contestant) with incentive as the dependent variable and a categorical variable that indicates

whether the contestant was selected being the predictor variable, along with random effects of the intercept and the categorical variable grouped by participants, we found that incentive increased when the contestant was selected. This effect was significant for all three contestants: Contestant A [$t(105.8) = 35.15, p < .0001$], Contestant B [$t(176.9) = 31.82, p < .0001$], and Contestant C [$t(102.1) = 9.544, p < .0001$].

1.6.3 Discussion

Experiment 3 asked participants to make judgments regarding team selection: who to select and how much incentive for each teammate. As predicted by the joint effort model, participants cared about the total strength of the contestants when deciding whether they would be able to lift the box together, therefore favored the strongest contestant (Contestant A) when the box was very heavy (box weight = 10). When the box was lighter (box weight = 6 or 7), however, participants favored contestants who were more hard-working, which entailed lower effort cost (Contestants B and C). The total incentive increased with box weight and the incentive participants provided to each contestant corresponded to their selection of teammates. For all three judgments, the joint effort model's predictions were qualitatively similar to the behavioral data, while the other three models all made predictions that deviated systematically from the data.

From Figure 1.7B, we see that humans exhibited a preference for certain contestants over others; for example, when the box weight was 6 and 7, some participants selected Contestant A over Contestant C, thereby showing a clear preference for Contestant B over Contestant C. This pattern was not observed in the joint effort model's predictions. It is worth noting, however, that there is an important methodological difference between Experiment 3 and Experiment 1. In Experiment 3, each contestant's strength is determined by the experimental setup; therefore the model does not infer contestants' strength, which makes it impossible for the joint effort model to choose probabilistically. In other words, the joint effort model cannot choose another partner because of the way

that the decision rule is set up. It is possible that, with a different decision rule, we could bring the quantitative pattern of the data arbitrarily close to human judgments, but that would not change the substance of the model. The important point is that the joint effort model captures human intuitions about which collaborative partner would be preferred.

1.7 General Discussion

We presented a Collaborative Utility Calculus framework which extends theory of mind reasoning to include an agent-specific representation of competence and effort. We applied this framework to collaborative cognition, where multiple agents work together in order to achieve goals that may be impossible for any individual. To succeed in these collaborative tasks, we argued that agents must know what other agents are capable of doing and how much effort they are willing to exert.

Through three studies, we demonstrated that people make judgments about competence, effort, incentives, and team structure that are in broad agreement with our framework. In Experiment 1, we found that people can infer others' competence through observations of behavior. More importantly, people can generalize individual agents' behavior to multi-agent contexts and make inferences about group behavior, including predicting whether a collaboration will be successful, and inferring the amount of effort collaborators exert, without any direct observation. In Experiment 2, we found that people assigned incentives to teams based on considerations of their strength, and in Experiment 3, we found that people allocated individuals to teams and selected incentives for them according to agent-specific attributes and agent-general tasks. Taken together, we showed that people are capable of making inferences regarding collaborative behavior and applying these inferences to tackle novel problems of incentive selection and team selection.

We showed that human judgments of collaborative behavior were well-predicted by a multi-agent joint effort model grounded in recursive reasoning of effort between agents. Alternative models

that did not invoke recursive reasoning failed to capture the data, suggesting that recursive theory of mind plays a central role in human collaboration.

While the joint effort model qualitatively resembled the data, people reported lower probabilities, greater effort, and higher incentives compared to the joint effort model. This raises the possibility that people are risk-averse, erring on the side of safety by preferring a slightly lower bonus over no bonus. We explored this possibility by creating a *safe joint effort model*, in which a hindrance factor is added to hedge the model's predictions, as we speculate that participants are doing. Instead of satisfying Equation 1.4, the safe joint effort model requires that the sum of agents' effort times strength has to be greater than or equal to the box weight plus this additional hindrance factor. This model yields predictions that are quantitatively closer to human judgments. Since the predictions of this safe joint effort model are qualitatively similar to the original joint effort model, we decided for exposition purposes to keep the simpler model in the main text but include the elaborated model in the supplemental material (Figures S1-S5). The safe joint effort model is also able to offer a plausible explanation regarding why participants in Experiment 1 were able to make strength judgments in Round 3 of scenario 'F,F;F,L' while the joint effort model could not: When the hindrance factor is large (greater than 1.5), the safe joint effort model is able to make predictions for this scenario.

Our work complements past research on collaboration. Evolutionary theories of collaboration address how collaboration could have emerged as evolutionarily stable strategies at the population level[101]. Economic theories of collaboration are concerned with the structure of the task: under what combinations of incentives agents could be expected to cooperate[148]. The machine learning literature on collaboration focuses on how agents carry out sub-tasks in cooperative multi-agents systems e.g.,[171]. Our work provides a different perspective on collaboration, grounded in a formal description of the cognitive capacities that enable individuals to coordinate collaborative actions, and validated through tasks with rich psychological structure.

One limitation of our work is that we constrained the experimental setup to teams of two agents;

in principle, however, our framework can generate predictions for larger teams. Moving forward, our framework may shed light not only on how humans reason about beliefs, desires, competence, and effort in larger collaborations, but also what the boundaries of these reasoning abilities may be. From a computational perspective, we might expect agents to replace optimal recursive reasoning (which is intractable for large groups) with simpler heuristics e.g.[81]. When viewed through a resource-rational lens[65,141], the heuristics used by the alternative models are plausible "shortcuts" to joint inference, and these models may be useful in diagnosing why collaborations sometimes fail. An open question for future work is how cognitively bounded groups of agents can realize resource-rational collaboration by optimally allocating their cognitive resources.

Understanding how well these inferences scale to larger teams—and when they begin to break down—may also provide a cognitive perspective on the design of organizations: Organizations make social reasoning easier by breaking down a large, sprawling structure into smaller work teams or chains of command, thus limiting the number of individuals with whom any one worker has to coordinate. Understanding the limitations of human social reasoning may inform the design of organizations and administrative structures that are adapted to those limitations[129].

Another limitation of our work is that we operationalized competence as physical strength. Moreover, we simplified our analysis by treating strength as a static, scalar quantity. In real-world problems, competence is instead multidimensional and dynamic. Even within the domain of physical strength, one might be able to lift more weight during deadlifts than during chest presses, and the heaviest weight that one can lift may increase with consistent practice. Moving beyond the physical domain, real-world collaborations unite the efforts of people with varying cognitive skills and expertise, and the most relevant expertise that one has to offer may depend on the composition of the rest of the group—for example, an individual researcher with expertise in both computational modeling and experimental design may tackle the computational modeling in a team composed mostly of experimentalists, and implement experiments in a team of mostly theorists. Thus, one intrigu-

ing extension of our framework would be to allow agents to have varying degrees of competence in varying domains, effectively operationalizing competence as a vector rather than as a scalar quantity. In doing so, our framework may then be extended to predict how teams divide labor based on the expertise of agents in different tasks.

Our framework can be extended to treat competence as a dynamic, rather than a static, property. Competence can both increase and decrease over long periods of work. On one hand, after performing the same task many times—such as lifting a heavy weight, or manufacturing pins, or building computational models—individuals may become more efficient, able to carry out the same work for less effort. Indeed, foundational economic accounts suggest that this increase in competence is one of the chief benefits of division of labor[206]: by specializing narrowly, individual workers can become more efficient at their particular task, thus outperforming generalists. On the other hand, there is also a downside to this repetition: Over time, an individual may grow bored with doing the same task, or carry out the same task less efficiently due to fatigue.

One particularly important problem for future work is understanding how skills are taught and transferred over long-term collaborations, such as in apprenticeships between craftspeople or in the mentoring relationship between an advisor and their student. Indeed, evolutionary accounts of teaching propose that teaching co-evolved with the use of complex tools, where more complex technical skills require more sophisticated, efficient means of transferring those skills between individuals[149,30]. However, little is known about how teachers select what skills to impart to their students, particularly in contexts where students can then hone these skills on their own through practice. Most laboratory experiments of teaching instead focus on how teachers select information that will change the mental states of a learner—such as what they believe[200] or value[105]—rather than their skills (but see Kleiman-Weiner et al.[125]).

1.8 Conclusion

In sum, we have provided evidence that collaborative decisions are scaffolded by recursive inferences about others' effort and competence. We presented a Collaborative Utility Calculus framework that captures qualitative patterns in human judgments across a variety of decisions that are critical for collaboration, such as predicting whether a joint task will succeed, deciding what incentives to provide to collaborators, and choosing whom to recruit for a collaborative task. Moving forward, important tasks for future work include understanding the limits of these inferences—such as how well they scale to larger groups—and building dynamic, multidimensional representations of competence into the model. Ultimately, our goal is to more completely formalize the rich cognitive capacities that support human collaboration.

2

Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks

2.1 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.

2.2 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 [191], and adults give a greater share of the blame for failed collaborations to collaborators who could have changed the outcome [2].

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 [251]. 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 [91].

Classical production-style theories attribute causation based on the force that one object or agent exerts on another. Dowe [54] 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 [254] 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 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 [85,166,243].

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 fam-

ily 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. [68] 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 [17,116]. Individuals also tend to be punished more if they fail for lack of effort, rather than lack of ability [250].

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 [147]. Chockler and Halpern [40] 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 [72,271].

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 mod-

els (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.

2.3 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.

Below we describe how each model computes responsibility based on a description of the lifting event. Table 2.1 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.

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

Table 2.1: 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**

2.3.1 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:

$$R_a^F \propto \begin{cases} F_a & \text{if } L = 1 \\ 10 - F_a & \text{if } L = 0 \end{cases} \tag{2.1}$$

- **Strength model**. The strength (S) model considers agents' strength as the metric for responsibility judgments. Gerstenberg et al. [68] 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:

$$R_a^S \propto S_a \tag{2.2}$$

- **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:

$$R_a^E \propto \begin{cases} E_a & \text{if } L = 1 \\ 1 - E_a & \text{if } L = 0 \end{cases} \tag{2.3}$$

2.3.2 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.[79] 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 [190], 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 [189].

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

$$P_a = \begin{cases} \sum_{E'_a} P(E'_a)\mathbb{I}[E'_a S_a + E_{/a} S_{/a} < W] & \text{if } L = 1 \\ \sum_{E'_a} P(E'_a)\mathbb{I}[E'_a S_a + E_{/a} S_{/a} \geq W] & \text{if } L = 0, \end{cases} \tag{2.4}$$

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:

$$R_a^{FA} \propto P_a \tag{2.5}$$

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:

$$R_a^{NFA} \propto 1 - P_{/a} \tag{2.6}$$

- **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:

$$R_a^{BA} \propto (R_a^{FA} + R_a^{NFA})/2 \tag{2.7}$$

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.

2.3.3 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:

$$R_a^{EBA} \propto wR_a^E + (1-w)R_a^{BA}, \tag{2.8}$$

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:

$$R_a^{EBA} \propto (R_a^E + R_a^{BA})/2 \tag{2.9}$$

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.

2.3.4 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

$$E'_A \times S_A + E_B \times S_B \geq W, \tag{2.10}$$

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 Equation 2.10, 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.

2.4 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 counterfactual-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.

2.4.1 Materials and Methods

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 minutes. Participants in Experiment 1b were paid $2.50 to complete 25 trials with an estimated completion time of 12 minutes. 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).

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 con-

tains 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).

Procedure

Figure 2.1 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 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

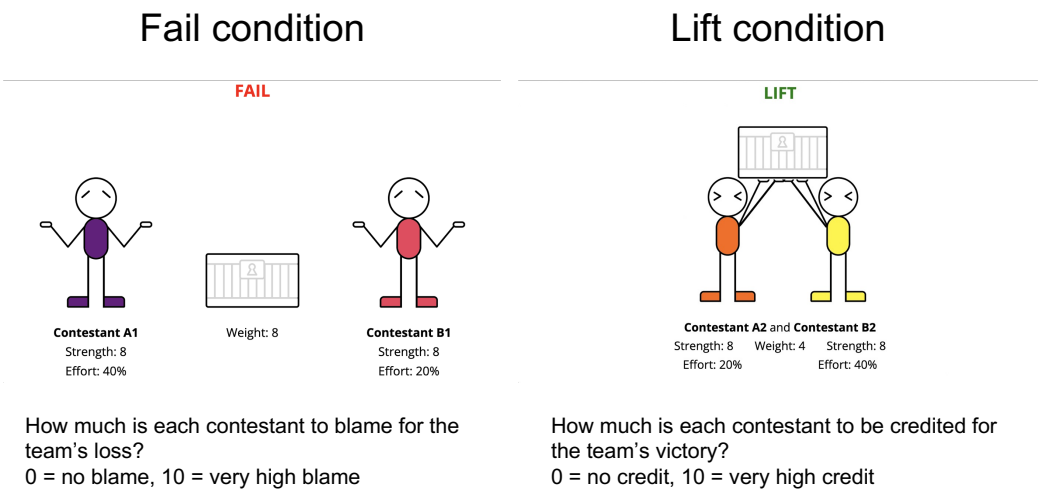


Figure 2.1: Experimental setup used in Experiments 1a, 1b, and 3. Participants observed each contestant's strength and effort, the weight of the box, and whether the contestants successfully lifted the box together (Lift condition) or not (Fail condition), and they then assigned blame or credit to each contestant. In Experiments 2a and 2b, participants also observed the force exerted by each contestant.

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.

2.4.2 Results

Participants made similar judgments in Experiments 1a and 1b, as shown in the leftmost columns of Figure 2.2 and Figure 2.3, 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 Figure 2.2 and Figure 2.3, 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 [$t(179.0) = -21.33, p < .0001$]. Figure 2.2 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.

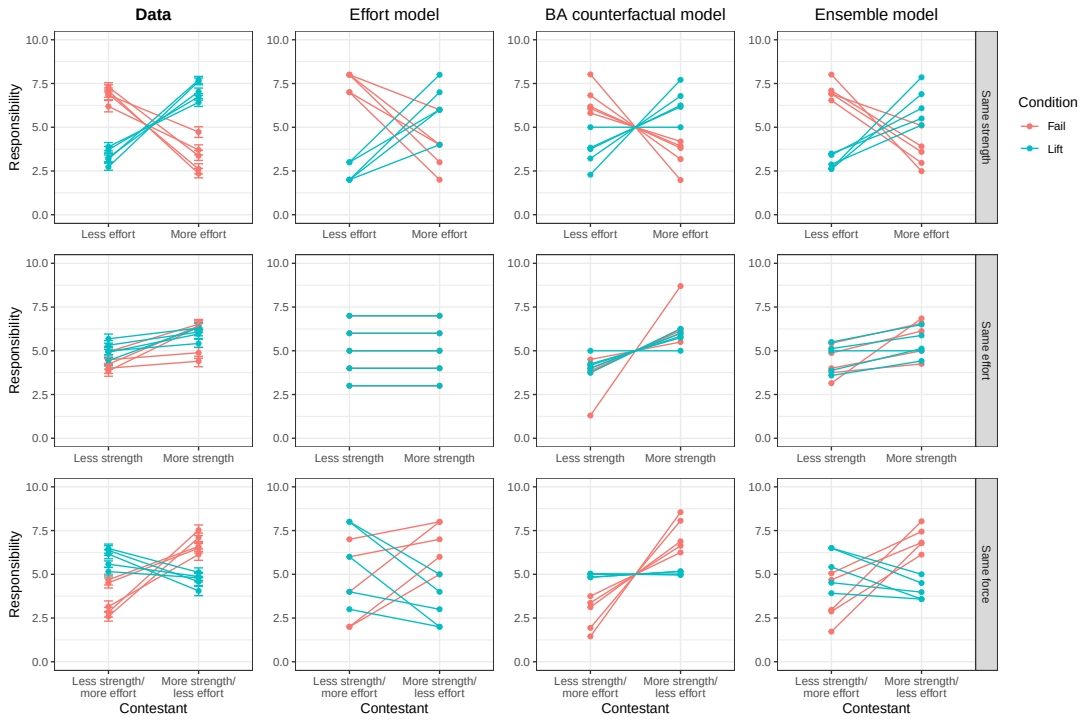


Figure 2.2: Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 1a. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

For every computational model, we fit a linear 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 Figure 2.4A, 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

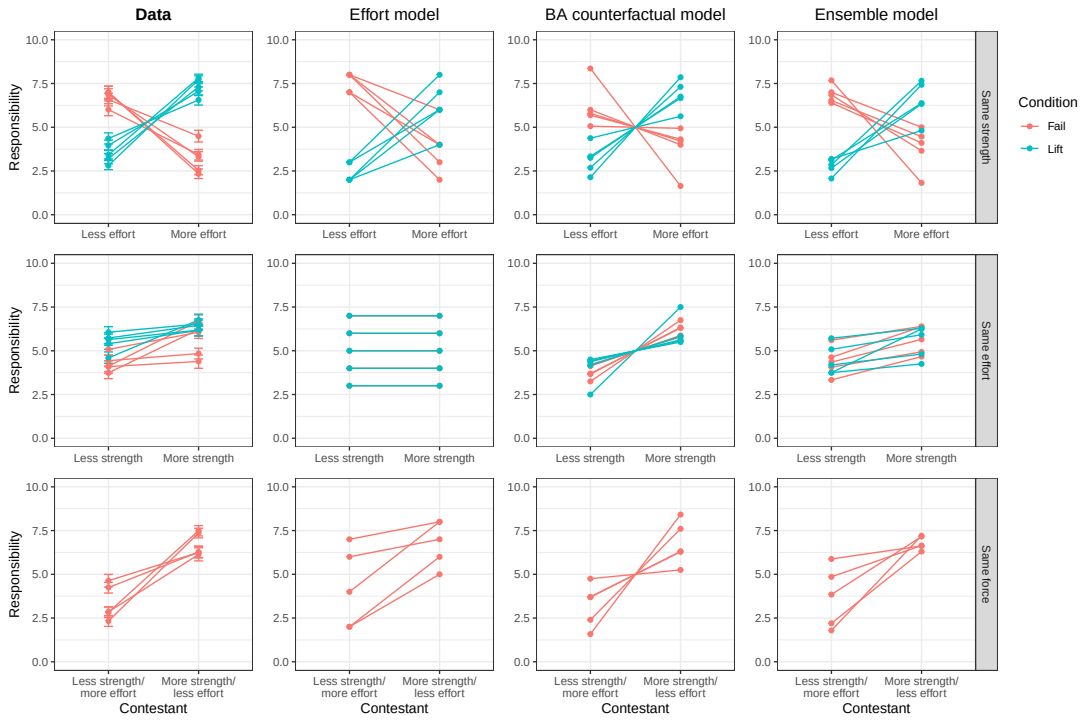


Figure 2.3: Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 1b. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

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'

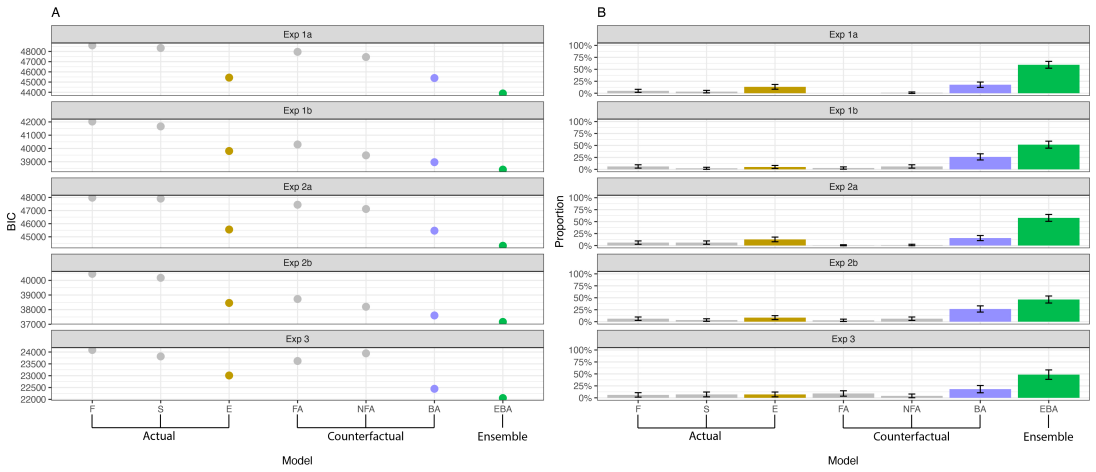


Figure 2.4: (A) Bayesian information criterion (BIC) for each mixed-effects regression model. (B) Proportion of participants best described by each model. Error bars indicate 95% confidence intervals of proportions. F = Force model, S = Strength model, E = Effort model, FA = Focal-agent-only counterfactual 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).

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.

Figure 2.2 and Figure 2.3 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.

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 Figure 2.4A). From Figure 2.2 and Figure 2.3, 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 Figure 2.5 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 Figure 2.4B, the Ensemble model explains the data of the largest number of participants (59.4% of the participants in Experiment 1a, 51.7% of the par-

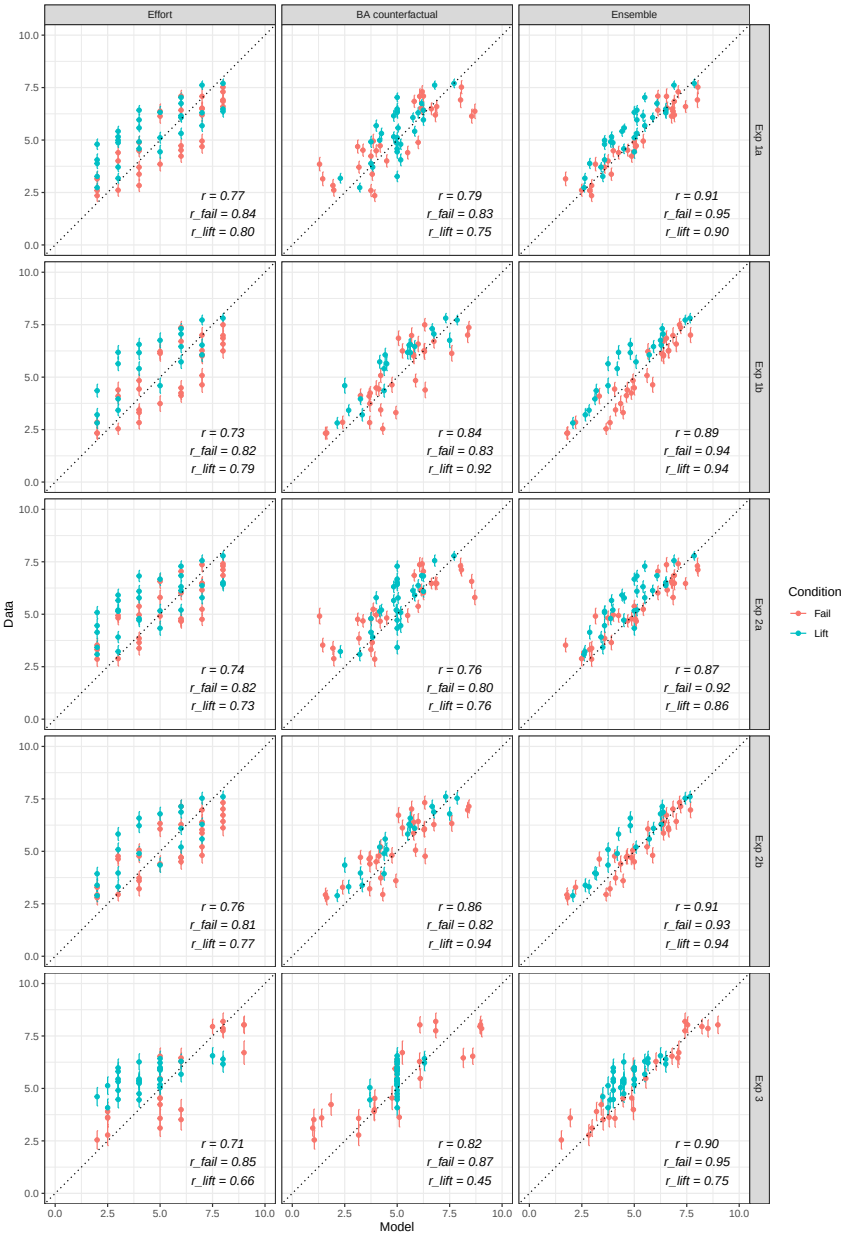


Figure 2.5: Comparison between behavioral data and model predictions. Pearson correlation coefficients for data and model predictions pooled across both conditions and for each condition are shown at the bottom right of each subplot. Each dot denotes one agent in one scenario; error bars indicate 95% confidence intervals.

ticipants 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.

2.4.3 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.

2.5 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.

2.5.1 Materials and Methods

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 minutes. Participants in Experiment 2b were paid $2.50 to complete 25 trials with an estimated completion time of 12 minutes. The experiments were approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/PNZ_2HW.

Stimuli

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

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).

2.5.2 Results

Participants' judgments in Experiments 2a and 2b were similar (see the leftmost columns of Figure 2.6 and Figure 2.7, 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 2b]. As shown in Figure 2.6 and Figure 2.7, 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 [$t(179.0) = -19.38, p < .0001$]. Figure 2.6 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 [$t(176.0) = 19.91, p < .0001$].

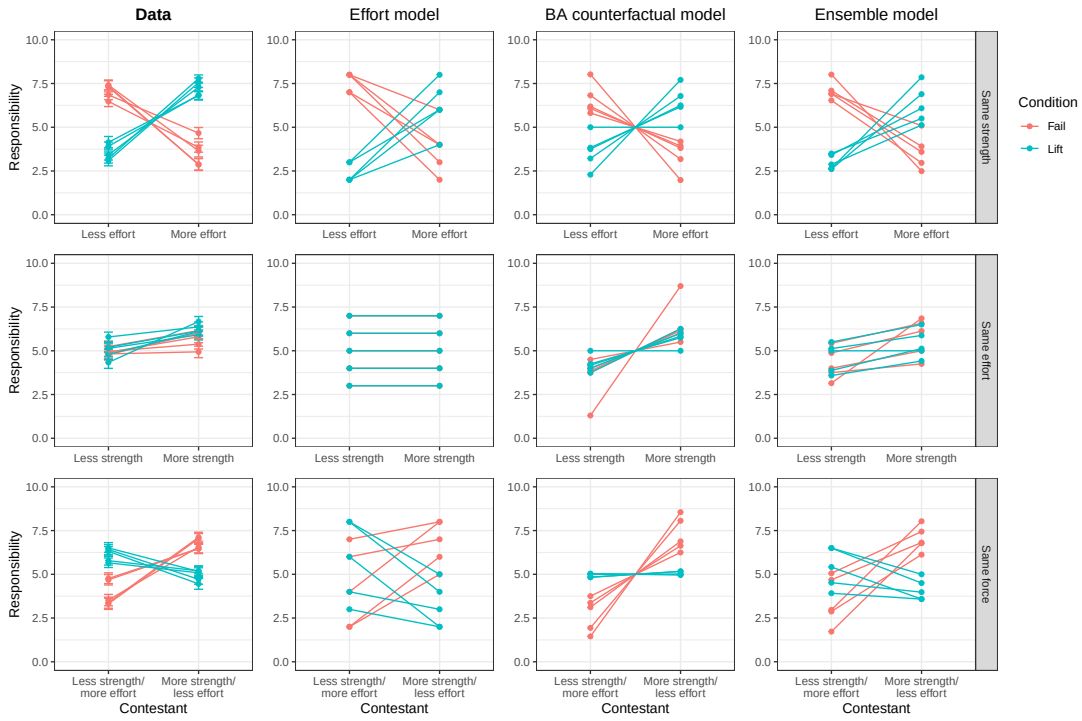


Figure 2.6: Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 2a. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

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 Figure 2.4A) 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,

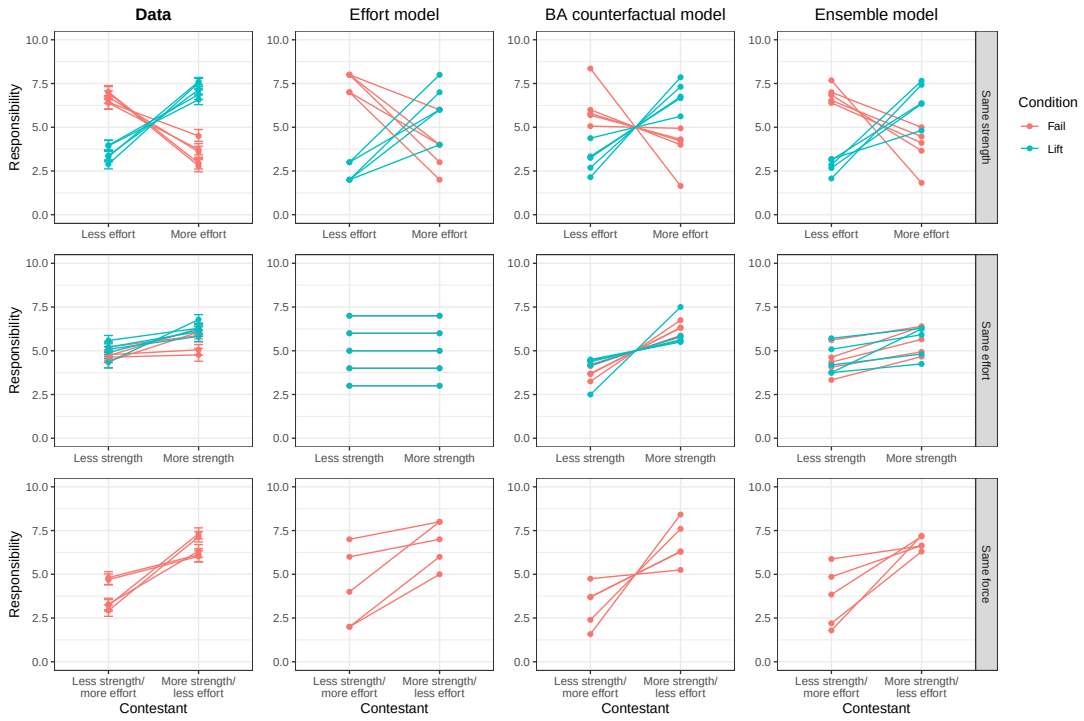


Figure 2.7: Data and predictions of the Effort model, Both-agent counterfactual model, and Ensemble model in Experiment 2b. Each line corresponds to a scenario. Error bars in the Data column indicate bootstrapped 95% confidence intervals.

$r = 0.76, p < .0001$ 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 Figure 2.5 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 Figure 2.4B).

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 $t(180.4) = -8.52, p < .0001$ 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].

2.5.3 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.

2.6 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.

2.6.1 Materials and Methods

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 minutes. The experiment was approved by the Harvard Institutional Review Board and preregistered at https://aspredicted.org/ZC3_VG4; we explain deviations from the preregistered analysis plan in the Supplement (see Deviations from pre-registrations).

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).

Procedure

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

2.6.2 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 Figure 2.4A) 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 Figure 2.5 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 [$t(101.7) = 11.63, p < .0001$].

2.6.3 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.

2.7 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[68], or a coordination game where fishers decide independently whether to catch fish or to clear the roads to go to market[2]. 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., [54,216,254]. 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[254]. 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 [267]. 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[87] or instrumental learning tasks[89], 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*[89]—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[145], and adults tend to judge more effortful prosocial actions as more praiseworthy[17], in part because effort signals the importance of the prosocial goal to the agent[4]. 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[75,131,210,228].

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[35]. 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[259]. 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 hadn't tripped[91]. 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[150,180]. For example, taking inspiration from noisy models of Newtonian physics[69], 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.

3

People reward others based on their willingness to exert effort

3.1 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 [266]. Past work has found that responsibility judgments are driven both by how much effort people actually contributed and how much they could have contributed [266]. 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.

3.2 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 [250,132,249,252,185]. 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[250]. 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 newtons of force to lift a box contributes, in a very literal sense, 10 newtons of output. Many companies provide monetary bonuses to employees ("pay for performance") in the form of piece rates or for achieving particular sales goals[26,117,1,161,135,230,124]. Researchers have also found that children allocate more rewards to collaborators who generated more output[13,93,119,191].

However, a direct mapping from outputs to rewards might be missing an important intuition: The same amount of work might require more effort from some people than others, because people have varying competencies[267] and tasks may vary in difficulty[17]. 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[116]. 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[140,256,71,192,74,104,69,73]. 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[163]. This hypothesis is supported by prior work on how people mete out punishment for *harmful* actions; both adults and children between 5–11 years of age choose harsher punishments if a person played a causal role in bringing out a harmful outcome[202,203].

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[266]. 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.[266]. This allowed us to examine whether reward is determined directly from judgments of responsibility, as suggested by prior theories[249,250,132,252,185].

To this end, we adapted the materials from Xiang et al.'s (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.

3.3 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. [266] to the bonus allocation problem. The models are summarized in Table 3.1. 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),

[*] All data and code are publicly available at https://github.com/yyyxiang/bonus_allocation.

Table 3.1: Summary of the models.

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

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.

3.3.1 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:

$$B_a^F \propto F_a \tag{3.1}$$

- **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:

$$B_a^S \propto \begin{cases} S_a & \text{if } L = 1 \\ 10 - S_a & \text{if } L = 0 \end{cases} \tag{3.2}$$

- **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:

$$B_a^E \propto E_a \tag{3.3}$$

3.3.2 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 [189], 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:

$$P_a = \begin{cases} \sum_{E'_a} P(E'_a)\mathbb{I}[E'_a S_a + E_{/a} S_{/a} < W] & \text{if } L = 1 \\ \sum_{E'_a} P(E'_a)\mathbb{I}[E'_a S_a + E_{/a} S_{/a} \geq W] & \text{if } L = 0, \end{cases} \tag{3.4}$$

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. [266], 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:

$$B_a^{FA} \propto \begin{cases} P_a & \text{if } L = 1 \\ 1 - P_a & \text{if } L = 0 \end{cases} \tag{3.5}$$

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:

$$B_a^{NFA} \propto \begin{cases} 1 - P_{/a} & \text{if } L = 1 \\ P_{/a} & \text{if } L = 0 \end{cases} \tag{3.6}$$

- **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. [266], we assign equal weights to the two components for simplicity:

$$B_a^{BA} \propto (B_a^{FA} + B_a^{NFA})/2 \tag{3.7}$$

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.

3.3.3 Ensemble model

Xiang et al. [266] 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:

$$R_a^{EBA} \propto wR_a^E + (1-w)R_a^{BA}, \tag{3.8}$$

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:

$$B_a^{EBA} \propto (B_a^E + B_a^{BA})/2 \tag{3.9}$$

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

3.4 Experiments 1a and 1b

In these experiments, we borrowed the task design and stimuli from Experiments 2a and 2b of Xiang et al. [266], 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. [266] 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.

3.4.1 Materials and Methods

Participants

We recruited 181 participants for Experiment 1a and 179 participants for Experiment 1b via Amazon's Mechanical Turk platform (MTurk).[†] We used the same sample size as in Xiang et al. [266] 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.

[†]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.

Stimuli

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

Procedure

Figure 3.1 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.

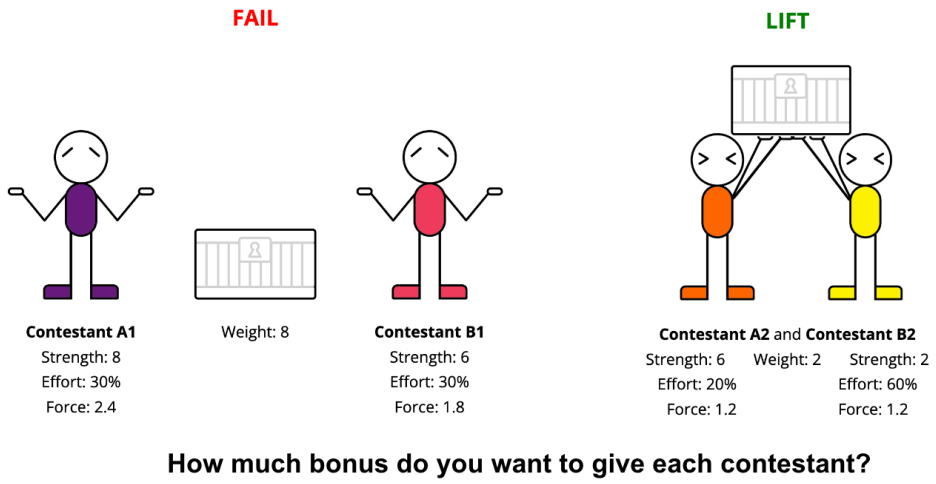


Figure 3.1: Example contest in Experiments 1a and 1b. Agents either failed (left panel) or succeeded (right panel) in lifting the box together. Participants observed the box weight and each agent's strength, effort, and force. They assigned a bonus between $0-10 to each agent separately. In each contest, 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.

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.

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[178] for all the linear mixed-effects models.

3.4.2 Results

Figure 3.2 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 Figure 3.2). 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 3.2 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 Figure 3.2, 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

Table 3.2: Bonus ∼ Condition + (1 + Condition | Participant)

	Estimate	SE	df	t-statistic	p-value
Exp 1a.					
(Intercept)	2.118	0.137	180.0	15.49	<.001
Condition	2.989	0.170	180.0	17.60	<.001
Exp 1b.					
(Intercept)	2.004	0.137	178.0	14.63	<.001
Condition	3.203	0.178	178.0	18.04	<.001

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 anal-

ysis 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 3.3 for the full regression output. Juxtaposing these plots with results from Xiang et al. [266]—the two "Xiang et al. [266] Responsibility" columns in Figure 3.2)[‡]—the biggest difference is in the "Same effort" trials: the stronger agent is judged to be more responsible, but both agents receive similar bonuses.

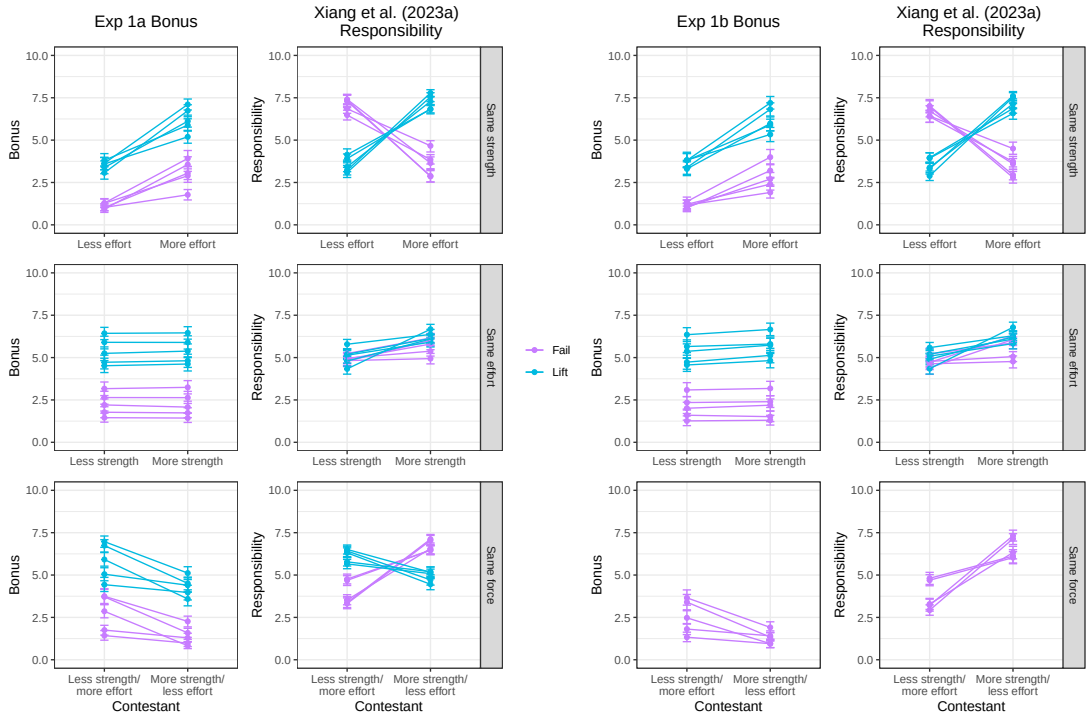


Figure 3.2: Results from Experiments 1a and 1b ("Exp 1a Bonus" and "Exp 1b Bonus") and from Xiang et al. [266] (the two corresponding "Xiang et al. [266] Responsibility" columns). Each line corresponds to a scenario. Error bars indicate bootstrapped 95% confidence intervals.

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

[‡]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.

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 percent 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 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 Figure 3.3A). 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 et al. [266], 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 Figure 3.3A). [§]

Plotting model predictions against the data ("Exp 1a Bonus" and "Exp 1b Bonus" columns in Figure 3.4), 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

[§]For consistency, we also used the same regression formula in the re-analyses of Xiang et al. [266].

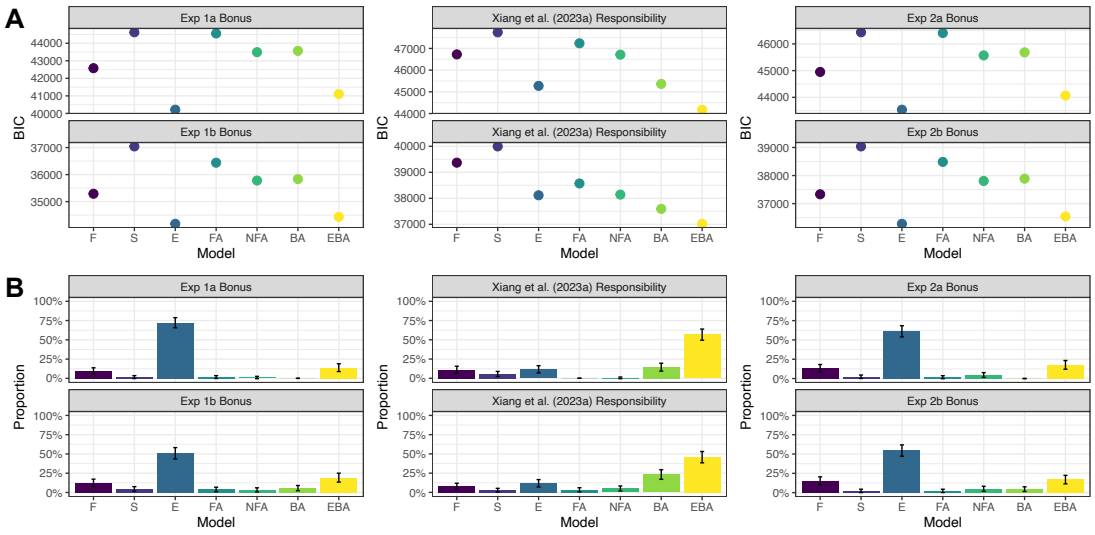


Figure 3.3: Model comparison in Experiments 1a and 1b ("Exp 1a Bonus" and "Exp 1b Bonus"), in Xiang et al.[266] (the two corresponding "Xiang et al.[266] Responsibility" columns), and in Experiments 2a and 2b ("Exp 2a Bonus" and "Exp 2b Bonus"). (A) Bayesian information criterion (BIC) for each mixed-effects 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.

[$r(28) = .97, p < .001$ 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.[266] Responsibility" columns in Figure 3.4).

The Ensemble model was introduced in Xiang et al.[266] 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.[¶]

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 Figure 3.3B). 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 partici-

[¶]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.

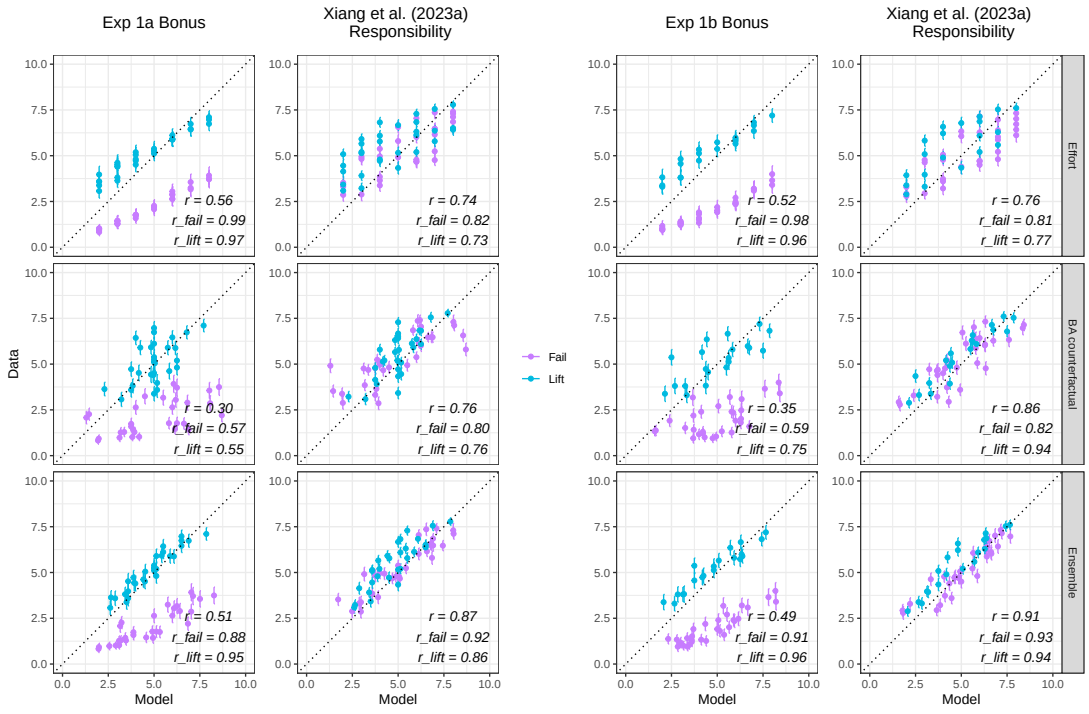


Figure 3.4: Data-model comparison in Experiments 1a and 1b ("Exp 1a Bonus" and "Exp 1b Bonus") and in Xiang et al.[266] (the two corresponding "Xiang et al.[266] Responsibility" columns). Pearson correlation coefficients for conditions combined and separate are shown at the bottom right of each subplot. Each point denotes one agent in one scenario; error bars indicate 95% confidence intervals.

pants in Experiment 1b were best described by the Both-agent counterfactual model. By contrast, in Xiang et al. [266], 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 Figure 3.3B). 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 3.4 for the full regression output.

3.4.3 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. [266], 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.

3.5 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.

3.5.1 Materials and Methods

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.

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.

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).

3.5.2 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 Figure 3.3A). Comparing the left and right panels of Figure 3.3, 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 Figure 3.3B). 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 3.5. 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 Figure 3.5. 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.

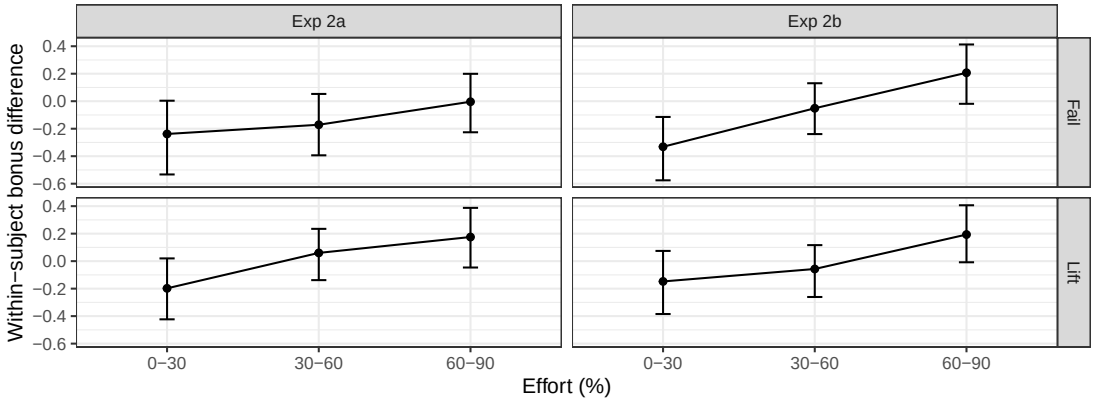


Figure 3.5: Within-subject bonus difference (non-secret agent's bonus minus secret agent's bonus). Agents' effort was binned to 0-30%, 30%-60%, and 60%-90% for visualization (the highest effort level in the stimuli is 80%). Error bars indicate bootstrapped 95% confidence intervals.

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 Figure 3.6) 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.

3.5.3 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

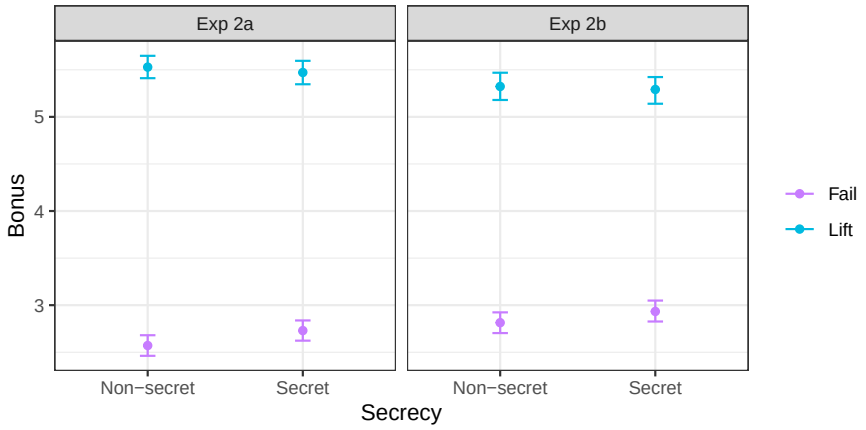


Figure 3.6: Bonus as a function of secrecy and Condition. Error bars indicate bootstrapped 95% confidence intervals.

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.

3.6 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. [266] 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 [17] and serves as a signal about an agent's underlying character [35,4]. 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 [142]. In addition, by exerting more effort, agents can also get better at the task and become better collaborators over longer timescales [268]. 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 [96,89]

where agents' competence may vary. Incentivizing traits that benefit the group may also have long-lasting benefits by promoting collaboration[98,100].

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. 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[67] 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[67,70,131,72]—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[181]. 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[135,41,117,52]. 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;[271,266], whereas the latter cares about how to motivate the agent to do the best they can[28]. 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[130,84]. 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[28,182,135]. 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.[266] did not find a similar effect in responsibility attribution. In fact, this pattern was also found in studies conducted decades ago[252,185]. 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 Figure 3.2 ("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 [25,232], adjust their reward allocations to restore equity in a dyad [139], 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 [110] and reward magnitude [187]. 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) [268]. 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.

Table 3.3: 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	<.001
Condition	2.780	0.169	180.0	16.46	<.001
Contestant	2.320	0.110	180.0	21.16	<.001
Exp 1b. Same strength.					
(Intercept)	0.925	0.131	178.0	7.05	<.001
Condition	2.932	0.179	178.0	16.42	<.001
Contestant	2.153	0.111	178.0	19.42	<.001
Exp 1a. Same effort.					
(Intercept)	2.225	0.156	180.0	14.23	<.001
Condition	3.163	0.183	180.0	17.30	<.001
Contestant	0.023	0.055	180.0	0.42	.678
Exp 1b. Same effort.					
(Intercept)	2.002	0.154	178.0	12.97	<.001
Condition	3.392	0.184	178.0	18.42	<.001
Contestant	0.174	0.064	178.0	2.72	.007
Exp 1a. Same force.					
(Intercept)	2.749	0.156	180.0	17.57	<.001
Condition	3.024	0.173	180.0	17.48	<.001
Contestant	-1.407	0.082	180.0	-17.24	<.001
Exp 1b. Same force.					
(Intercept)	2.534	0.166	178.0	15.31	<.001
Contestant	-1.222	0.093	178.0	-13.12	<.001
Note. For the "Same force" trials in Experiment 1b, the Condition regressor was dropped because these trials could only produce a "Fail" outcome.					

Table 3.4: Bonus ∼ Trial + (1 + Trial | Participant)

	Estimate	SE	df	t-statistic	p-value
Exp 1a.					
(Intercept)	3.612	0.127	180.0	28.36	<.001
Trial	0.036	0.037	180.0	0.98	.328
Exp 1b.					
(Intercept)	3.285	0.123	178.0	26.61	<.001
Trial	0.043	0.046	178.0	0.94	.347

Table 3.5: 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	<.001
Effort	1.030	0.056	177.3	18.28	<.001
Secrecy	0.068	0.076	177.2	0.89	.374
Condition	2.849	0.169	177.0	16.85	<.001
Effort × Secrecy	-0.136	0.043	177.8	-3.17	.002
Exp 2b.					
(Intercept)	2.849	0.162	179.0	17.64	<.001
Effort	1.105	0.058	179.8	19.16	<.001
Secrecy	0.085	0.072	179.7	1.17	.243
Condition	2.397	0.161	179.0	14.87	<.001
Effort × Secrecy	-0.185	0.048	178.6	-3.89	<.001

4

Optimizing competence in the service of collaboration

4.1 Abstract

In order to efficiently divide labor with others, it's important to understand what our collaborators can do (i.e., their *competence*). However, competence is not static—people get better at particular jobs the more often they perform them. This plasticity of competence creates a challenge for collaboration: For example, is it better to assign tasks to whoever is most competent now, or to the person who can be trained most efficiently "on-the-job"? We conducted four experiments ($N = 396$)

that examine how people make decisions about whom to train (Experiments 1 and 3) and whom to recruit (Experiments 2 and 4) to a collaborative task, based on the simulated collaborators' starting expertise, the training opportunities available, and the goal of the task. We found that participants' decisions were best captured by a *planning* model that attempts to maximize the returns from collaboration while minimizing the costs of hiring and training individual collaborators. This planning model outperformed alternative models that based these decisions on the agents' current competence, or on how much agents stood to improve in a single training step, without considering whether this training would enable agents to succeed at the task in the long run. Our findings suggest that people do not recruit and train collaborators based solely on their current competence, nor solely on the opportunities for their collaborators to improve. Instead, people use an intuitive theory of competence to balance the costs of hiring and training others against the benefits to the collaboration.

4.2 Introduction

Division of labor is central to collaboration: Even young children can appropriately assign tasks based on the relative competence of their collaborators [154,7], or take their collaborator's physical constraints into account when planning their own actions [245]. One of the chief advantages of division of labor is that it allows individual workers to get better at specialized tasks through practice. In *The Wealth of Nations*, Smith [205] gives the example of a contest between a teenager and a smith: Through sheer practice, a teenager who solely makes nails, day in and day out, can vastly outperform a smith who only occasionally makes nails as one of the many products of her forge. Smith [205] argued that these increases in efficiency from division of labor could directly account for extraordinary increases in the wealth of nations. Indeed, many of the benefits of collaboration and division of labor stem precisely from the fact that competence is not static.

Taking into account how people's competence changes with practice is particularly important when making decisions about whom to recruit for a collaborative task, or how much to invest in their training. For example, a company might hire a less experienced candidate who—with appropriate training—could qualify for the job, as opposed to a highly-skilled candidate who requires a hefty salary. Similarly, graduate school advisors might admit a student who shows great potential, given opportunities to train and learn. Training also allows current team members to become better suited for their roles. Organizations investing in effective training programs that enhance employees' capabilities tend to achieve both short- and long-term benefits for both the employees and the organizations [169]. Effective training fills the gap between desired performance and actual employee performance; by improving employees' performance and capabilities, companies increase their organizational productivity [56]. Formal employee training programs are also unique in their ability to bring below-average firms up to the performance level of comparable businesses [11]. In short, people often make decisions in real-world collaborations by considering not only how capable people are already, but also how capable they *will become* over the course of a long-term collaboration. However, little is known about the lay intuitions behind how people make decisions about whom to recruit, whom to train, and how to train them.

Answers to these questions hinge on many factors, including what we know about the team structure, what training resources are available, how much time we have for training, etc., and also require that we take into account how our collaborators' competence may change on the job. These decisions can be characterized as a sequential decision problem that requires long-term prospection. Recent work suggests that—much as people can use *planning* to solve sequential decision problems [50,109] in non-social settings—they can also plan joint actions efficiently [225,226,48] and use intuitive theories of other people's minds to plan interventions on their mental states [106], including beliefs and desires [179,10,9,114], and also representations of competence and effort [267]. Thus, one possibility is that people may make decisions about whom to recruit and train for a collaborative task

by anticipating how others' competence may change over time and planning accordingly. However, planning in general is computationally expensive, and these costs may be prohibitive in the context of collaboration, where optimal planning requires recursive theory of mind[237].

Thus, people may instead use simpler heuristics to decide whom to assign to what task without planning ahead[53,174,175]. For example, people may simply assign tasks based on who is more competent now, without considering how their competence will change as a result of training. Indeed, past work has found that adults tend to select teams whose combined expertise suffices for a particular job[267], and that children tend to use relative differences in ability to assign people to tasks[7,154]. Another possibility is that—rather than treating training as a means to an end—people might instead train agents who need it the most. If this is the case, then they may selectively train the weakest agent, or the agent who stands to improve the most from a single bout of training, without planning ahead to consider what the group's overall competence will be after training.

Across four experiments (N = 396), we show that people engage in planning when they make training and hiring decisions regarding simulated agents. The first two experiments are a collaborative box-lifting task framed as a game show. In Experiment 1, participants trained agents to prepare them to lift a heavy box together. In Experiment 2, participants first selected two agents out of a pool of four candidates, then trained the selected agents to lift a heavy box together. Experiments 3 and 4 generalized the task to a more education-oriented context, where participants recruited and trained students for a math Olympiad. We compared the human judgments to a *planning* model and four alternative models that base training decisions on heuristics: an *exploitation* model that selects the strongest agent, an *equity* model that selects the weakest agent, a *learning* model that trains agents who will improve the most after a single training step, and an *equality* model that invests equally in every agent's training. We found that the planning model qualitatively and quantitatively provided the best match to the data. * In the following section, we describe these models in detail.

*All data and code are publicly available at https://github.com/yyyxiang/competence_training.

4.3 Theoretical framework

In Experiments 1 and 2, pairs of simulated agents play a contest where they attempt to lift a heavy box (i.e., the target) together to win a prize. Before the contest, participants can increase the agents' strength by assigning them boxes to train with on their own. Experiments 3 and 4 apply the same framework to a similar problem with a different framing: Participants—imagining themselves as high school math coaches—recruit and train student contestants for a math Olympiad. Participants can increase students' math abilities by assigning them problem sets to complete. Since these two paradigms differ only in the framing, we describe below how we set up the models for the collaborative box-lifting task in Experiments 1 and 2. We start by describing how individual "trainees" (i.e., simulated agents) gain strength, based on the effort they invest into the lift, and then describe a suite of models of how "trainers" (i.e., participants) decide whom to train and how to train them.

4.3.1 Modeling trainees' behavior

Suppose each agent a has a strength S_a and exerts effort $E_a \in [0, 1]$ to lift a training box of weight W. Effort E_a defines the proportion of an agent's strength that is applied to lifting.[†] Therefore, agent a applies a force of $S_a \cdot E_a$ to lift, and succeeds if this force equals or exceeds the box weight W (i.e., $S_a \cdot E_a \geq W$). The agent's utility function is defined as the reward from lifting the box minus the effort cost[267]:

$$U(E) = R \cdot L - C(E), \tag{4.1}$$

where R is the reward the agent receives from successfully lifting the box, $C(E)$ is the effort cost, and $L = \mathbb{I}[E_a \cdot S_a \geq W]$. The indicator function $\mathbb{I}[\cdot] = 1$ if its argument is true, 0 otherwise; thus, $L = 1$ indicates that the lift was successful. We model agents as deterministic utility maximizers (i.e., agents

[†]For computational tractability, we constrained the effort space to discrete values, ranging from 0 to 1 with increments of 0.01.

always choose the effort level that maximizes the utility function):

$$E_a^* = \underset{E_a}{\operatorname{argmax}}\, U(E_a). \tag{4.2}$$

Note that in this phase, where agents are lifting boxes individually, we do not yet need to consider collaborative optimization of effort.

For simplicity, we assume that agents are highly motivated to lift the box (i.e., $R \gg C(E)$). Therefore, E_a^* is effectively:

$$E_a^* = \begin{cases} W/S_a & \text{if } S_a \geq W\,(L=1) \\ 0 & \text{if } S_a < W\,(L=0) \end{cases} \tag{4.3}$$

This means that an observer can predict an agent's effort once they know the agent's strength and the weight of the box.

We make a simple yet plausible assumption about how strength increases with training: Strength increases in proportion to effort, which can be formalized by the following update rule:

$$\Delta S_a = \rho \cdot E_a, \tag{4.4}$$

where $\rho \in [0,1]$ is a learning rate. Here, we set ρ to 1, such that strength increases by the level of effort exerted. If we assume that agents exert the optimal effort for the task (Equation 4.3), we see that failed lifts do not increase agents' strength—as the optimal effort is 0—whereas a successful lift increases it in proportion to the effort exerted. Because agents' strength increases after each training bout, assigning the same weight a second time will require less effort and increase the agents' strength less, consistent with the "diminishing returns" principle of strength training[165].

4.3.2 Modeling trainers' decisions

In Experiment 1, participants trained pairs of agents so that they could lift a target box of weight W together. We define *training strategies* as the sequences of actions taken by participants during training; at each step, participants can choose to pay a small training cost to train one agent using a box selected from an array, or they can choose to not train any agents at all to save on training costs. We use S'_{ai} to denote agent a's strength after training according to training strategy i. If the agents' combined strength equals or exceeds W by the end of training ($\sum_a S'_{ai} \geq W$), then the joint lift succeeds ($L = 1$), and participants receive a reward of R minus the cumulative training costs $K(i)$ of training strategy i. We set $K(i)$ as a small, fixed fee k for every unit of weight used during training; heavier boxes thus deduct higher costs from participants' bonus.

In Experiment 2, participants additionally chose which agents to recruit prior to training. Participants first selected two agents out of a pool of four candidates to form a team t, then trained the two selected agents ($a \in t$) to prepare the two of them to lift a heavy box together. The training process was identical to Experiment 1, although there was no cost associated with training (i.e., $K(i) = 0$). Instead, participants paid a hiring fee for each agent recruited on the team, which we set as a small, fixed fee h for every unit of strength. We denote the total fee for team t by $H(t)$.

To capture how participants make these decisions, we first describe a planning model that anticipates how agents' strengths will change as a result of the training strategy, then describe four alternative models that use heuristics to make decisions.

Planning model

For the *planning model*, the trainer's goal is to obtain the reward with minimal cost. Thus, the action value of training strategy i is defined as:

$$Q_i^{planning} = R \cdot L - K(i), \tag{4.5}$$

where L indicates whether the joint lift is successful by comparing the sum of agents' forces to the target box weight. L is defined as:

$$L = \mathbb{I}\left[\sum_a E'_a \cdot s'_{ai} \geq W\right], \tag{4.6}$$

where E'_a indicates the level of effort agent a exerts during the joint lifting.

Since we assume that agents are highly motivated to exert effort when necessary (i.e., $E'_a \approx 1$), the lift function becomes:

$$L \approx \mathbb{I}\left[\sum_a s'_{ai} \geq W\right] \tag{4.7}$$

In many cases, some strategies are equally preferred, and some others are only slightly less preferred. In light of this, we decided to add stochasticity to the choice process, instead of keeping the training strategy with the highest action value. We passed action values through a softmax function so that the model chooses training strategies stochastically. Strategies with higher action values are more likely to be chosen. The probability of selecting training strategy i is:

$$P_i = \frac{e^{\gamma Q_i}}{\sum_{j=1}^{K} e^{\gamma Q_j}}, \tag{4.8}$$

where γ is the inverse temperature parameter that controls the stochasticity of the predictions by scaling the action values before applying softmax, and K is the total number of training strategies.

The prediction of agents' strength after training is therefore the expected value of the agent's updated strength—in other words, a weighted average of updated strengths from all the possible

training strategies:

$$S'_a = \sum_i P_i \cdot S'_{ai} \tag{4.9}$$

In the team selection problem in Experiment 2, the action value for choosing each team is defined as:

$$Q_t^{planning} = R \cdot L - H(t), \tag{4.10}$$

where L is again an indicator function returning whether the agents lifted the target box together:

$$L = \mathbb{I}\left[\sum_{a \in t} S'_a \geq W\right] \tag{4.11}$$

Here S'_a indicates the selected agents' strengths post-training following the same calculation described above (Equation 4.9), except that we add the constraint that only agents selected on the team ($a \in t$) can receive training.

Alternative models

We compare the planning model to four alternative models that use heuristics in lieu of planning: an *exploitation model*, an *equity model*, a *learning model*, and an *equality model*. Same as above, we use S_a to denote agent a's strength before training, and S'_a to denote agent a's strength after training. Below, we describe how each of these models computes the action values of each strategy. The action values are then passed through the softmax function.

The ***exploitation model*** maximizes the strongest agent's strength after training while considering the costs of training and hiring. The action value of each training strategy is therefore:

$$Q_i^{exploitation} = \max_a S'_{ai} - K(i) \tag{4.12}$$

And the action value of selecting each team is:

$$Q_t^{exploitation} = \max_a S'_a - H(t) \tag{4.13}$$

Note that here and elsewhere, S'_a is implicitly dependent on the team t since only agents in t are trainable.

The ***equity model*** minimizes the strength difference between the strongest agent and the weakest agent after training while considering the costs of training and hiring:

$$Q_i^{equity} = \min_a S'_{ai} - \max_a S'_{ai} - K(i) \tag{4.14}$$

$$Q_t^{equity} = \min_a S'_{ai} - \max_a S'_a - H(t) \tag{4.15}$$

The ***learning model*** maximizes the total strength gain from training while considering the costs of training and hiring:

$$Q_i^{learning} = \sum_a S'_{ai} - S_a - K(i) \tag{4.16}$$

$$Q_t^{learning} = \sum_a S'_a - S_a - H(t) \tag{4.17}$$

The ***equality model*** invests equally in every agent's training by indiscriminately choosing training strategies and teams inspired by [157,102] while considering the costs of training and hiring:

$$Q_i^{equality} = -K(i) \tag{4.18}$$

$$Q_t^{equality} = -H(t) \tag{4.19}$$

We also considered a Rawlsian model that maximizes the strength of the weakest agent after training [183] and a utilitarian-style model that maximizes the sum of the agents' strengths after train-

ing[162,14]. The Rawlsian model makes similar predictions as the equity model because maximizing the weakest agent's strength after training is essentially reducing the strength difference between the strongest and the weakest agents (for example, both models tend to focus entirely on training the weakest agent). The utilitarian-style model makes similar predictions as the learning model because maximizing the sum of the agents' strengths after training is equivalent to maximizing total strength gain. Since these two models make similar predictions as the ones we described above, we focus on the four aforementioned alternative models (exploitation, equity, learning, and equality) in the interest of parsimony.

In Experiments 3 and 4, we extend this framework into a math Olympiad task, where agents' strength is replaced with math abilities, and box weights are replaced with difficulties of math problems. Everything else carries over into the new task.

4.3.3 Parameter fitting

Our theoretical framework contains a single free parameter: the inverse temperature parameter γ in Equation 4.8. For every model, we fit this parameter to each participant's choices (training strategies in Experiments 1 and 3 and teams in Experiments 2 and 4) using maximum likelihood estimation through a grid search with search space ranging from 5 to 15, with increments of 0.5. We chose this range because, below the lower bound, the alternative models approach uniform distributions (specifically, exploitation, equity, and learning models in Experiments 1 and 3, and all four alternative models in Experiments 2 and 4).

4.4 Experiment 1

The purpose of Experiment 1 was to examine how participants train teammates in preparation for a collaborative task. Specifically, we asked whether participants' decisions are better captured by the

planning model or by alternative heuristic models. We designed four scenarios in which the models described above make qualitatively distinct predictions. The planning model predicts that people will train agents to be *just strong enough* to clear the task, while the exploitation, equity, and learning models will try to maximize agents' final strength by maximizing the strongest agent's strength, reduction in the strength difference between the strongest and weakest agents, and the total strength gain, respectively, after accounting for training costs. The equality model will essentially only consider training costs since, otherwise, it prefers every training strategy equally. To ground the scenarios in a realistic setting, we framed the experiment as a game show consisting of four contests.

4.4.1 Materials and Methods

Participants

We recruited 99 participants via Amazon's Mechanical Turk platform (MTurk).[‡] In this and subsequent experiments, to ensure data quality, our task was only available to MTurk workers who have had at least 1,000 studies approved and an approval rate higher than 98%, and participants were only allowed to proceed to the experiment after answering all questions in a comprehension check correctly.

Participants received a base pay of $1 and a potential bonus payment of up to $4. The bonus payment depended on the training boxes participants used to train the contestants and whether the lifting turned out successful. Specifically, for every successful lift, participants received $1; for every weight unit they used, $0.04 was deducted from their bonus payment, regardless of the lift outcome. The experiment was approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/1GF_TVQ.

[‡]In Experiments 1, 3, and 4, we stopped data collection once 100 participants submitted their responses on MTurk, as stated in the pre-registration, but only received 99 participants' data due to a server error.

Procedure

Figure 4.1 shows the task setup. The game show consisted of four contests, each with a different pair of contestants. In each contest, participants first saw the contestants' starting strengths and the weight of a heavy box, and were asked to train contestants to prepare them to lift the heavy box together. In order to make the strengths and weights intuitive to the participants, we expressed them using the same units. 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.

Participants were provided with four training boxes in all contests (Weights 2, 4, 6, and 8) and were told that contestants would gain strength based on the effort they put into lifting each box. For example, if a contestant with a starting strength of 6 was assigned a training box of weight 3, they would only need to put in a 50% effort to lift the box; consequently, the contestant's strength would increase by 0.5. However, if the box is too heavy for them to lift, contestants will give up and not gain any strength. To ensure that participants understood the task setup, they completed three practice trials where they observed how a contestant's strength changed by lifting different boxes.

In each contest, participants had up to three turns to train the contestants. On each turn, they could choose which agent to train and which training box to assign to them, or they could choose to not train anyone at all. Participants saw the updated strengths of the contestants after each turn. After training, participants saw the two contestants attempt to lift the heavy box together; the contestants were able to do so only if their combined strength was equal to or greater than the weight of the heavy box. Links to the experiments can be found at `https://github.com/yyyxiang/competence_training`.

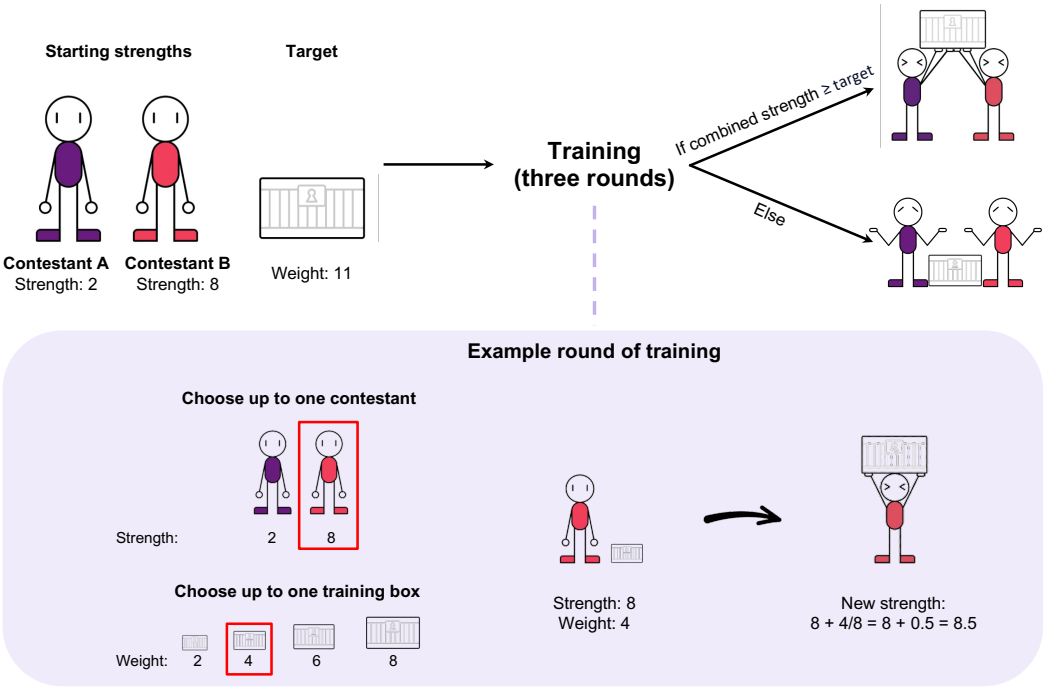


Figure 4.1: Example contest in Experiment 1. Participants saw two agents of varying strengths and a target box. They had three rounds to train the agents. In each round of training, they could either assign a box to an agent or not train anyone. Agents gained strength equivalent to the level of effort they put forth. After training, agents attempted to lift the target box together.

4.4.2 Results

Figure 4.2A shows average behavioral data and model predictions in a series of 2-D plots. Each subplot corresponds to a different scenario, and each line shows how both agents' strength changed, on average, over the course of training in that scenario; the weaker agent's strength is plotted on the x-axis, and the stronger agent's strength is plotted on the y-axis. Each point in the participants' responses (Data) denotes the average strength of the two agents in each round of training. Each point in the model predictions denotes the weighted average strength calculated from Equation 4.9, using values of γ fit to individual participants. Because agents' strengths can only increase or stay the same, their starting strengths are at the origin of each plot, and points along a line that are farther from

the origin occurred later in training. The dashed black lines indicate the minimum total strength required to lift the target box. In other words, agents are strong enough to complete the task once their combined strength passes the dashed line.

From these plots, we can see that participants' average responses aligned closely with the predictions of the planning model, and they both have endpoints close to the dashed line; both participants and the planning model tended to train agents to be just strong enough for the task, weighing rewards and costs. By contrast, the alternative models failed to capture this pattern: The exploitation model trained the stronger agent as much as possible, the equity model trained the weaker agent as much as possible, and the learning model trained the two agents for the largest strength gain possible, all without considering the target weight. The equality model, on the other hand, did not train agents to be strong enough to lift the box in three out of four scenarios.

To quantitatively assess how well each model captures the data, we computed correlations between the participant-averaged data and model predictions of agents' strength at every step of training. The planning model and the equality model have the highest correlation with the data (Pearson's $r = 0.999, p < .0001$ for both), followed by the learning model ($r = 0.997, p < .0001$), exploitation model ($r = 0.992, p < .0001$), and equity model ($r = 0.983, p < .0001$). Note that these correlation coefficients are all very high because agents' strength can only increase or stay the same with training, and because agents' relative strength does not change much during scenarios, as the maximum strength increase possible during training is small compared to the difference in starting strengths between agents. Quantitatively comparing the average model-predicted *strengths* obscures differences between our models. Therefore, we deviated from our pre-registered plan and instead compared individual participants' *training strategies* to model predictions using random-effects Bayesian model selection [186]. For each participant, we first evaluated the fit of each model using $BIC = k\ln(n) - 2\ln(L)$, where k is the number of free parameters, n is the number of scenarios, and L is the maximized value of the log likelihood. We then approximated the log model evidence

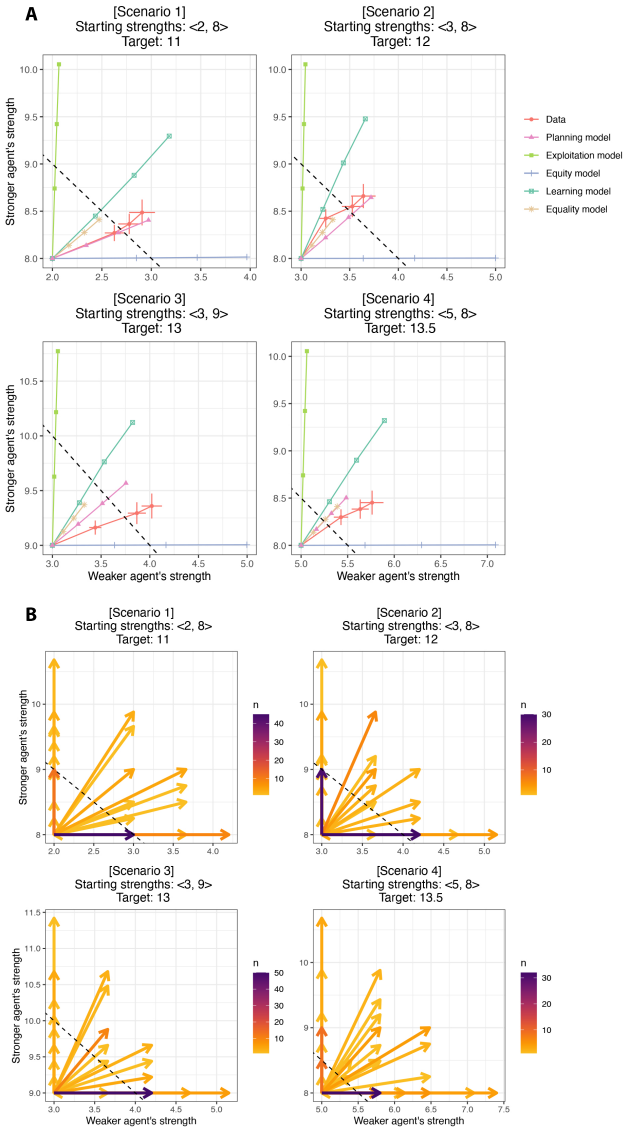


Figure 4.2: Experiment 1 results. (A) Data and model predictions of agents' strength in each round of training. Each subplot corresponds to one scenario, where the x-axis shows the weaker agent's strength, the y-axis shows the stronger agent's strength, and each line traces agents' average strength in each training trial. The dashed black line marks the threshold where agents' combined strength exceeds the target box weight; intuitively, if participants are balancing the costs and benefits of training, they should train agents until their combined strength reaches this threshold, and no further. Error bars indicate 95% confidence intervals. (B) Individual participants' training trajectories. Each subplot denotes a scenario, and each arrow denotes a particular training strategy; the color of the arrow indicates how many participants adopted this strategy. For simplicity, here we represent each training strategy by showing only agents' strength before (origin) and after training (arrowhead). The dashed black line marks the threshold where agents' combined strength exceeds the target box weight.

for each participant as $-0.5BIC$[18] and used it to compute the protected exceedance probability for each model, which can be interpreted as the probability that each model is the most frequently occurring in the population.

Overall, the planning model has a protected exceedance probability of approximately 1, which means that the planning model is the most likely model in the population. To compare how well each model fits individual participants' responses, Figure 4.3A shows the distribution of model evidences for each model for each participant. The planning model had the highest model evidence. It is worth noting that Experiment 1 was not originally designed to discriminate between the planning model and the equality model; we differentiate between them in Experiment 3, below, with four new scenarios.

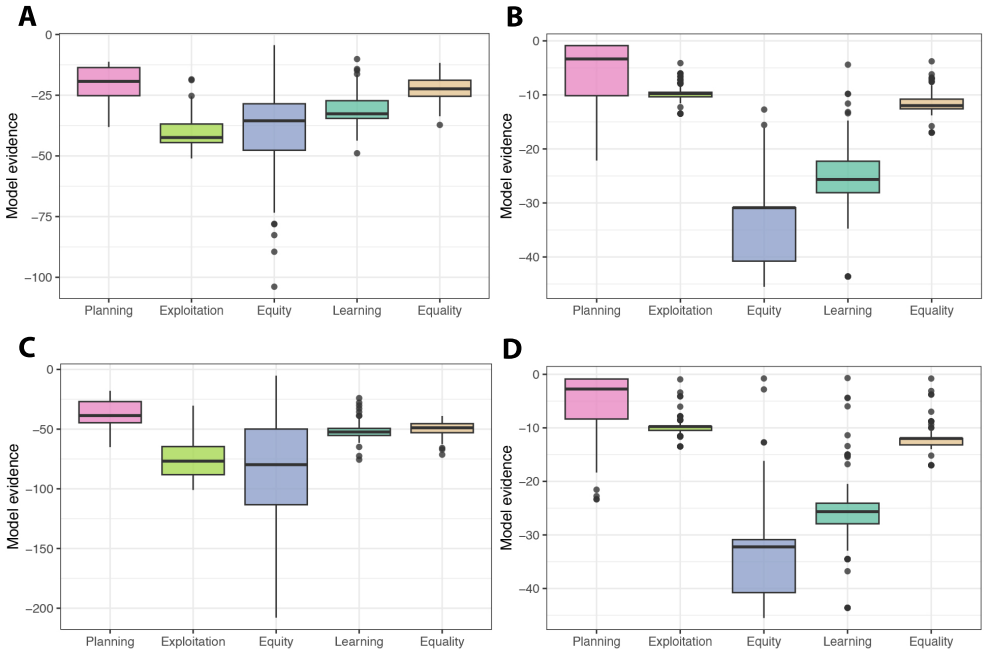


Figure 4.3: Distribution of model evidences for each participant (approximated as $-0.5BIC$) in (A) Experiment 1, (B) Experiment 2, (C) Experiment 3, and (D) Experiment 4.

Finally, we took a closer look at individual participants' data (Figure 4.2B). Each arrow corre-

sponds to a particular training strategy, and the color of the arrow indicates how many participants adopted this strategy. For simplicity, here we represent each training strategy by showing only agents' strengths before (origin) and after training (arrowhead). Intuitively, diagonal arrows indicate that the participant trained both agents, horizontal arrows indicate training the weaker agent only, and vertical arrows indicate training the stronger agent only. While some participants trained agents more than what was necessary for goal achievement—either training both agents in turn or training a single agent over multiple rounds—overall, we found that the modal trajectories for all the scenarios show that participants trained either the stronger agent or the weaker agent until agents' combined strength reached the threshold for success (indicated by the dashed black line). Here, three out of four scenarios in Experiment 1 required only one round of training for success. In Experiment 3, we explored scenarios where three rounds of training are needed.

4.4.3 Discussion

In this experiment, we compared the planning model and four alternative models to participants' training strategies using random-effects Bayesian model selection. Overall, the planning model is the closest to the data both qualitatively and quantitatively and occurs the most frequently in the population.

4.5 Experiment 2

Experiment 1 addressed the question of how people decide which team members to train and how to train them. However, it does not address the problem of whom to recruit to the team in the first place. This question is important because, in real-world scenarios, a team doesn't necessarily need to recruit members who are already capable of a collaborative task to succeed at it; they can also recruit members who can be trained to be good enough for the task. For example, a company can

either choose to poach a highly-skilled employee, or hire someone less experienced who will require training on the job. In Experiment 2, we modified our design to study how training considerations affect people's decisions about whom to recruit. Instead of giving participants two agents to train as in Experiment 1, here we asked participants to first hire two contestants from a pool of candidates with varying strengths and hiring costs that scale with strength and then train contestants to lift a heavy box together. We designed three scenarios with varying target box weights to test which model best predicts which candidates participants tend to hire. Intuitively, the planning model selects teams based on the target, forming different teams as needed for different scenarios, while alternative models each tend to select the same teams in all scenarios.

4.5.1 Materials and Methods

Participants

We recruited 100 participants via Amazon's Mechanical Turk platform (MTurk). Participants received a base pay of $1 and a potential bonus payment of up to $6. The bonus payment depended on which contestants participants selected to train and complete the task, and whether the lifting turned out successful. Specifically, for every successful lift, participants received $2; for every strength unit they used, $0.02 was deducted from their bonus payment, regardless of the lift outcome. There were no costs associated with training boxes. The experiment was approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/6LJ_2D9.

Procedure

The game show consisted of three contests. At the start of each contest, participants were shown four agents who auditioned to appear in the game show. Each potential contestant had a different starting strength and cost a different amount of money to hire. Figure 4.4 shows the contestants and

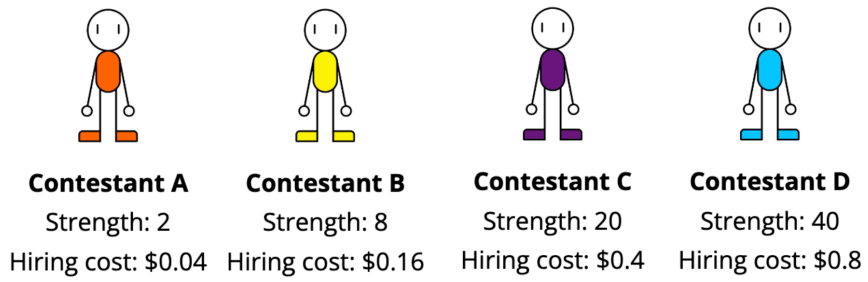


Figure 4.4: The four candidates in Experiment 2. Hiring costs are proportional to agents' strength.

the hiring fees; in general, stronger contestants cost more to hire. In each contest, participants chose two contestants and trained them separately to prepare them to lift a heavy box together. Once the participants selected two contestants, the training procedure was identical to Experiment 1.

4.5.2 Results

We compared the proportion of participants that chose each team to the model-predicted probabilities of choosing each team. As shown in the first row of Figure 4.5A, participants hired agents who were either just strong enough for the task (e.g., agents with strengths 8 and 20 for a target of 24) or that could become strong enough for the task with some training (e.g., agents with strengths 2 and 20 for a target of 24). When both options were available, participants favored the latter—hiring agents that needed training over agents that did not—to reduce hiring costs. The planning model (second row) captured these response patterns. Similarly to participants, the planning model balanced the hiring costs against the reward of successful collaborations. Thus, it preferred teams that would succeed (either with or without training) over teams that would not, and it preferred teams that were cheaper to recruit among teams that would succeed. The alternative models (last four rows), on the other hand, did not predict these patterns in the data. Instead, the alternative models each preferred the same teams in all scenarios, regardless of whether this team composition was appropriate for the target box weight, and all alternative models had some tendency to hire weaker

agents to lower hiring costs. This tendency was strongest in the learning model, which selected the two weakest agents to maximize strength gain, and weakest in the exploitation model, which had some tendency to hire the strongest agent to maximize the strongest agent's strength after training. By contrast, the equity model preferred to hire the single weakest agent to minimize the gap between the strongest and weakest agents, and the equality model dispreferred hiring the strongest agent due to hiring costs; however, neither of them had strong preferences among the remaining agents.

To evaluate each model's quantitative fit to the data, we computed correlations between the data and model predictions. The planning model has a very strong correlation with the data (Pearson's $r = 0.951, p < .0001$), whereas the alternative models all have weak correlation with the data: the exploitation model ($r = -0.256, p = .31$), equity model ($r = -0.207, p = .41$), learning model ($r = -0.195, p = .44$), and equality model ($r = -0.218, p = .38$).

Consistent with Experiment 1, we also compared each of the models to behavioral data using random-effects Bayesian model selection. We found that the planning model has a protected exceedance probability of approximately 1, suggesting that the planning model is the most likely model in the population. The planning model also had the highest model evidence (Figure 4.3B).

We additionally plotted individual participants' training trajectories. For every scenario, we zoomed in on the training trajectories of the modal teams: <2,20> in Scenario 1, <8,40> in Scenario 2, and <20,40> in Scenario 3 (Figure 4.5B). As before, each arrow represents agents' pre-training and post-training strength based on a specific training strategy. Diagonal arrows indicate that the participant trained both agents, horizontal arrows indicate training the weaker agent only, and vertical arrows indicate training the stronger agent only. The color of each arrow shows the number of participants who adopted each strategy. We see that the majority of participants trained only one agent, to the point where agents' combined strength exceeded the threshold for success (the dashed black line).

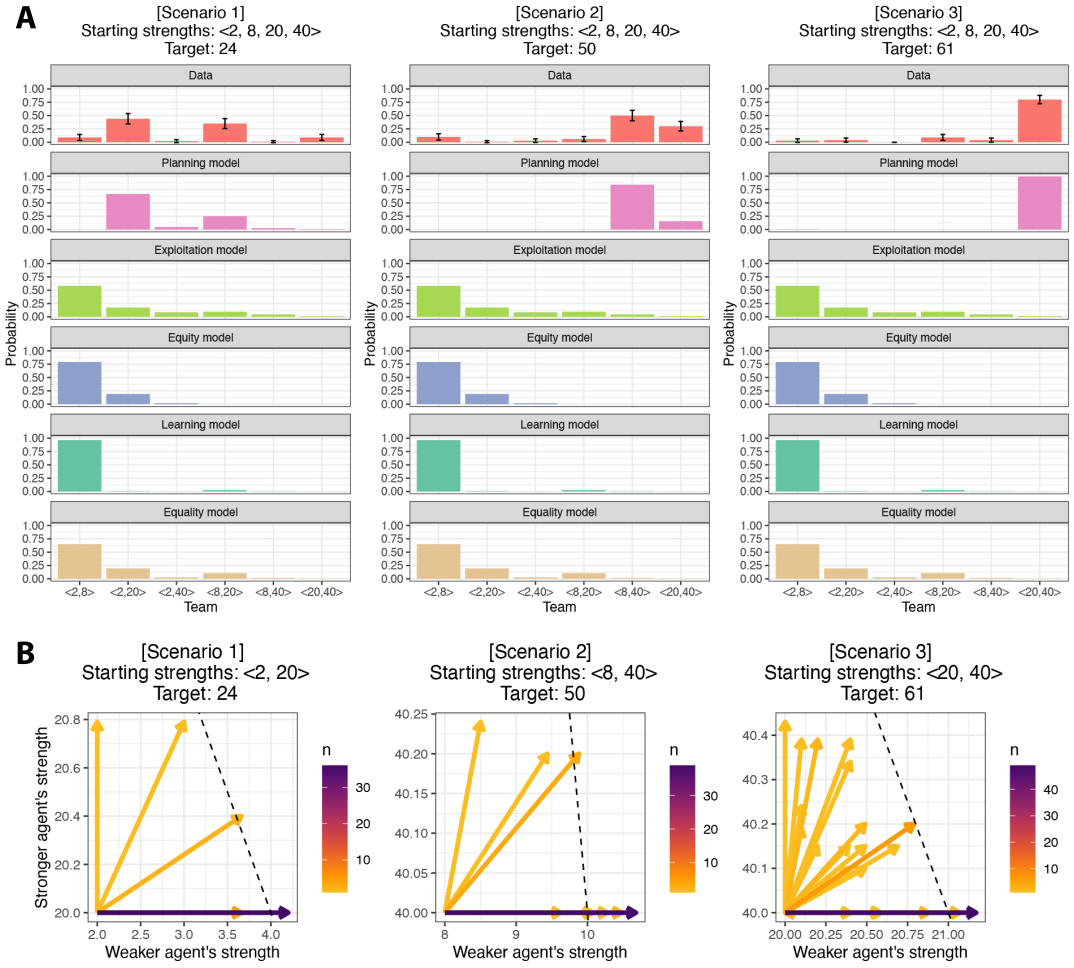


Figure 4.5: Experiment 2 results. (A) Data and model predictions, showing the probability of choosing each team. Each subplot corresponds to one scenario; the brackets on the x-axis labels refer to the strength combinations of the two hired agents. Error bars indicate 95% confidence intervals of proportions. (B) Individual participants' training trajectories of the modal teams. Each subplot denotes a scenario, and each arrow denotes a particular training strategy; the color of the arrow indicates how many participants adopted this strategy. For simplicity, here we represent each training strategy by showing only agents' strength before (origin) and after training (arrowhead). The dashed black line marks the threshold where agents' combined strength exceeds the target box weight.

4.5.3 Discussion

In Experiment 2, we extended our task setup and models to the problem of team selection. The planning model captures participants' decisions about training *and* recruitment both qualitatively and quantitatively, whereas the alternative models deviate from the data substantively. Put together, these results suggest that people anticipate how their collaborators' competence will change and what problems they will face in the future when they decide whom to recruit.

4.6 Experiment 3

To test the generalizability of our findings, in Experiments 3 and 4, participants completed tasks that had the same underlying structure as Experiments 1 and 2, but were presented in a different context. Specifically, instead of asking participants to hire and train contestants of different strengths for a physical task, we asked them to recruit and train students with different math abilities for a math Olympiad. Though the underlying task structure remains the same, these two contexts differ from one another in potentially important ways; for example, rather than training contestants in one-off games, participants in this task assumed the role of an educator, in a context where planning may be superseded by concerns about allocating training equally or to those who need it the most. These experiments thus provide a test of whether the cognitive processes that underlie decisions about recruitment and training apply more broadly to sequential social decision-making situations, beyond one context. Experiment 3 is a replication of Experiment 1 and Experiment 4 is a replication of Experiment 2. We included all combinations of starting strengths and target box weights used in the first two experiments; here, they were reframed as starting math abilities and target problem difficulties, respectively.

In Experiment 3, to discriminate between the planning model and equality model, we added four new scenarios (Scenarios 5-8) where more training is required to reach the target, as opposed

to Scenarios 1-4 (same as Experiment 1) where only a little training is needed. The new scenarios had the same starting math abilities as the original scenarios, but higher target problem difficulties. In these scenarios, the planning model predicts that participants will train agents more as the target increases, whereas alternative models predict the same training strategies despite the target increase. If participants weigh the costs of training against the benefits of successful collaboration, then we should expect that they should also train agents differently to achieve different targets.

4.6.1 Materials and Methods

Participants

We recruited 98 participants via Amazon's Mechanical Turk platform (MTurk). Participants received a base pay of $2 and a potential bonus payment of up to $8. The bonus payment depended on the problem sets participants used to train the student contestants and whether the student contestants won. Specifically, for every win, participants received $1; for every difficulty level unit of the problem sets they used, $0.04 was deducted from their bonus payment, regardless of the contest outcome. The experiment was approved by the Harvard Institutional Review Board and preregistered at https://aspredicted.org/LJ7_YJD.

Procedure

Participants were instructed to imagine themselves as high school math coaches. Their school was sending eight different pairs of students to participate in eight different math Olympiads, and the participants' job was to train students to prepare them to work out a difficult math problem together to win the contest. Students can get better at math by completing problem sets; their abilities, indexed by their "math levels", increase by the level of effort they put into solving the problem sets, and a problem set that is too hard for them will not increase their math levels. The training proce-

dure was the same as in Experiment 1.

4.6.2 Results

Figure 4.6 visualizes the data and model predictions in a series of 2-D plots. Plots on the same row show scenarios with the same starting math levels but different targets. As in Experiment 1, each line shows how both agents' math levels changed, on average, over the course of training; the math level of the agent who was worse at math is plotted on the x-axis, and the math level of the agent who was better at math is plotted on the y-axis. Each point in the participants' responses denotes the participant-averaged math level of the two agents in every round of training. Each point in the model predictions denotes the weighted average math level calculated from Equation 4.9 using γ fit to individual participants. The dashed black lines indicate the minimum total math level required to win the contest. Participants' average responses aligned closely with the planning model; both have endpoints just past the dashed line, indicating that participants tended to train agents to be just good enough for the task, to maximize rewards and minimize training costs. By contrast, the alternative models failed to capture this qualitative pattern: The exploitation model trained the agent who was better at math as much as possible, the equity model trained the agent who was worse at math as much as possible, and the learning model trained the two agents for the largest possible gain in math levels. The equality model, on the other hand, did not train agents to be good enough to win the contest in seven out of eight scenarios. In addition, by horizontally comparing scenarios with the same starting math levels but different targets, we see that participants adjusted their training decisions when the target changed. The planning model was the only one that captured this pattern; all alternative models predicted the same training strategies when the target changed.

We computed correlations between the participant-averaged data and model predictions of agents' math level at every step of training. The planning model has the highest correlation with the data (Pearson's $r = 0.999, p < .0001$), followed by the equality model ($r = 0.998, p < .0001$),

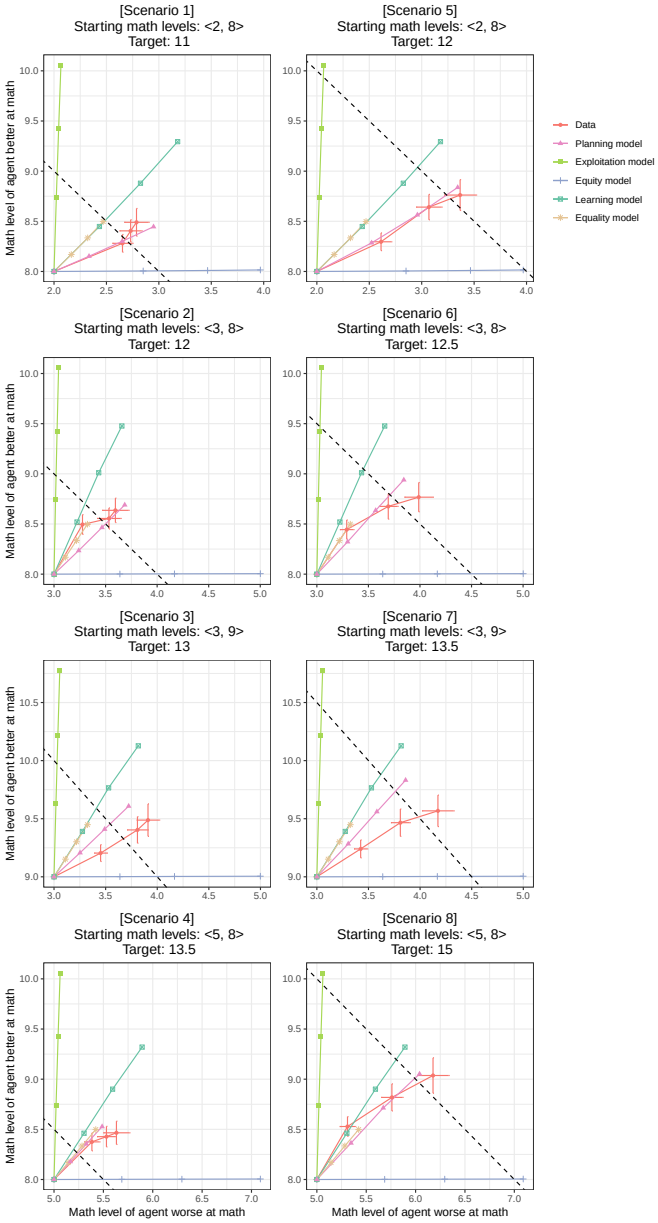


Figure 4.6: Experiment 3 data and model predictions of agents' math level in every round of training. Each subplot corresponds to one scenario, where the x-axis shows the math level of the agent who was worse at math, the y-axis shows the math level of the agent who was better at math, and each line traces agents' average math level in each training trial. The dashed black line marks the threshold where agents' combined math level exceeds the target difficulty level of the contest; intuitively, if participants are balancing the costs and benefits of training, they should train agents until their combined math level reaches this threshold, and no further. Error bars indicate 95% confidence intervals.

learning model ($r = 0.996, p < .0001$), exploitation model ($r = 0.992, p < .0001$), and equity model ($r = 0.983, p < .0001$). Note that all correlations between model predictions and data are high because agents' math levels increase monotonically, which obscures differences in how well each model captures qualitative patterns in the data. Therefore, we conducted a formal model comparison using random-effects Bayesian model selection to quantitatively assess how well each model captures participants' training strategies. We found that the planning model's protected exceedance probability was approximately 1, meaning that the planning model is the most frequently occurring model in the population. In individual participants, the planning model also had the highest model evidence (Figure 4.3C). Compared to Experiment 1, here the differences between the evidence for the planning model and equality model were more pronounced because of the four new scenarios designed to discriminate between them.

Finally, looking at individual participants' training trajectories (Figure 4.7), we see that, overall, the modal trajectories in all the scenarios were to train agents to be just good enough to win. More specifically, in scenarios where only a little training was required for winning (Scenarios 1-4), participants tended to train just one agent, whereas when a lot of training was needed (Scenarios 5-8), participants tended to either solely train the agent who was worse at math, or to provide training to both agents.

4.6.3 Discussion

In Experiment 3, we replicated the findings in Experiment 1 and showed that the findings generalized to a new and more education-orientated framing—training students for a math Olympiad. Each alternative model captures aspects of participants' responses—e.g., the equality model mirrors participants' responses in contexts where only a little training is required—but only the planning model captures the full pattern of the data.

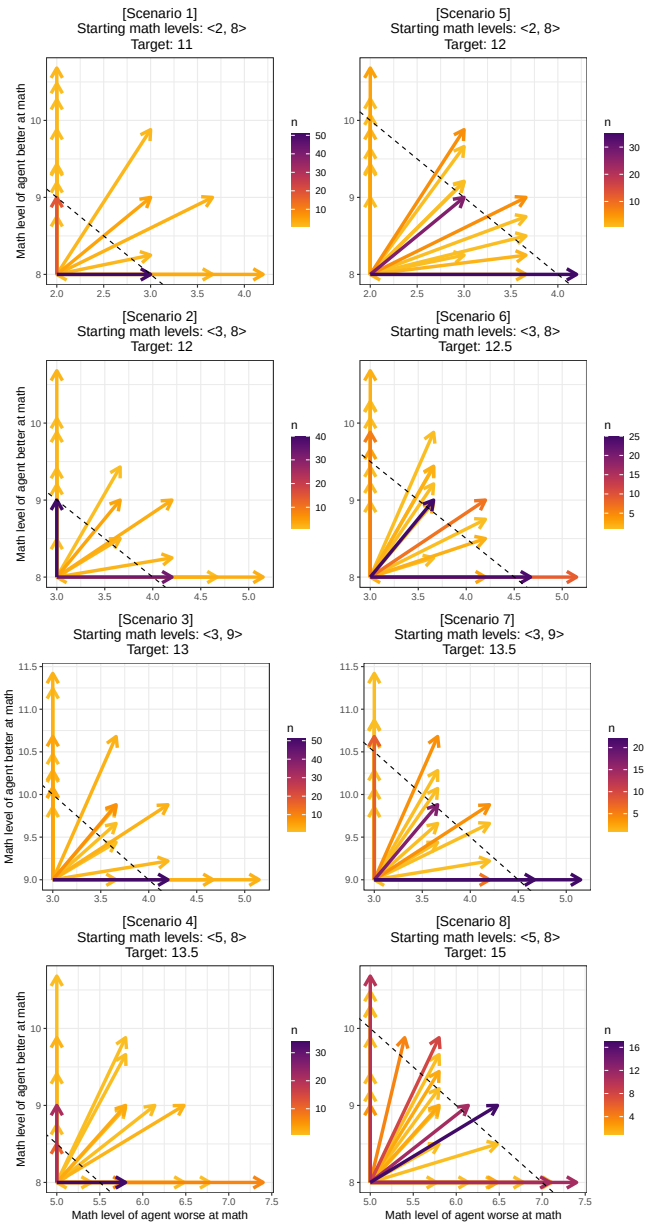


Figure 4.7: Experiment 3 individual participants' training trajectories. Each subplot denotes a scenario, and each arrow denotes a particular training strategy; the color of the arrow indicates how many participants adopted this strategy. For simplicity, here we represent each training strategy by showing only agents' math levels before (origin) and after training (arrowhead). The dashed black line marks the threshold where agents' combined math level exceeds the target difficulty level of the contest.

4.7 Experiment 4

Experiment 3 successfully replicated the results in Experiment 1 in a new context. In Experiment 4, we replicated Experiment 2 to ask whether the planning model also captures participants' decisions about whom to recruit for math Olympiads.

4.7.1 Materials and Methods

Participants

We recruited 99 participants via Amazon's Mechanical Turk platform (MTurk). Participants received a base pay of $1 and a potential bonus payment of up to $6. The bonus payment depended on which students participants recruited, and whether training enabled them to win the contest. Specifically, for every win, participants received $2; for every math level unit in the students they recruited, $0.02 was deducted from their bonus payment, regardless of the contest outcome. There were no costs associated with problem sets. The experiment was approved by the Harvard Institutional Review Board and pre-registered at https://aspredicted.org/T6Z_XCD.

Procedure

As in Experiment 3, participants imagined themselves as high school math coaches. Their school was planning to send pairs of students to participate in different math Olympiad contests; participants were asked to first recruit two students for each contest among four students who auditioned to participate, and then train them to prepare them to work out a difficult math problem together. Each student had a different starting math level and required a different amount of money to recruit. The stimuli and structure of the game were the same as in Experiment 2.

4.7.2 Results

As in Experiment 2, overall, participants recruited students who were either just good enough for the contest, or could win the contest with the least training. Among the two options, participants favored recruiting students who weren't good enough and training them to win (see the first row of Figure 4.8A), in order to minimize hiring costs. The planning model was the only model that captured these patterns (second row); none of the alternative models based their choices on whether a team would win the contest or not (third to last rows).

Correlation analyses revealed that the planning model has a very strong correlation with the data (Pearson's $r = 0.974, p < .0001$), whereas the alternative models all have weak correlation with the data (exploitation model: $r = -0.289, p = .25$; equity model: $r = -0.246, p = .32$; learning model: $r = -0.234, p = .35$; equality model: $r = -0.257, p = .30$). Random-effects Bayesian model selection further showed that the planning model has a protected exceedance probability of approximately 1, suggesting that the planning model is the most likely model in the population. The planning model also had the highest model evidence (see Figure 4.3D).

Zooming in on individual participants' training trajectories of the modal teams (<2,20> in Scenario 1, <8,40> in Scenario 2, and <20,40> in Scenario 3 (Figure 4.8B), we see that the majority of participants trained the agent who was worse at math exclusively, and just enough to win the contest.

4.7.3 Discussion

Experiment 4 replicated Experiment 2 in a new context where participants recruited and trained students for a math Olympiad. As in Experiment 2, the planning model captures participants' decisions about training *and* recruitment both qualitatively and quantitatively, but none of the alternative models does. These results suggest that the cognitive processes we uncover apply more broadly

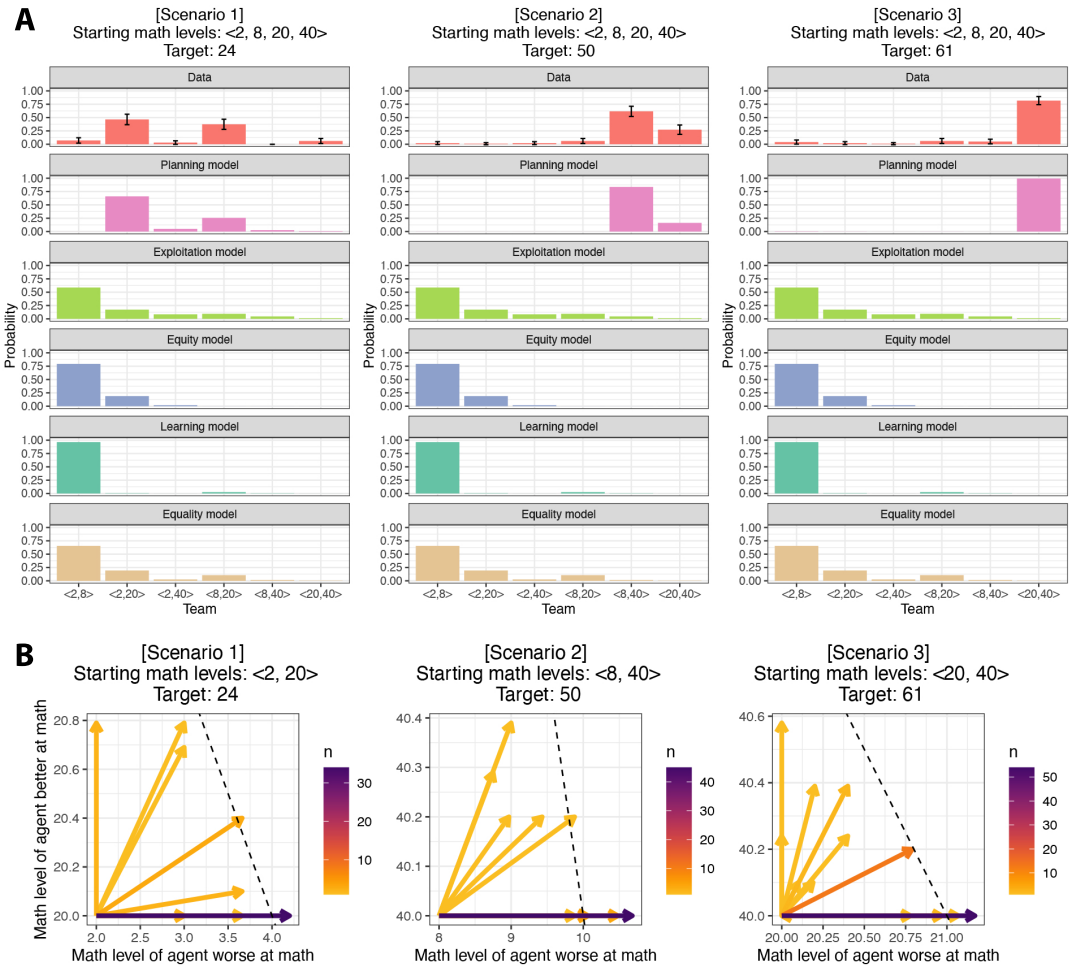


Figure 4.8: Experiment 4 results. (A) Data and model predictions, showing the probability of choosing each team. Each subplot corresponds to one scenario; the brackets on the x-axis labels refer to the math levels of the two recruited students. Error bars indicate 95% confidence intervals of proportions. (B) Individual participants' training trajectories of the modal teams. Each subplot denotes a scenario, and each arrow denotes a particular training strategy; the color of the arrow indicates how many participants adopted this strategy. For simplicity, here we represent each training strategy by showing only agents' math levels before (origin) and after training (arrowhead). The dashed black line marks the threshold where agents' combined math level exceeds the target difficulty level of the contest.

to sequential social decision-making situations, beyond the context of hiring and training workers.

4.8 General discussion

Members of a collaboration often gain and hone their skills over the course of collaborative tasks—for example, a student may become a stronger writer while drafting a manuscript, or an employee may learn to carry out her tasks more effectively on the job. However, little is known about how people account for this plasticity in collaborative decision-making. In the current work, we conducted four experiments to investigate how people decide which team members to train and how to train them (Experiments 1 and 3), and whom to recruit to a team (Experiments 2 and 4). We compared a planning model that anticipates how agents' competence will change through training to four alternative models that do not plan ahead: an exploitation model, an equity model, a learning model, and an equality model. By comparing human judgments to model predictions, we found that the planning model is the only model that captures the qualitative patterns in participants' responses and provides the best quantitative fit overall. These findings were consistent in both a physical task where contestants lifted a box together and an educational task where students solved a math problem together, demonstrating that the cognitive processes we uncover apply broadly to sequential social decision-making situations. Our results suggest that, in collaborations, people do not solely recruit the most competent agents for the job, nor do they invest solely in training while losing sight of the goal; instead, people plan ways to optimize training in the service of collaboration, maximizing the benefits to collaborators while minimizing the costs of training.

Though the alternative models performed worse than the planning model in our study, there are scenarios where the planning model makes predictions similar to the alternative models, and the heuristics of these alternative models can serve as valid shortcuts to the decision problem. For example, when the time horizon is very short, people might simply pick the more competent agents

since there is less payoff for skill development through training. In the face of a difficult task, people might opt for maximizing total improvement. And when a task requires that each team member execute an individual subtask, people might prioritize training the less competent agents. These heuristics may also be a reasonable way of hedging one's bets when it is difficult to anticipate exactly to what degree a person's competence will improve with training, or the exact difficulty of the problems that they will face in the future. However, even if these heuristics may explain how people make decisions about training within particular contexts, they do not explain why specific heuristics are appropriate within specific scenarios. By contrast, the planning model captures variability in participants' decisions across different training contexts. Thus, it is possible that the planning model provides a theoretical framework to understand the *variability* in training strategies across different, idealized scenarios. Future work needs to be done to understand how these planning models fare under uncertainty, and to develop process models that approximate planning over longer-running collaborations with larger action spaces.

Understanding variability in training strategies sheds light on the relationship between teaching and collaboration. Teaching is often defined as a modification in a teacher's behavior that enables a learner to acquire a skill or knowledge more rapidly than they would on their own, at a cost or for no immediate benefit to the teacher[34]. While this definition helps to distinguish teaching from instrumental behavior, what it misses is that the two are often intertwined: Humans often pass on skills through collaborative relationships that also benefit the mentor, such as guilds, apprenticeships, and job training. The current work therefore motivates new questions about how teachers and learners who are enmeshed in collaborative relationships balance the pedagogical benefits to a learner against the constraints and costs of the collaborative task. Viewed in the context of this work, by assigning tasks to the most competent agent, the exploitation model prioritizes meeting the demands of the collaborative task without considering who needs training, whereas the equity model and the learning model focus solely on the benefits of training and ignores the demands of the task.

Moving forward, how the nature of the collaborative relationship—such as the size of the collaborative group, or the time horizon of the collaboration—affect the degree to which we invest in training collaborators, and how these additional, instrumental concerns affect decisions about who and what is most helpful to teach, remain open questions for future research.

Our framework used a simple setting to study questions related to training and competence. However, it can be straightforwardly extended to address more complex settings. In our setup, team members had only one task to complete, whereas, in multi-task collaborative settings, team members typically have to complete a variety of tasks. For example, a research group often works on multiple different projects simultaneously and, within each project, there are usually a myriad of tasks to complete, such as devising theories, running experiments, analyzing data, and writing up the findings. Therefore, good teams need to possess some degree of versatility, which may either be from each individual being able to fit multiple roles or from building up a diverse team where different people can satisfy the demands of different tasks. Besides being versatile, realistic settings also require that teams be robust to turnover. Our design did not include attrition, but in reality, team members might drop out at any time—for example, a researcher might leave the group, and someone else on the team might need to pick up where she left off. It is thus important to train people with skills that are generally adaptive, rather than just the ability to do one thing. Further work is needed to understand how teams figure out the optimal degree of specialization.

Another limitation is that we made improvements in competence directly observable and quantifiable. These assumptions serve their purpose in our study, but realistic settings are more complex. For example, completing a project won't make a researcher look evidently more knowledgeable right away, and the researcher herself might find it difficult to tell exactly how many research skills she gained over the past year. In addition, we simplified how training changes competence by only considering increases in competence and assuming that competence increases by the level of effort people put forth in training. In the real world, people's performance may also *decrease* as a function

of fatigue and boredom, and the relationship between effort and subsequent improvements is more complex and can vary from skill to skill.

An open question for future work is how well these findings scale in more realistic conditions. One intriguing direction is to incorporate collaborators' theory of mind into the models e.g., [267]. We deliberately simplified the theory of mind requirements in our task so that we could focus on understanding how people use their intuitive theory of mind to guide training decisions. However, it leaves open the question of whether and how participants take into consideration their collaborators' beliefs and desires when making hiring and training decisions. For example, people differ in their willingness to exert effort, and if we allow that to vary in Experiment 3, we might expect students of the same math level to benefit from the same problem set differently depending on how diligent they are. Another possible extension of this work is to apply the same task structure to hiring and training real people. Our task used artificial agents; this allowed us to study the plasticity of competence in a tightly controlled experimental context, where changes in competence happen in a matter of seconds right after an action, are directly visible, and can be accurately anticipated. Even though we made an effort to simulate the experience of collaborating with real people (e.g., by framing agents as contestants in a game show or students in a math Olympiad), and our previous work [267,266] and many studies in the field e.g., [9,229,75,114,236,145] have demonstrated that adults, children, and even infants are capable of viewing artificial agents as humans, it is possible that participants might weigh their own payoffs and other factors differently, such as caring more about collaborators' learning, or about bridging the competence gap between them. Because improving competence takes time, multiple rounds of evident improvement might not be feasible in lab experiments; researchers may need to resort to large-scale long-term observational studies, such as tracking the training decisions of a company or a sports team.

Here, we present a planning model of collaboration that makes decisions about whom to recruit and how to train them by anticipating how collaborators' competence will change and what prob-

lems they'll face in the future. This work provides a first step towards understanding how humans facilitate and invest in other people's learning in a wide range of collaborative relationships, e.g., as pupils, apprentices, and trainees.

5

Discussion

In this chapter, I discuss how the Collaborative Utility Calculus connects with group-level processes, its implications for classic questions that cut across areas of cognitive science and computer science, and open questions for future research.

5.1 Relation to group-level processes

The Collaborative Utility Calculus is an individual-level account of collaboration: it specifies how individuals represent costs, rewards, competence, and effort when collaborating with others. Collaboration has also been studied at broader levels of analysis, including dyads [195,127,24,156,8], small groups [133,3,15,257], and even joint actions spread over massive networks such as care teams [55], ship

crews [108], intelligence networks [90], and self-governing online communities [172,238]. Scaling up individual-level theories of collaboration to groups introduces new challenges, including understanding how collaborators align on shared representations and what institutional and social tools they have available to maintain motivation among the group. In this section, I situate the Collaborative Utility Calculus within this problem space, clarifying how characterizing individual-level cognitive mechanisms may provide new ways to understand and intervene on group-level behaviors.

5.1.1 Understanding social motivations to collaborate

Why do people commit to collaborative goals in the first place? Many collaborative activities cannot be explained solely in terms of immediate instrumental benefits to all participants. Individuals often collaborate when acting alone is more effective to achieve the goal [49], and even young children expect agents to collaborate in situations where the same outcome could be achieved by themselves with less costs [241]. These patterns raise a foundational puzzle: What motivates people to work together, even when collaboration does not bring additional benefits to the individual or entails additional costs?

One influential answer comes from social economics, which argues that individuals' preferences extend beyond their own material payoffs: people may also derive utility from others' welfare [36], from fair distributions of payoffs [59], or from the reputational benefits of cooperation [170]. By incorporating these social concerns directly into the utility function [59,20,62,36], seemingly altruistic collaboration can remain instrumentally rational: individuals collaborate because doing so maximizes a broader conception of utility that includes social goals.

The Collaborative Utility Calculus takes a different approach by endogenizing certain forms of social information. For example, rather than adding a term for other-regarding preference in individual utility functions, the framework shows how this preference can arise from maximizing one's immediate benefits in an interdependent setting, where success requires everyone to pull their

weight. Beyond describing social concerns by adding new preference parameters, this approach provides a mechanistic explanation for how certain cooperative behaviors emerge from individual social reasoning.

5.1.2 Coordinating joint actions

How do individuals coordinate their actions in real time to achieve shared goals? Even in simple dyadic tasks—lifting a couch, dancing a waltz, passing a basketball—successful collaboration requires partners to anticipate one another, align their behaviors, and adjust dynamically as circumstances change. A large body of research on joint action has shown that collaborators form shared representations of the task and of each other[240,127], and use these representations to infer joint commitments[158] and predict upcoming actions[196,195]. This line of work identifies a core principle of shared action plans: they tend to minimize collective costs, reflecting sensitivity to joint efficiency[225,226,155].

While this literature provides a rich characterization of the structure and dynamics of coordinated behavior, it does not fully explain how individual agents arrive at a shared action plan, particularly when costs differ across collaborators. What internal trade-offs are evaluated when deciding how much effort to exert? How do differences in competence or willingness to bear costs shape the division of labor? The Collaborative Utility Calculus complements joint action research by specifying how individuals represent costs, rewards, competence, and effort within a unified utility-based framework. In doing so, it highlights that the best solution for a team depends on how costly a given action is for each collaborator. These asymmetries can be built into real-time coordination tasks to better understand how people divide work when they differ in ability or in how much effort they are willing to contribute.

5.1.3 Embedding collaboration within cultural and institutional structures

In recent years, humans have been able to collaborate at a scale that is unprecedented in our evolutionary history. We can now pursue projects that span decades and involve thousands of collaborators spread all over the world. For example, the ATLAS collaboration—a large particle physics project at CERN involved in the discovery of the Higgs boson—involves more than 5000 researchers worldwide supported by massive technological and institutional infrastructure[42]. These large-scale collaborations have been uniquely enabled by both technological innovations that facilitate communication and by institutional innovations to fairly distribute tasks, decisions, and credit among the group. Yet collaborating at this grand scale is not trivial. For every success story enabled by these technological and institutional structures, there are also untold numbers of teams floundering under unclear direction, clashing priorities, and ineffective communication. What makes large-scale collaborations possible, and why do they sometimes fail?

One influential approach to this question treats collaboration as a *distributed cognitive system* embedded in cultural and institutional structures. Theories in distributed cognition and team cognition argue that collaborations include not just the minds of individual collaborators, but also structural supports external to individual minds—such as institutional roles, routines, and material artifacts[107,108,60,248,211,46,45]. Hutchins's (1995) ethnography of Navy ship navigation provides a canonical example: the computational problem of navigation is solved by combining the efforts of individual crew members with a suite of navigational tools and communication protocols. No single crew member represents "the whole plan" in their mind. In this sense, teams can function as cognitive systems whose operations emerge from structured interaction rather than from centralized planning in any one individual. More broadly, distributing cognition across individuals and across generations also allows for rapid and transformative cultural evolution. The body of cultural

knowledge is too vast for any single individual to possess[61], and technology can be too complex for individuals to operate on their own [164]. By dividing the cognitive labor, each individual only needs to learn a subset of knowledge or skills and can act as a node in a vast distributed system that enables collaborations at scale[234].

The Collaborative Utility Calculus complements these system-level accounts by specifying the individual-level reasoning that operates within such structures. It helps explain why even collaborators embedded in the same institutional structure may behave differently due to differences in their abilities or costs. For example, two sailors occupying the same monitoring role may adopt different reporting thresholds depending on differences in expertise, fatigue, or incentives from their supervisor. Distributed cognition frameworks describe how roles and artifacts structure coordination, but they typically do not model how asymmetric costs translate into individual effort allocation or decision thresholds. The Collaborative Utility Calculus helps fill this gap by modeling how individuals reason about costs within these larger structures.

5.2 Implications

So far, I have shown that the Collaborative Utility Calculus expands the expressive power of Bayesian mentalizing models to capture a wide range of decisions that are central to collaboration, such as allocating effort, building a team, and investing in training team members. The modeling framework grounds collaboration in basic mechanisms of human social cognition, and is thus compatible with emerging work in computational cognitive science and cognitive development. In this section, I examine the developmental origins of these cognitive capacities, their implications for pedagogy and communication, and their role in supporting large-scale collaborations. I conclude by considering how the Collaborative Utility Calculus may provide a principled basis for comparison between human collaborations and emergent behaviors in artificial systems.

5.2.1 Developmental Emergence of the Collaborative Utility Calculus

The Collaborative Utility Calculus builds on three key ingredients that emerge early in development: an understanding of action efficiency, a motivation to act prosocially and jointly with others, and the ability to reason about one's own and others' competence. I trace how each of these components develops during infancy and early childhood to form the foundation for collaborative decision-making.

The first key ingredient is an early-emerging understanding of action efficiency—the expectation that agents minimize the costs they incur to achieve their goals. This principle is foundational to the Collaborative Utility Calculus, which assumes that agents weigh effort costs against rewards to make decisions. To support this reasoning, infants must first learn to interpret others' actions not just as movements, but as goal-directed behaviors. By 9 months, infants interpret others' actions as goal-directed: for example, they understand that an experimenter reaching for a toy intends to grab that toy—rather than to reach for that location—and are surprised when the experimenter continues reaching towards that location after the toy is moved. This suggests that they understand the experimenter's action as directed toward a specific goal object rather than a spatial location[255]. This understanding is accompanied by expectations about how rational agents should act. Infants as young as 12 months expect agents to take efficient paths to a goal. They are not surprised when an agent hops over a wall to get to a goal but are surprised when the agent makes unnecessary detours, such as hopping over a clear path (teleological stance[63]).

This understanding of action efficiency depends on infants representing other minds and bodies interacting with the physical world. That is, infants interpret actions as a product of an agent's internal properties (e.g., their intent) and external constraints (e.g., the location of the goal)[144]. This understanding is supported by two key insights. The first is an understanding of the physical demands of goal-directed action. Past work has found that training 3-month-old, prereaching infants

to grab objects with "sticky mittens" enables them to better recognize when other people use inefficient means to reach for an object, compared to infants who do not have this early training[209,204,66]. The second is an early-emerging understanding of people as causal agents. Prereaching infants are surprised when someone uses inefficient means to light up a toy—which causes a visible change—but not when they use inefficient means to simply lift the toy without changing its state[143].

By the end of the first year, infants can flexibly apply this understanding of action efficiency to predict how much effort an agent will exert to achieve a goal, and conversely, to infer how much an agent values a goal based on the costs they are willing to incur[146]. These inferences indicate that infants do not simply learn statistical associations between actions and outcomes; rather, they form a generative model of agents as intentional actors. This capacity is foundational to the Collaborative Utility Calculus, which posits that effort is not a generic cost, but rather emerges from the interaction between internal properties and external constraints. The early emergence of this generative model suggests that, from infancy, children are equipped to use effort as a window into others' minds.

The second ingredient is a basic set of capacities and motivations that support collaborative action. The first is an early-emerging motivation to help: toddlers will help others achieve their goals, even when they receive no immediate benefit and the person helped is a stranger[246] (but see[47]). The second is *shared intentionality*, a motivation to follow through on shared goals—instead of viewing others as tools for achieving their own goals, children commit to a joint activity and view themselves and their partners as bound by an obligation to complete it[222]. When this obligation is not met—for example, when a collaborator stops participating—toddlers make attempts to re-engage them[246]. Third, children grasp the interdependent nature of collaborative tasks: they understand that one person's contribution depends on what another person does[97] or can do[245], and that success often requires coordination between the two parties[27]. Together, these capacities create the foundation for the Collaborative Utility Calculus, enabling children to move beyond individual goal pursuit

toward coordinated action with others.

The third ingredient is the ability to reason about one's own and others' competence. While infants show early sensitivity to action costs, reasoning about competence seems to emerge later in childhood. 4- and 5-year-olds understand that a person is more competent than their partner if they complete the same task faster, or complete a more difficult task within the same time [136]. By age 5, children begin dividing physical and cognitive labor strategically, assigning easier tasks to less competent partners and harder tasks to more competent partners [153,7]. Younger children struggle with these judgments [7], suggesting that this ability requires more advanced social-cognitive reasoning. The ability to assess differences in competence and divide labor accordingly represents an important step towards the Collaborative Utility Calculus, which depends on understanding both individual abilities and how to combine them effectively.

The Collaborative Utility Calculus adds a crucial fourth ingredient: an understanding of effort in relation to one's competence. While past developmental studies have largely examined the two separately, with effort often treated as a general cost term, the Collaborative Utility Calculus highlights the need for a richer account. It explains why individuals with similar competences may contribute very different amounts of work—due to differences in their willingness to exert effort—and how effort is redistributed across team members. Thus, a promising direction for future research is to investigate how children come to represent competence and effort as distinct but interacting components, and how they use these representations to guide collaborative decisions.

5.2.2 Pedagogy, communication, pragmatics

As we have seen, the Collaborative Utility Calculus captures how people incentivize collaborators to exert effort and train them to accomplish goals just beyond their current capabilities. These properties have important implications for understanding how people teach and acquire complex skills in collaborative relationships—such as between advisor and student, or artisan and apprentice—and

thus have the potential to enrich existing theories of pedagogy and communication.

Bayesian models of cooperative communication formalize pedagogy as a series of recursive inferences between a teacher and learner. A key implication of this principle is that the cognitive processes underlying teaching and social learning are two sides of the same coin[88]: The teacher (or speaker) selects information to maximize the learner's belief in a target hypothesis, while the learner (or listener) interprets the information to infer the target hypothesis. This process can be described using mathematical theories of *optimal transport*, where the goal of communication is to efficiently transfer beliefs into the listener's mind while minimizing costs of providing information[244,201].

This principle explains an impressive range of communicative behaviors. Consistent with this framework, teachers strategically select information that is tailored to the learner's beliefs and goals[38,200,233], while learners can actively draw inferences that go beyond what teachers explicitly demonstrate[21] to evaluate whether teachers are reliable, knowledgeable, and engaged[199,208,12]. These teaching and learning behaviors emerge across a variety of modalities, including learning from examples[200], demonstrations[105], advice[235], and verbal descriptions[214]. In the domain of language, this principle is related to the *rational speech act* model[82,51], a theoretical framework that explains many instances of flexible language use, including hyperbole, generics, and polite speech[270,218,120].

Incorporating representations of competence and effort into Bayesian models of communication and pedagogy opens the door to addressing deeper puzzles about how we teach and communicate. Prior empirical work suggests that even young children can identify competent individuals. Preschoolers selectively learn from reliable teachers and understand that competence is context-specific[128,208]. For example, they recognize that other children can be better informants than adults about which toys are most fun[231], and that people who are experts in one domain might not be experts in another[152,122]. Yet, to date, little work has considered how competence changes over the course of teaching, or how competence and effort interact. Communicating with someone who is highly competent but disengaged, or highly motivated but incompetent, calls for different strategies,

yet current models rarely distinguish between these cases (but see [12]. In addition to addressing these moment-to-moment challenges, the Collaborative Utility Calculus offers an account of how humans develop and nurture each others' competence through long apprenticeships or mentorships—for example, deciding what to teach by considering not only how to convey the most information at the moment, but also anticipating how a learner's competence might evolve with practice, and what kinds of challenges are appropriate at different stages [263,38,92].

5.2.3 Large-scale collaborations

The Collaborative Utility Calculus relies on individuals reasoning about others' mental states and abilities. But in large-scale collaborations, these processes—reasoning about what each of our collaborators thinks, and what they think other collaborators think—can quickly become computationally intractable. How do these individual-level principles scale up to reasoning about many minds over an extended collaboration?

This problem is particularly salient in larger collaborations. Suppose that you are a theorist collaborating with a molecular biology lab that runs experiments. Tracking the mental state of each biologist individually may quickly become intractable. One way to simplify the process of reasoning about other minds is to abstract away from individuals and to treat whole groups as a single entity. You can represent the biology lab as an abstracted "mind" that can think, have intentions, and make plans [247,126,212], or you can represent lab members' mental states in terms of their institutional roles [112]. Another way to simplify this process is to constrain whom you reason about, and how deeply to think. You can structure the biology lab into sub-teams to limit the number of collaborators you need to represent [64], or selectively engage in recursive reasoning only when necessary [168,37,258]. Each of these strategies reduces cognitive load, but at the cost of potentially overlooking important information within the group.

To understand how individuals shift between these shortcuts and the more costly recursive rea-

soning in large-scale collaboration, new methodological approaches are required. Currently, it is difficult to create large-scale social groups in traditional laboratory experiments, or test theories of large-scale social phenomena against real-world data. Recent advances have introduced tools to study collaborations at an unprecedented scale, allowing researchers to explore how humans adapt to work together effectively in larger groups. In one study, researchers analyzed data from more than 20,000 players in One Hour One Life—a multiplayer online game where players build communities and develop technologies [238]. They found that the composition of communities was closely tied to their growth and decline, and that communities that balanced between maintaining and developing technologies tended to thrive. These findings open the door to understanding the strengths and weaknesses of different team structures, and using new approaches such as multiplayer online games as testbeds for theories of large-scale social dynamics.

The challenges of scaling collaboration are not limited to human teams. They also arise in human-AI and multi-agent artificial systems, where coordination must occur among agents with varying capabilities and access to information. The Collaborative Utility Calculus offers a benchmark for evaluating whether these systems behave like human teams: it predicts how agents with different capabilities should collaborate and distribute tasks based on inferred utility. This would allow for testing whether multi-agent artificial systems truly replicate the principles underlying human collaboration [159,160]—or merely reproduce superficial patterns (e.g., [262,16]). Conversely, multi-agent systems also offer a powerful theoretical tool for cognitive science. By engineering artificial agents with systematically varied cognitive constraints—such as limited planning horizons or restricted access to others' internal states—we can simulate the downstream consequences of individual limitations on group behavior. Multi-agent systems also allow us to examine the extent to which large-scale collaboration can emerge purely from simple task constraints (as in many multi-agent reinforcement learning systems), versus requiring the rich social reasoning instantiated in the Collaborative Utility Calculus. In this way, multi-agent systems are not only beneficiaries of cognitive theories, but also

experimental platforms for testing and refining them.

5.3 Concluding Remarks

Collaboration is a fundamental part of human life. Efforts to understand how humans collaborate have spanned decades and disciplines. Philosophical theories define the essential features that distinguish individual actions from truly collaborative activities. Evolutionary theories illuminate how collaboration can emerge and persist as a stable population-level adaptation to enhance group survival and success. Economic theories formalize the incentive structures that enable cooperation among self-interested agents. Existing psychological theories explore the cognitive mechanisms that allow individuals to infer how individuals think and act. Each of these literatures has made important contributions to a different piece of collaboration. What remains missing is a framework that connects these pieces.

Here, I propose the *Collaborative Utility Calculus*, a framework that explains how reasoning about other minds enables humans to collaborate. According to this framework, humans have an intuitive theory of competence and effort, which they knit together with their intuitive theory of belief-desire reasoning to make joint inferences and plan joint actions. I reviewed the empirical evidence supporting this framework throughout the cognitive sciences and discussed its broad implications for cognitive science, particularly in understanding how the cognitive capacities that support collaboration emerge over development, how they shape pedagogy and communication, and how they may scale to support large-scale collaborations in both human and artificial systems.

The framework rests on the premise that people have some information about their collaborators' competence and mental states. When this assumption holds, I have found that people are capable of making flexible inferences about others' competence and effort from minimal observations. Yet, when it does not hold, the Collaborative Utility Calculus may also be applied to understand

how and why collaborations break down. It remains an open question how people acquire this information when it is impossible to observe others' contributions directly, as in cross-continental scientific collaborations, or when this information may be less reliable, as when collaborators lie about their ability or effort or strategically mislead others' inferences[264]. Under these conditions of ambiguity, people may also use simpler heuristics instead[81]—such as assuming that more confident individuals are more competent, or relying on past roles as a proxy for current ability—which may circumvent the need for directly inferring these variables and require fewer computational resources. The Collaborative Utility Calculus can serve as a benchmark to understand when certain heuristics are appropriate and what they may miss in more natural, ambiguous settings—for example, when changes in a person's competence are difficult to observe directly.

Rather than looking at these inferences at a single timepoint, another fruitful direction for future work is to consider how they evolve over the course of real-time social interaction. In dynamic collaborations, people often signal their intentions through action—by taking paths that are most direct to their goals[10] or unambiguous in context[39]. Observers use these cues, along with environmental constraints, to infer what others are capable of doing and what they are trying to do[260]. Using these cues, people can coordinate effectively even without explicit communication: In real-time collaborative games, individuals converge on joint goals by tracking one another's actions[217,33] and can predict future locations even when partners briefly disappear[269]. Over time, partners can self-organize into stable, complementary roles, even when they cannot directly observe each other's actions[80]. Modeling these interactions through the lens of the Collaborative Utility Calculus may help reveal how beliefs about others' competence and effort are updated iteratively, as people observe and act with one another in real time.

The Collaborative Utility Calculus focuses on the cognitive mechanisms that support collaboration at the individual level—how people represent their own and other people's competence, effort, and mental states to plan joint actions. Complementary traditions in psychological game theory

highlight several principles that facilitate coordination at the group level. These include *common knowledge*, a shared understanding of what all parties know[219]; *salience*, the use of unique "focal points" to guide convergence on a single solution without explicit communication[193]; and *group identification*, the tendency to treat the group's success as a shared payoff[44]. Integrating these principles with the Collaborative Utility Calculus may yield a richer account of how humans navigate the challenges of collaboration under cognitive constraints. For example, common knowledge may help prevent runaway recursive reasoning about effort, while salience can reduce coordination failures by allowing collaborators to align quickly on the same plan.

Collaboration has long been seen as a hallmark of human intelligence—what separates us from other species and enables achievements that no individual could accomplish alone[99]. Today, collaboration is changing rapidly: People now work in large teams on a global scale, and artificial systems have the potential to transform how we work together. As the ways we collaborate evolve, so must our theories. The Collaborative Utility Calculus opens up new avenues for understanding the cognitive tools we have at our disposal to navigate this shifting landscape, and how we can better tailor collaborations to the strengths and limitations of the human mind.

A

Chapter 1: Supplemental Information

A.1 A safe joint effort model

Recall that in the joint effort model, the lift outcome is determined by each agent's effort, strength, and the box weight:

$$L = \mathbb{I}[\textstyle\sum_a E_a S_a \geq W], \tag{A.1}$$

where a indexes agents. To explore the hypothesis that people hedge their predictions, we create a *safe joint effort model*, where we add a hindrance factor H to Equation A.1:

$$\widetilde{L} = \mathbb{I}[\textstyle\sum_a E_a S_a \geq (W + H)]. \tag{A.2}$$

The hindrance factor H is defined as:

$$H = k \cdot \sum_a (1 - E_a), \tag{A.3}$$

where $k \geq 0$ is a scaling parameter.

The optimal effort is thus given by:

$$\mathbf{E}^* = \underset{\mathbf{E}}{\operatorname{argmax}}\, R \cdot P(\widetilde{L} = 1 | \mathbf{E}, W) - C(\mathbf{E}), \tag{A.4}$$

where $\mathbf{E} = [e_1, \ldots, e_n]$ is the vector of efforts for n agents and $C(\mathbf{E})$ is the cost function:

$$C(\mathbf{E}) = \sum_a \alpha_a E_a + \beta G(\mathbf{E}), \tag{A.5}$$

where $\beta \geq 0$ is a scaling parameter and $G(\mathbf{E})$ is the Gini coefficient, defined as half of the relative mean absolute difference in effort. In the next section, we contrast the predictions of the safe joint

effort model (k set to 3.5), the joint effort model, the joint effort model without the Gini coefficient, and behavioral data.

A.2 Variations of the joint effort model

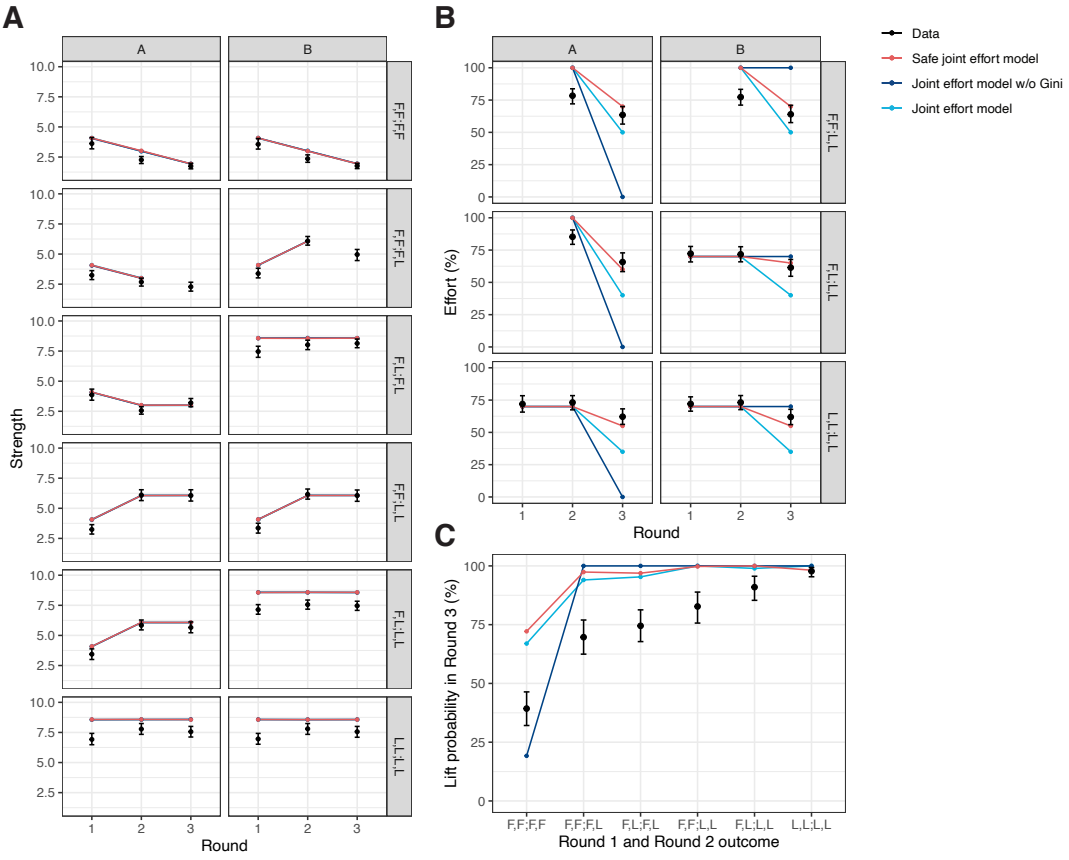


Figure A.1: (A) Predictions of contestants' strength in Experiment 1. (B) Predictions of contestants' effort in Experiment 1. Effort judgments were not elicited when the lift outcome was Fail. (C) Predictions of the lift probability in Round 3 of Experiment 1. Model simulations averaged over 10 runs. Error bars indicate bootstrapped 95% confidence intervals.

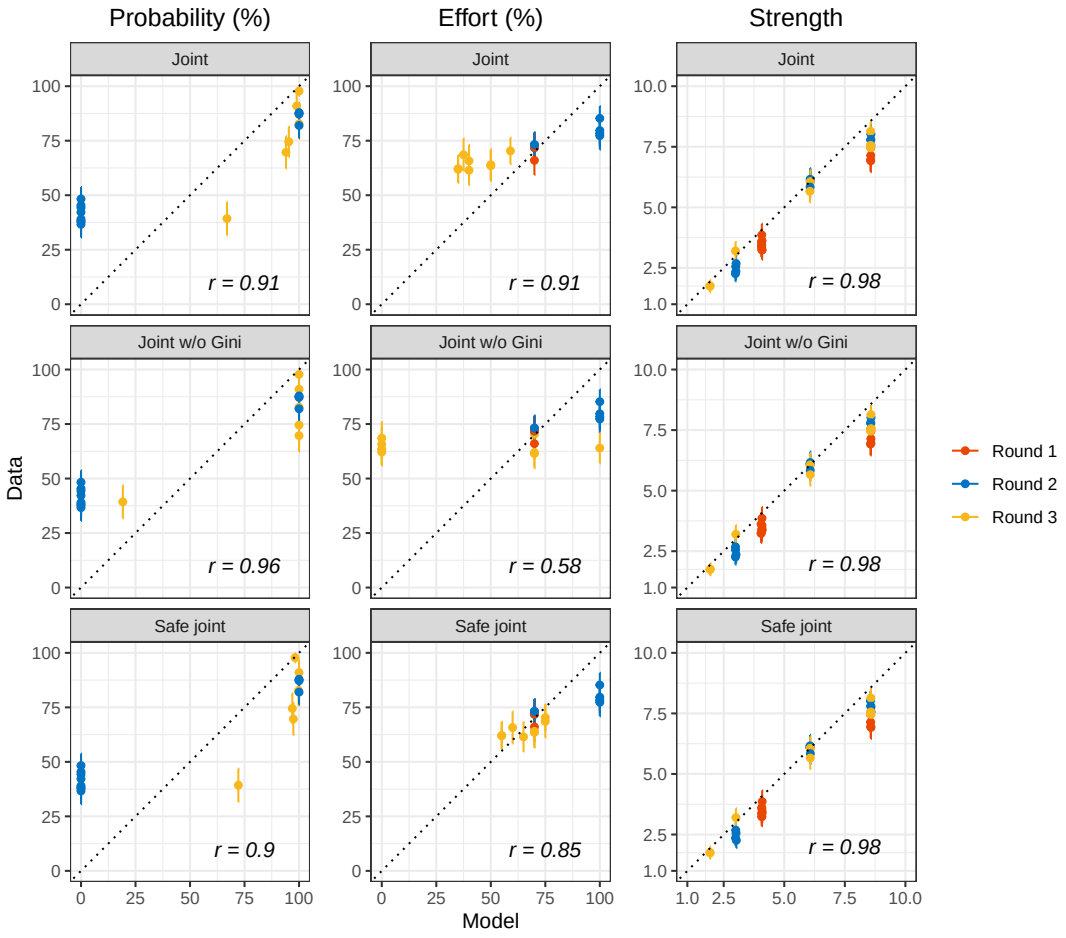


Figure A.2: Comparing model predictions to behavioral data across predictions of lift probability, effort, and strength in Experiment 1. Model simulations averaged over 10 runs. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

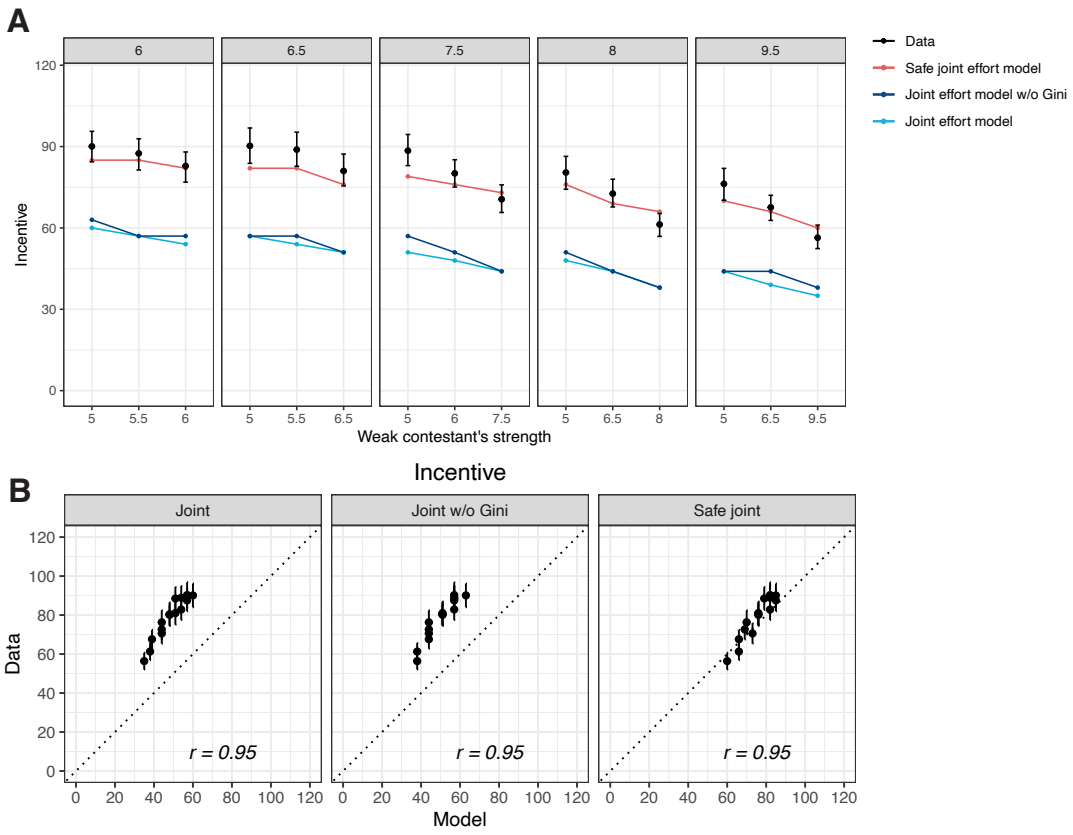


Figure A.3: (A) Incentive provided to pairs of contestants in Experiment 2. Each panel corresponds to the strongest contestant's strength in each contest. (B) Comparison between behavioral data and model predictions of incentive in Experiment 2. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars in (A) indicate bootstrapped 95% confidence intervals; error bars in (B) indicate 95% normal confidence intervals.

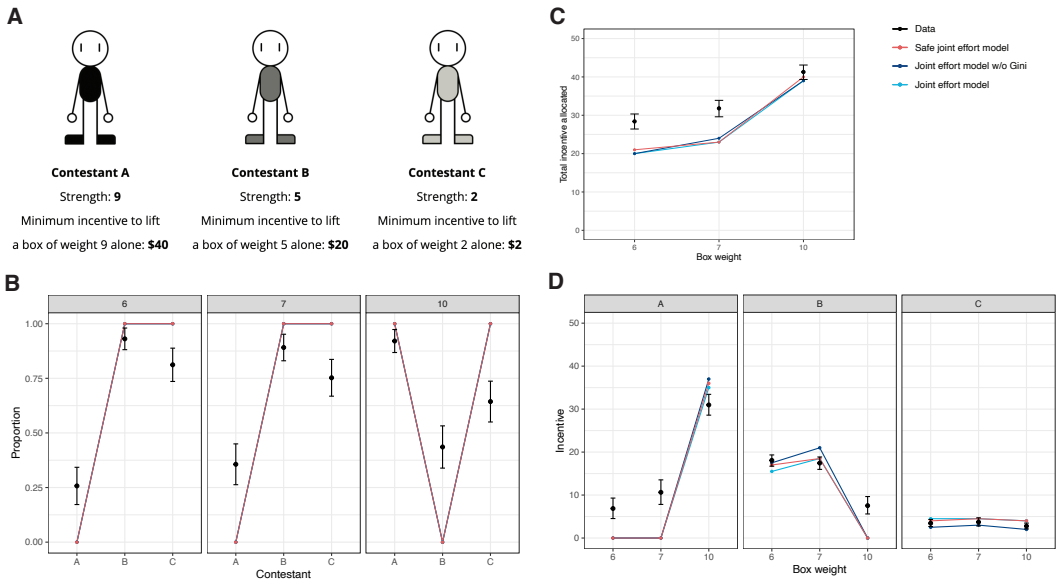


Figure A.4: (A) Contestants in Experiment 3. (B) Proportion of times each contestant was selected for each contest in Experiment 3. Contests are identified by their box weights, which appear above each plot. (C) Total incentive provided to contestants in Experiment 3. (D) Incentive provided to each contestant in Experiment 3. Each plot corresponds to an individual contestant. Error bars in (B) indicate 95% confidence intervals of proportions; error bars in (C) and (D) indicate bootstrapped 95% confidence intervals.

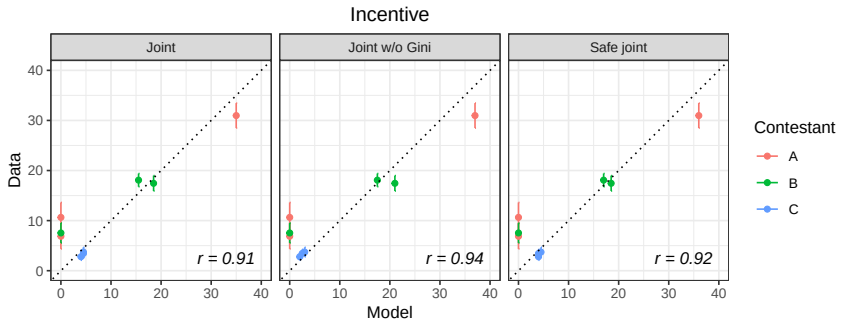


Figure A.5: Comparison between behavioral data and model predictions of incentive in Experiment 3. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

A.3 Including Gini coefficient in all models

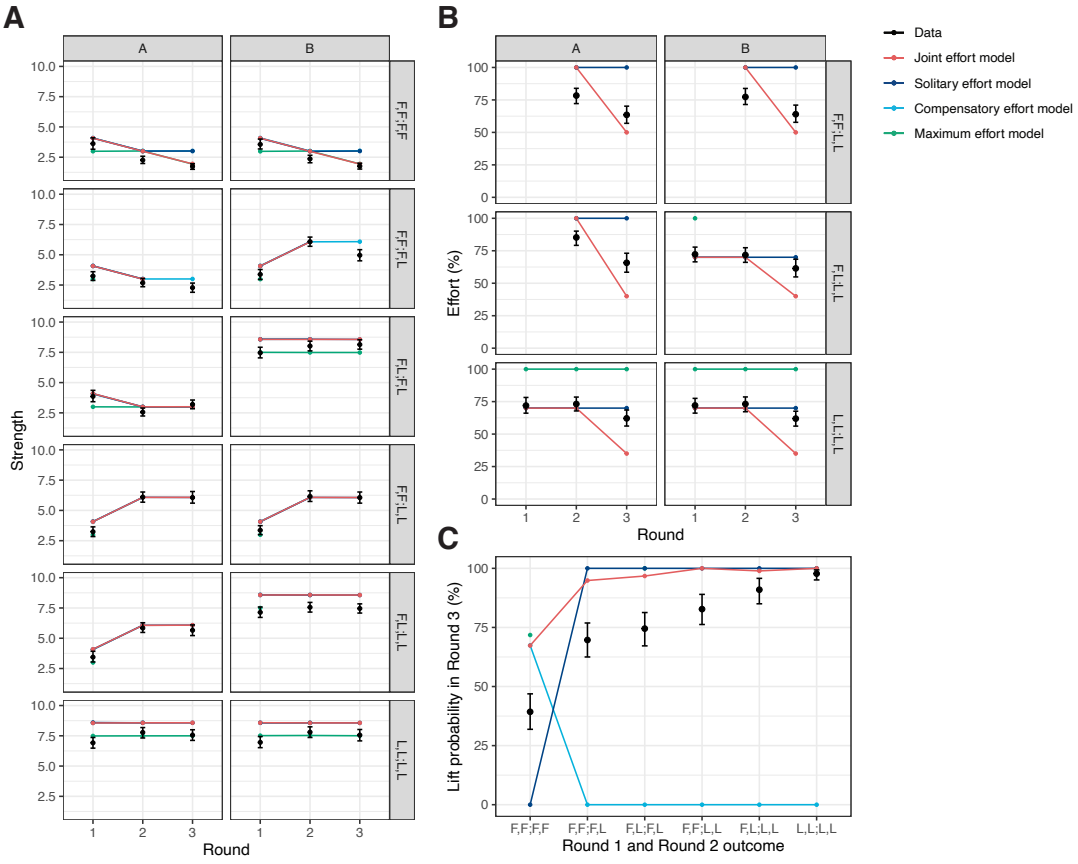


Figure A.6: (A) Predictions of contestants' strength in Experiment 1. (B) Predictions of contestants' effort in Experiment 1. Effort judgments were not elicited when the lift outcome was Fail. (C) Predictions of the lift probability in Round 3 of Experiment 1. Model simulations averaged over 10 runs. Error bars indicate bootstrapped 95% confidence intervals.

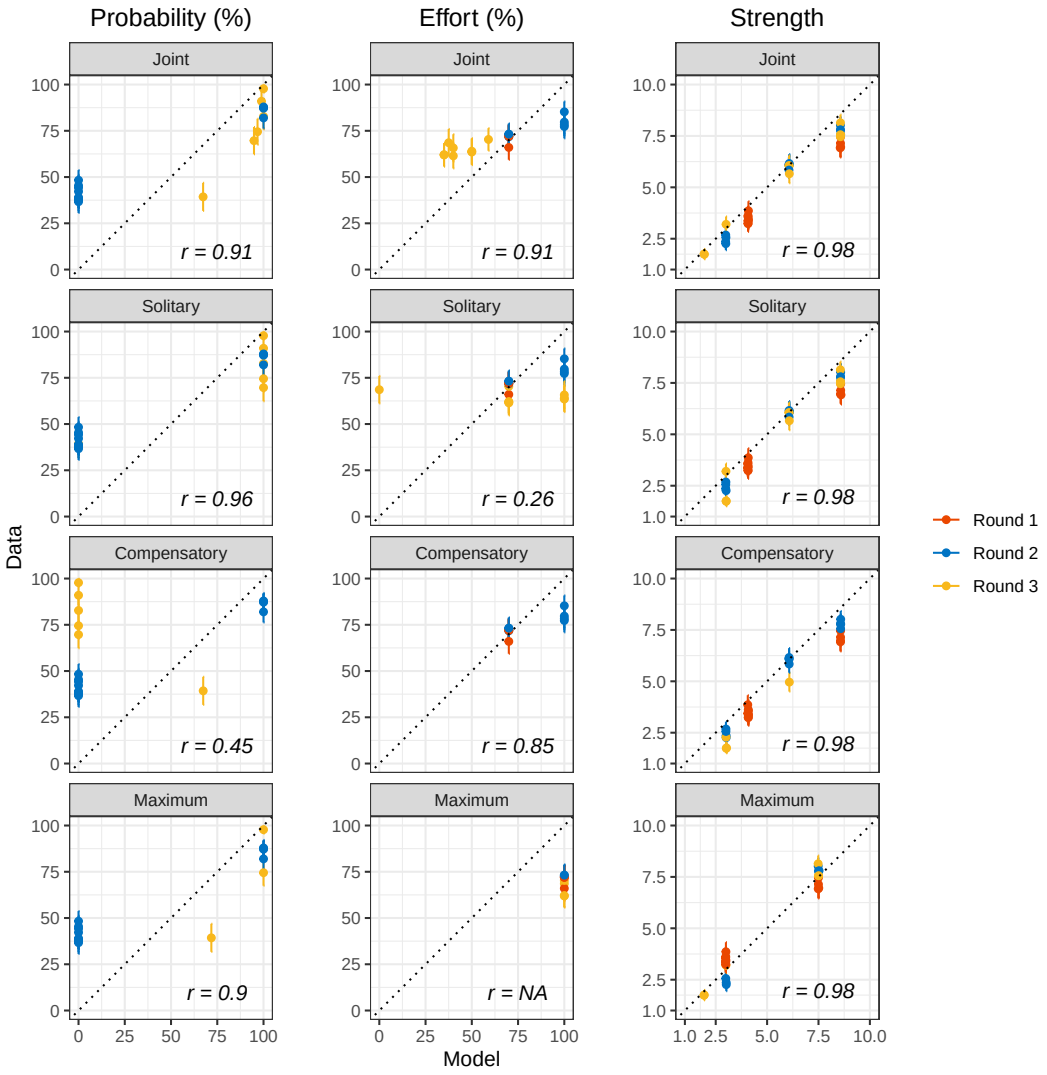


Figure A.7: Comparing model predictions to behavioral data across predictions of lift probability, effort, and strength in Experiment 1. Model simulations averaged over 10 runs. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

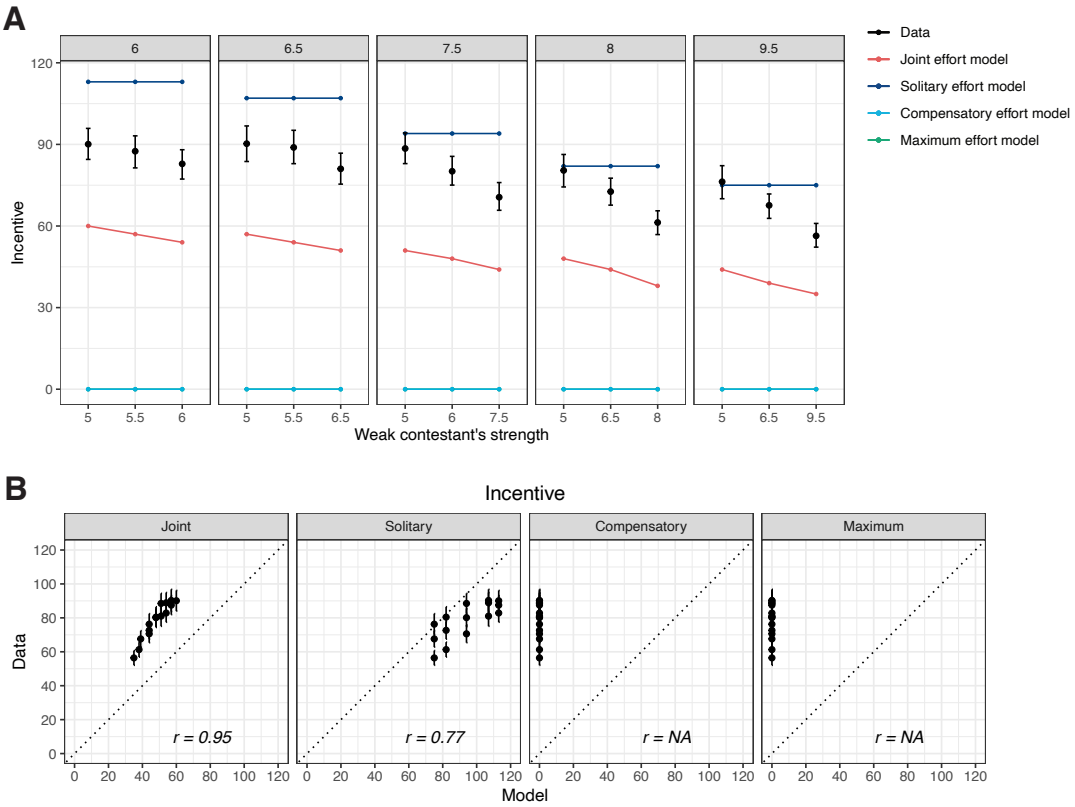


Figure A.8: (A) Incentive provided to pairs of contestants in Experiment 2. Each panel corresponds to the strongest contestant's strength in each contest. (B) Comparison between behavioral data and model predictions of incentive in Experiment 2. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars in (A) indicate bootstrapped 95% confidence intervals; error bars in (B) indicate 95% normal confidence intervals.

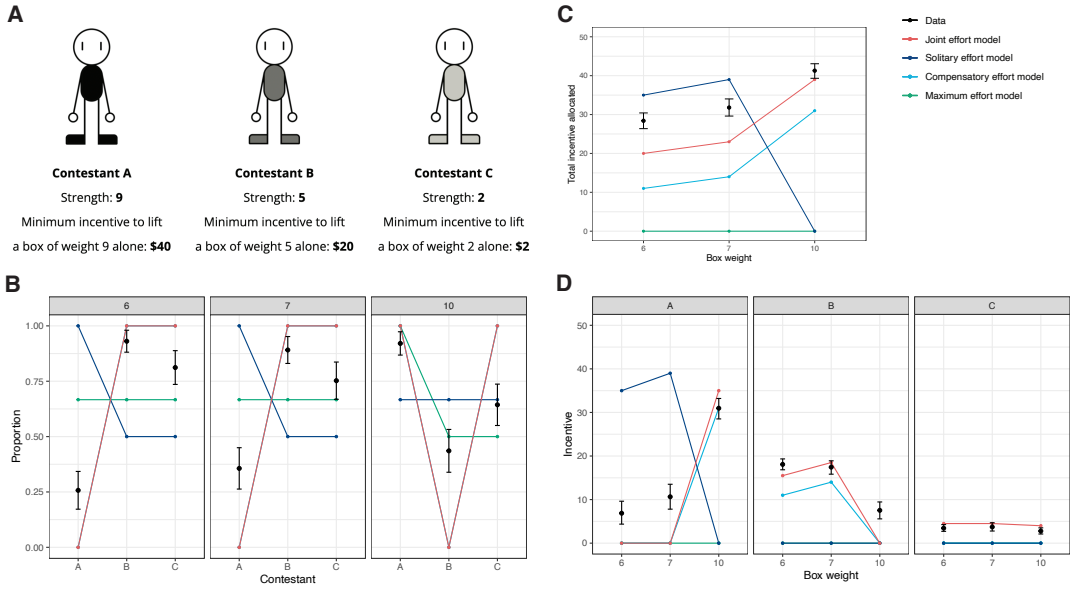


Figure A.9: (A) Contestants in Experiment 3. (B) Proportion of times each contestant was selected for each contest in Experiment 3. Contests are identified by their box weights, which appear above each plot. (C) Total incentive provided to contestants in Experiment 3. (D) Incentive provided to each contestant in Experiment 3. Each plot corresponds to an individual contestant. Error bars in (B) indicate 95% confidence intervals of proportions; error bars in (C) and (D) indicate bootstrapped 95% confidence intervals. Predictions of the compensatory effort model in (B) is covered by the joint effort model. This is because when the agent chooses not to act, the penalty for unequal effort allocation is very high, such that the compensatory effort model predicts that an agent would act because they need to avoid the high penalty, rather than a result of recursive reasoning. Decreasing the β value will decrease the penalty, therefore resulting in a similar prediction as in the main text. Predictions of the compensatory effort model in (D) is covered by the solitary effort model (predicting $0 for Contestant C across all contests).

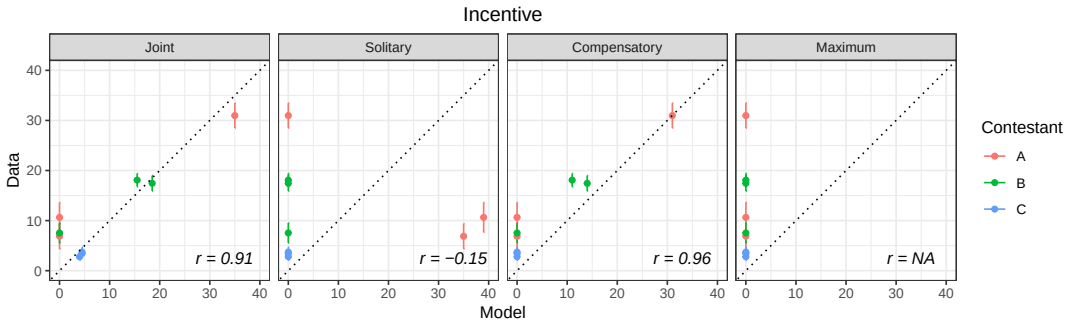


Figure A.10: Comparison between behavioral data and model predictions of incentive in Experiment 3. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

A.4 Experiment instructions

A.4.1 Experiment 1

Welcome!

In this experiment, you will watch a game show called *Lift That Box!* and answer some questions.

The game show consists of different contests between different pairs of contestants.

In each contest, the contestants will be given **three attempts** to lift a box.

We will show you whether they succeeded in lifting the box after each attempt.

In the **final round** of each contest, the two contestants will try to lift the box **together**.

Page Break

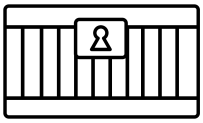
The weight of the box is always the same, across all rounds and contests.

To give you an idea of how heavy the box is:

Suppose we rated the weights of objects on a scale of 1 to 10, where 1 is a weight so light that virtually anyone can lift it, and 10 is a weight so heavy that only the strongest humans can lift it.

We loaded the box so that it would measure a 5 on this scale.

In other words: An average person (strength 5) should be able to lift the box if they put in an all-out, 100% effort, and an extremely strong person (strength 10) should be able to lift the box with only a moderate, 50% effort.



Page Break

Contestants can win **$10** or **$20** for lifting the box, as indicated on the box.

In the final round, if they lift the box together, **each of them gets the prize money** shown on the box.

The prizes change from round to round, but the box doesn't change.

Page Break

You will predict how likely the contestants are to lift the box, observe the actual outcome, and answer questions regarding their strength and effort.

When making your guesses, you will see a table of all the previous outcomes.

By the end of the game, we will randomly pick a round to calculate your bonus.

If the outcome on that round is Lift, then your bonus will be the probability of lifting that you guessed.

If the outcome on that round is Fail, then your bonus will be 1 minus the probability of lifting that you guessed.

Contestant	I	J
Round 1 ($10)	Fail	Lift
Round 2 ($20)	Lift	Lift
Round 3 ($20 for each)	?	

Page Break

Keep in mind that some people are stronger and some are weaker.

Even two contestants with the same strength may have different outcomes (some people are lazier and need higher reward to do the same thing!).

The contestants know very well their own strength and the other contestant's strength.

They also know how heavy the box is.

Page Break

Also remember that **the box is always the same**.

Different pairs of contestants will come in to lift the box; each pair will play the game for three consecutive rounds.

If the contestant does not feel the prize is worth their effort, they might not lift the box even if they are strong enough to do so.

Note that effort is costly; If the contestant can lift the box with 50% effort, they would not exert more effort than needed.

A.4.2 Experiment 2

Welcome!

In this experiment, you will watch a game show called *Lift That Box!* and answer some questions.

The game show consists of different contests between different teams of contestants.

We will change all of the contestants at the end of each contest.

In each round of the contest, two contestants will be given an attempt to lift a box together.

Page Break

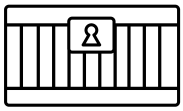
The weight of the box is always the same, across all rounds and contests.

To give you an idea of how heavy the box is:

Suppose we rated the weights of objects on a scale of 1 to 10, where 1 is a weight so light that virtually anyone can lift it, and 10 is a weight so heavy that only the strongest humans can lift it.

We loaded the box so that it would measure a 5 on this scale.

In other words: An average person (strength 5) should be able to lift the box if they put in an all-out, 100% effort, and an extremely strong person (strength 10) should be able to lift the box with only a moderate, 50% effort.



Page Break

Contestants can win a certain amount of prize money as an incentive for lifting the box.

We will tell you the minimum incentive that each contestant would accept to lift the box alone.

Your job is to decide how much incentive you would like to provide each pair of contestants to lift the box together.

If they lift the box together, **each contestant gets the incentive** you provide.

For example, if you provide an incentive of $50, then each contestant will get $50 if they lift the box together.

No matter how the prizes change, the box is always the same.

Page Break

For every successful lift, you will get a bonus of $0.4.

For every dollar of incentive you give, $0.002 will be deducted from your bonus payment, regardless of whether the result is Lift or Fail.

Suppose you provide $50 for each contestant to lift a box.

If they succeed, then your bonus payment increases by $0.4-50x$0.002 = $0.3.

If they fail, you lose 50x$0.002 = $0.1.

Your total bonus payment can go up to $6, and will also not be less than $0.

You will see whether the contestants lift the box or not with the incentive you gave by the end of each contest.

Page Break

Keep in mind that some people are stronger and some are weaker.

The box is always the same, its weight equivalent to strength of 5.

The contestants know very well their own strength and the other contestant's strength.

They also know how heavy the box is.

Page Break

Also remember that the more effort a contestant needs to lift the box, the higher the incentive they will require.

In other words, stronger contestants can lift the box with less effort, therefore they require lower incentives.

Note that **effort is costly**; If the contestant does not feel the prize is worth their effort, they might not lift the box even if they are strong enough

A.4.3 Experiment 3

Welcome!

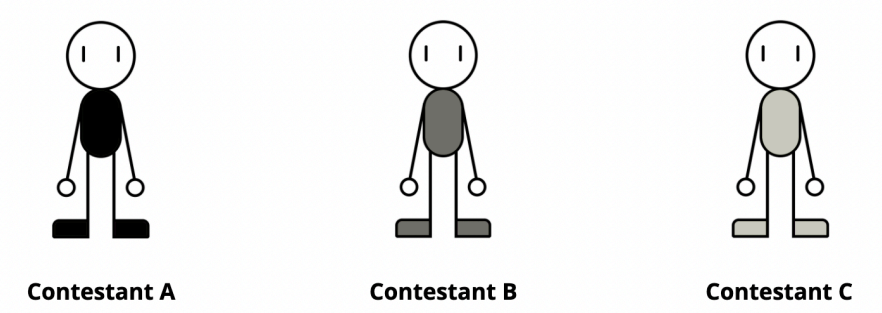
In this experiment, you will watch a game show called *Lift That Box!* and answer some questions.

The game show consists of different contests between different teams of contestants.

There are three different contestants altogether.

In each contest, two of the three contestants will be given an attempt to lift a box together.

We will change the box at the end of each contest.



Page Break

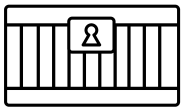
The weight of the box is converted onto the same scale as people's strength.

Suppose we rated the weights of objects on a scale of 1 to 10,

where 1 is a weight so light that virtually anyone can lift it, and 10 is a weight so heavy that only the strongest humans can lift it.

A box of weight 1 would be extremely easy to lift, and a box of weight 10 would be extremely hard to lift.

A box of weight 5 would be liftable by an average person (strength 5) if they put in an all-out, 100% effort, and an extremely strong person (strength 10) should be able to lift the box with only a moderate, 50% effort.



Page Break

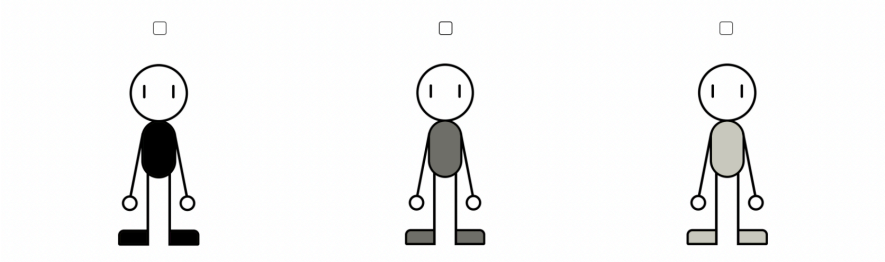
Contestants can win a certain amount of prize money as an incentive for lifting the box.

We will tell you the strength of each contestant and the minimum incentive they would accept to lift the heaviest box they can lift by themselves.

Your job is to **select two contestants** to lift the box together in each contest, and **decide how much incentive** you would like to provide each contestant you selected.

You will only provide incentives to the two contestants you selected.

The total incentive you can provide to the two contestants of your choice is **$50**.



Contestant A	**Contestant B**	**Contestant C**
Strength: **9**	Strength: **5**	Strength: **2**
Minimum incentive to lift a box of weight 9 alone: **$40**	Minimum incentive to lift a box of weight 5 alone: **$20**	Minimum incentive to lift a box of weight 2 alone: **$2**

Page Break

For every successful lift, you will get a bonus of $1.

For every dollar of incentive you give, $0.01 will be deducted from your bonus payment, regardless of whether the result is Lift or Fail.

Suppose you provide $30 to one contestant and $10 to the other contestant. If they succeed, then your bonus payment increases by $1-(30+10)x$0.01 = $0.6. If they fail, you lose (30+10)x$0.01 = $0.4.

Your total bonus payment can go up to $3, and will also not be less than $0.

You will see whether the contestants lift the box or not with the incentive you give by the end of each contest.

Page Break

Keep in mind that some people are stronger and some are weaker.

The available contestants are always the same, but **the box changes from contest to contest**.

The contestants know very well their own strength and the other contestant's strength.

They also know how heavy the box is.

Page Break

Also remember that some contestants consider their effort more costly than others.

In other words, some contestants might require more incentive to exert a certain amount of effort.

If the contestant does not feel the prize is worth their effort, they might not lift the box even if they are strong enough to do so.

B

Chapter 2: Supplemental Information

B.1 Comparing Ensemble models with different weightings

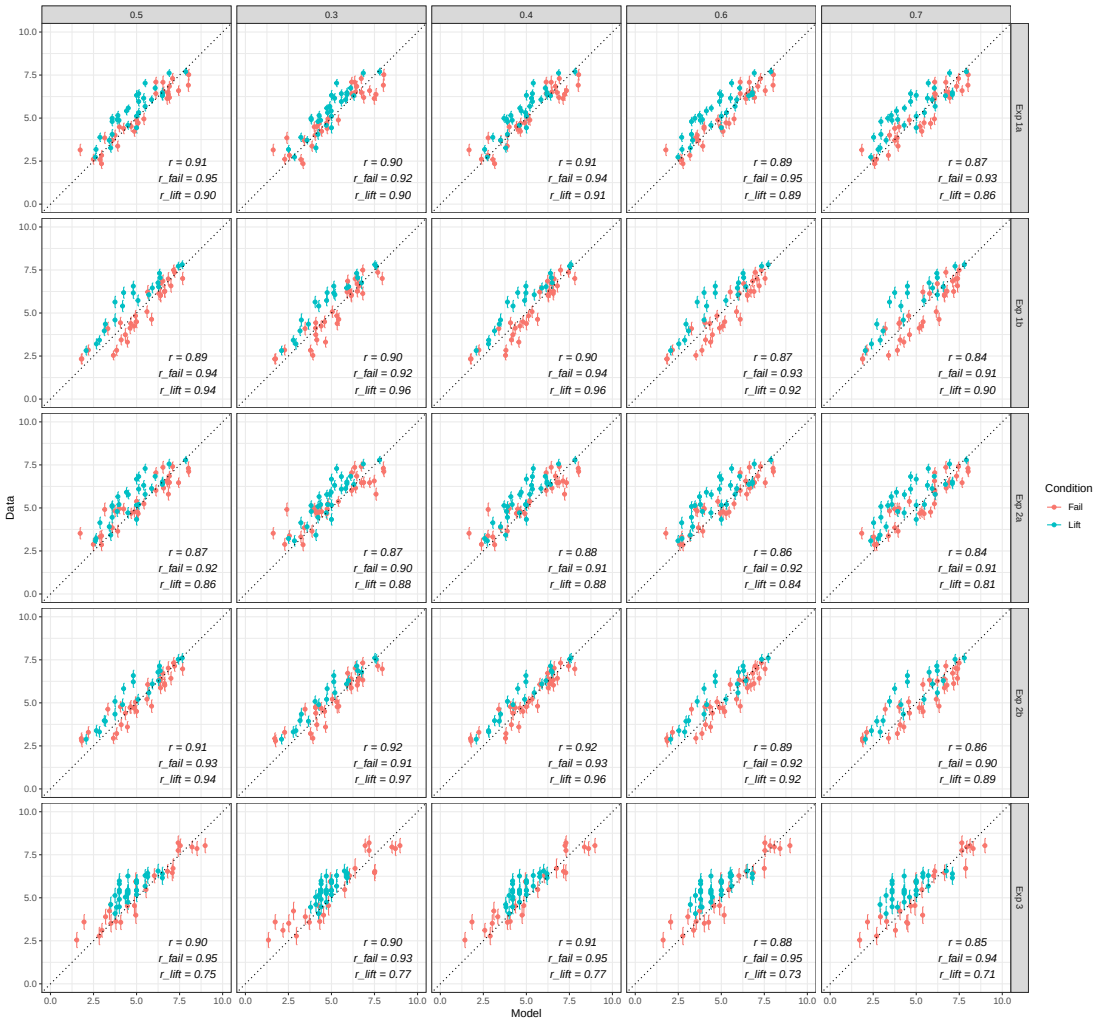


Figure B.1: Comparison between behavioral data and predictions of Ensemble models with different weightings. Each panel corresponds to the *w* in each model. The leftmost panel ($w = 0.5$) is the equal-weighting Ensemble model we implemented in the main text. Pearson correlation coefficient for data 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% normal confidence intervals.

B.2 Stimuli

Table B.1: Experiments 1a and 2a stimuli.

Scenario	Condition	Match	Box weight	Strength 1	Strength 2	Effort 1	Effort 2
1	fail	strength	5	5	5	0.2	0.6
2	fail	strength	7	6	6	0.3	0.8
3	fail	strength	5	4	4	0.3	0.6
4	fail	strength	6	5	5	0.2	0.7
5	fail	strength	8	8	8	0.2	0.4
6	fail	effort	6	4	8	0.4	0.4
7	fail	effort	7	4	6	0.6	0.6
8	fail	effort	6	2	8	0.5	0.5
9	fail	effort	9	5	7	0.7	0.7
10	fail	effort	8	6	8	0.3	0.3
11	fail	force	3	2	6	0.6	0.2
12	fail	force	5	3	6	0.8	0.4
13	fail	force	5	4	6	0.3	0.2
14	fail	force	9	5	8	0.8	0.5
15	fail	force	8	6	8	0.4	0.3
16	lift	strength	4	5	5	0.2	0.6
17	lift	strength	5	6	6	0.3	0.8
18	lift	strength	3	4	4	0.3	0.6
19	lift	strength	4	5	5	0.2	0.7
20	lift	strength	4	8	8	0.2	0.4
21	lift	effort	4	4	8	0.4	0.4
22	lift	effort	5	4	6	0.6	0.6
23	lift	effort	5	2	8	0.5	0.5
24	lift	effort	6	5	7	0.7	0.7
25	lift	effort	3	6	8	0.3	0.3
26	lift	force	2	2	6	0.6	0.2
27	lift	force	3	3	6	0.8	0.4
28	lift	force	2	4	6	0.3	0.2
29	lift	force	6	5	8	0.8	0.5
30	lift	force	4	6	8	0.4	0.3

Note: For this and subsequent experiments, Strength 1 and Strength 2 refer to the strength of the two contestants, and Effort 1 and Effort 2 refer to the effort of the two contestants. The order of the stimuli was randomized during the experiment. The positioning of contestants was randomized.

Table B.2: Experiments 1b and 2b stimuli.

Scenario	Condition	Match	Box weight	Strength 1	Strength 2	Effort 1	Effort 2
1	fail	strength	8	5	5	0.2	0.6
2	fail	strength	8	6	6	0.3	0.8
3	fail	strength	6	4	4	0.3	0.6
4	fail	strength	8	5	5	0.2	0.7
5	fail	strength	10	8	8	0.2	0.4
6	fail	effort	8	4	8	0.4	0.4
7	fail	effort	8	4	6	0.6	0.6
8	fail	effort	8	2	8	0.5	0.5
9	fail	effort	10	5	7	0.7	0.7
10	fail	effort	9	6	8	0.3	0.3
11	fail	force	6	2	6	0.6	0.2
12	fail	force	6	3	6	0.8	0.4
13	fail	force	7	4	6	0.3	0.2
14	fail	force	10	5	8	0.8	0.5
15	fail	force	9	6	8	0.4	0.3
16	lift	strength	2	5	5	0.2	0.6
17	lift	strength	4	6	6	0.3	0.8
18	lift	strength	2	4	4	0.3	0.6
19	lift	strength	3	5	5	0.2	0.7
20	lift	strength	2	8	8	0.2	0.4
21	lift	effort	2	4	8	0.4	0.4
22	lift	effort	3	4	6	0.6	0.6
23	lift	effort	3	2	8	0.5	0.5
24	lift	effort	4	5	7	0.7	0.7
25	lift	effort	2	6	8	0.3	0.3

Table B.3: Experiment 3 stimuli.

Scenario	Condition	Box weight	Strength 1	Strength 2	Effort 1	Effort 2
1	fail	5	2	8	0.5	0.4
2	fail	8	4	8	0.75	0.5
3	fail	5	4	10	0.5	0.2
4	fail	7	5	8	0.8	0.25
5	fail	4	5	10	0.4	0.1
6	fail	9	8	10	0.75	0.2
7	fail	5	3	10	0.5	0.2
8	fail	8	6	8	0.5	0.4
9	fail	4	3	10	0.4	0.1
10	fail	8	4	6	0.5	0.75
11	fail	10	4	8	0.5	0.75
12	fail	10	6	10	0.1	0.5
13	lift	4	4	5	0.5	0.4
14	lift	4	4	10	0.5	0.2
15	lift	5	4	10	0.5	0.3
16	lift	6	4	10	0.75	0.3
17	lift	8	5	8	0.8	0.5
18	lift	5	5	10	0.4	0.3
19	lift	6	5	10	0.6	0.3
20	lift	7	5	10	0.6	0.4
21	lift	9	5	10	0.8	0.5
22	lift	8	8	10	0.5	0.4
23	lift	5	4	8	0.5	0.5
24	lift	4	4	10	0.25	0.3
25	lift	5	5	8	0.4	0.5
26	lift	6	5	10	0.4	0.4
27	lift	5	8	10	0.25	0.3
28	fail	5	5	16	0.2	0.1
29	fail	12	6	8	0.75	0.75
30	fail	14	6	10	0.7	0.9

Note: Scenarios marked gray were impossible scenarios excluded from our analysis.

B.3 Actual- and Counterfactual-contribution model predictions (line plots of Experiments 1a and 1b)

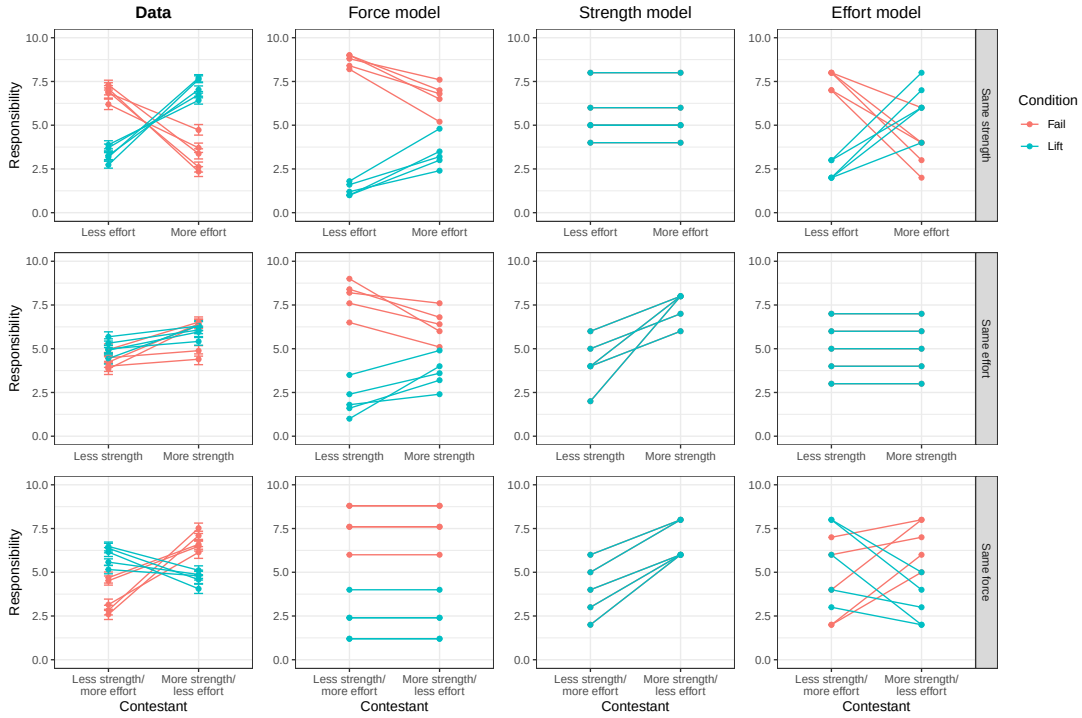


Figure B.2: Data and predictions of the Force model, Strength model, and Effort model in Experiment 1a. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

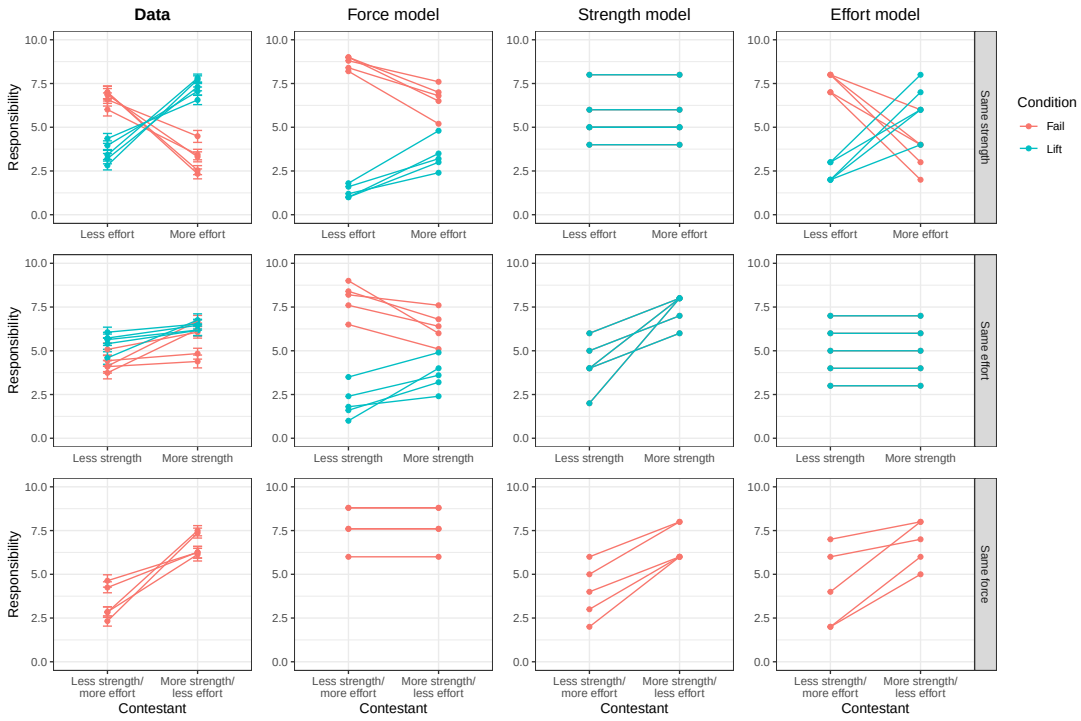


Figure B.3: Data and predictions of the Force model, Strength model, and Effort model in Experiment 1b. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

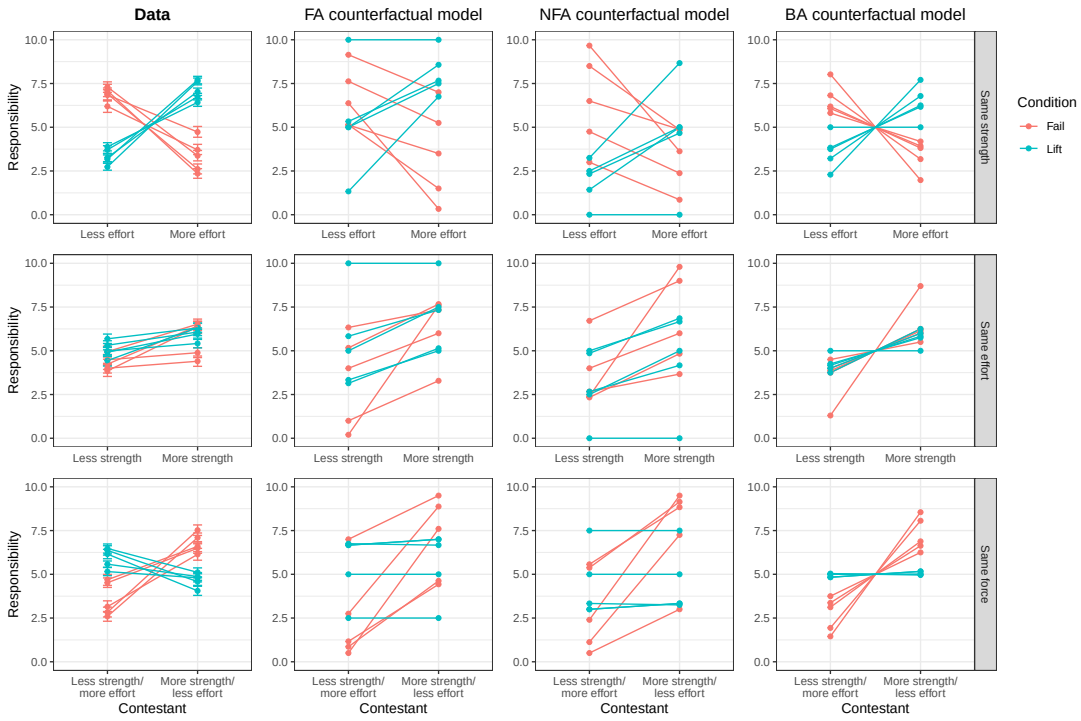


Figure B.4: Data and predictions of the Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models in Experiment 1a. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

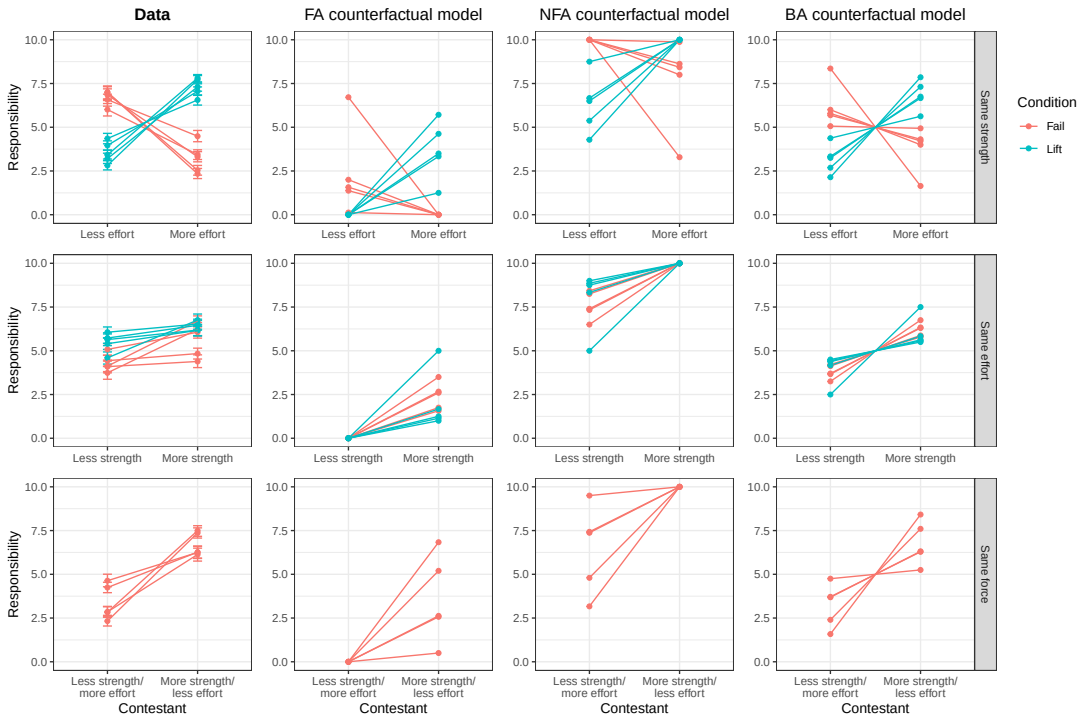


Figure B.5: Data and predictions of the Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models in Experiment 1b. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

B.4 Actual- and Counterfactual-contribution model predictions (scatter plots)

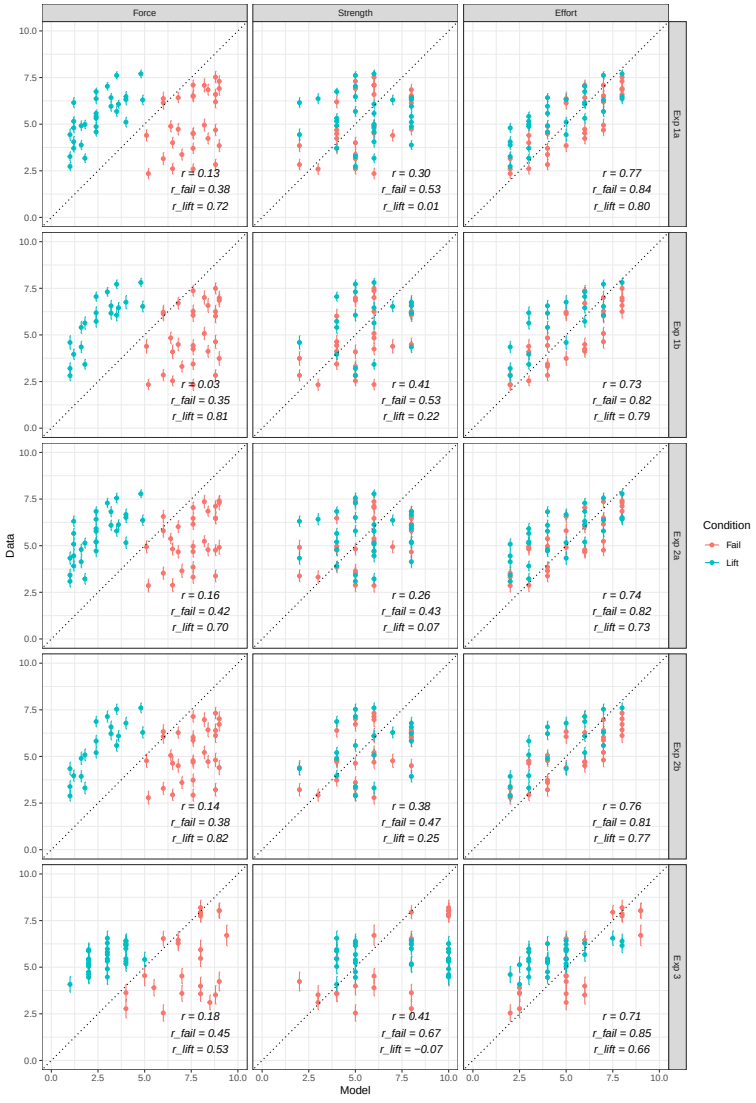


Figure B.6: Comparison between behavioral data and predictions of the Force model, Strength model, and Effort model. Pearson correlation coefficient for data and model predictions pooled across both conditions and for each condition are shown at the bottom right of each subplot. Each dot denotes one agent in one scenario; error bars indicate 95% normal confidence intervals.

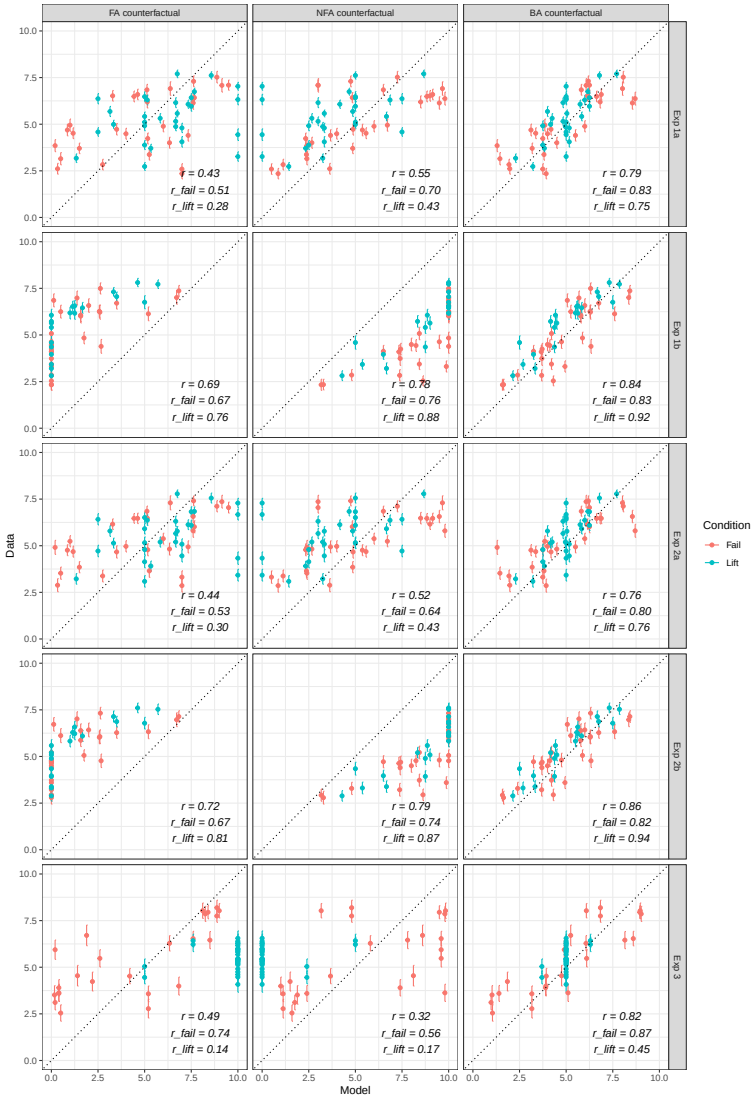


Figure B.7: Comparison between behavioral data and predictions of the Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models. Pearson correlation coefficient for data and model predictions pooled across both conditions and for each condition are shown at the bottom right of each subplot. Each dot denotes one agent in one scenario; error bars indicate 95% normal confidence intervals.

B.5 Experimental setup in Experiments 2a and 2b

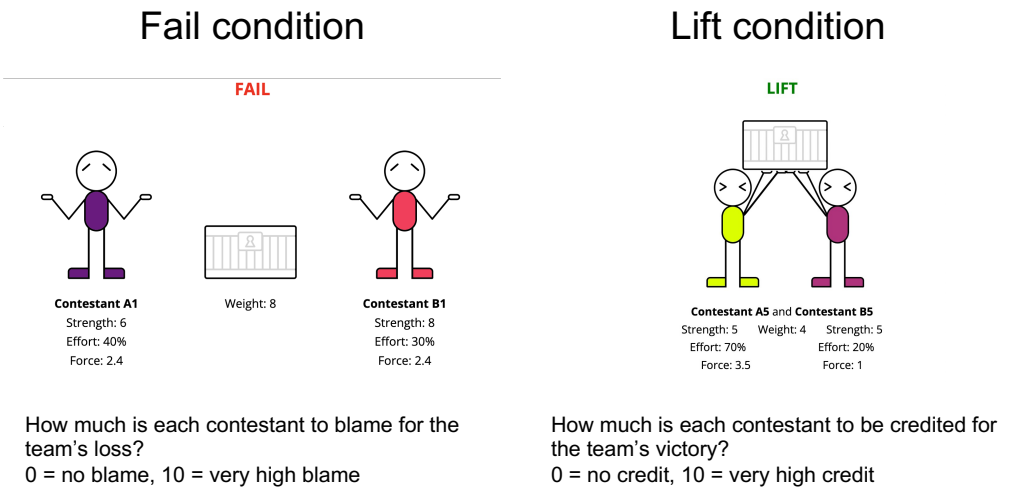


Figure B.8: Experimental setup used in Experiments 2a and 2b. Participants observed each contestant's strength, effort, and force, the weight of the box, and whether the contestants successfully lifted the box together (Lift condition) or not (Fail condition), and they then assigned blame or credit to each contestant.

B.6 Actual- and Counterfactual-contribution model predictions (line plots of Experiments 2a and 2b)

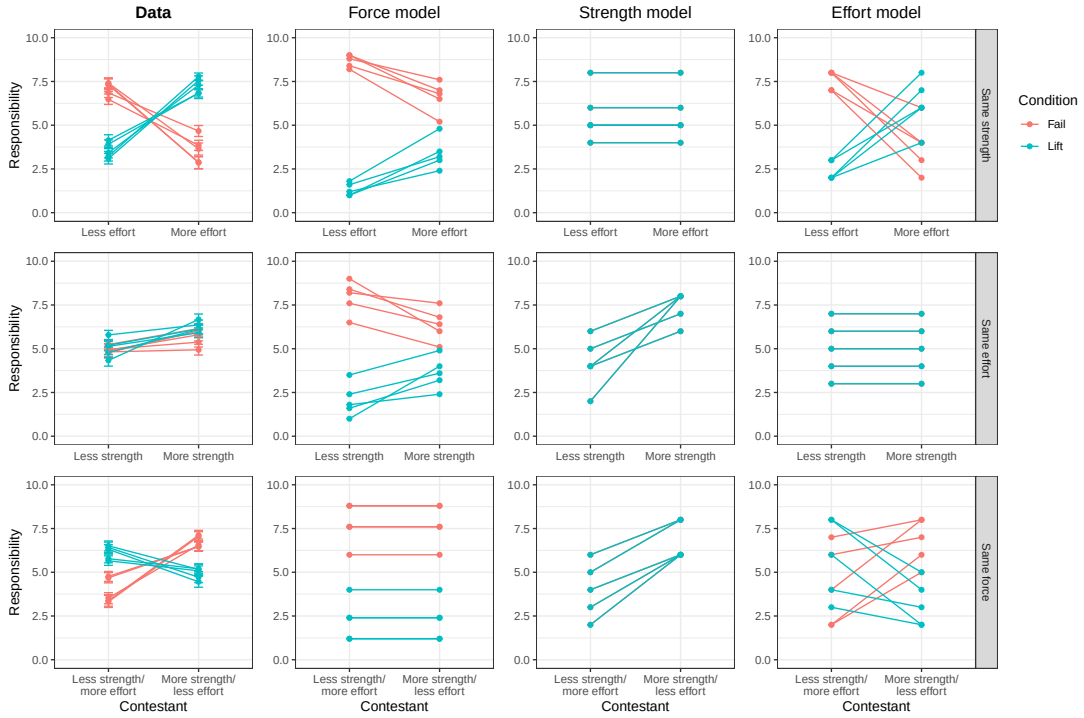


Figure B.9: Data and predictions of the Force model, Strength model, and Effort model in Experiment 2a. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

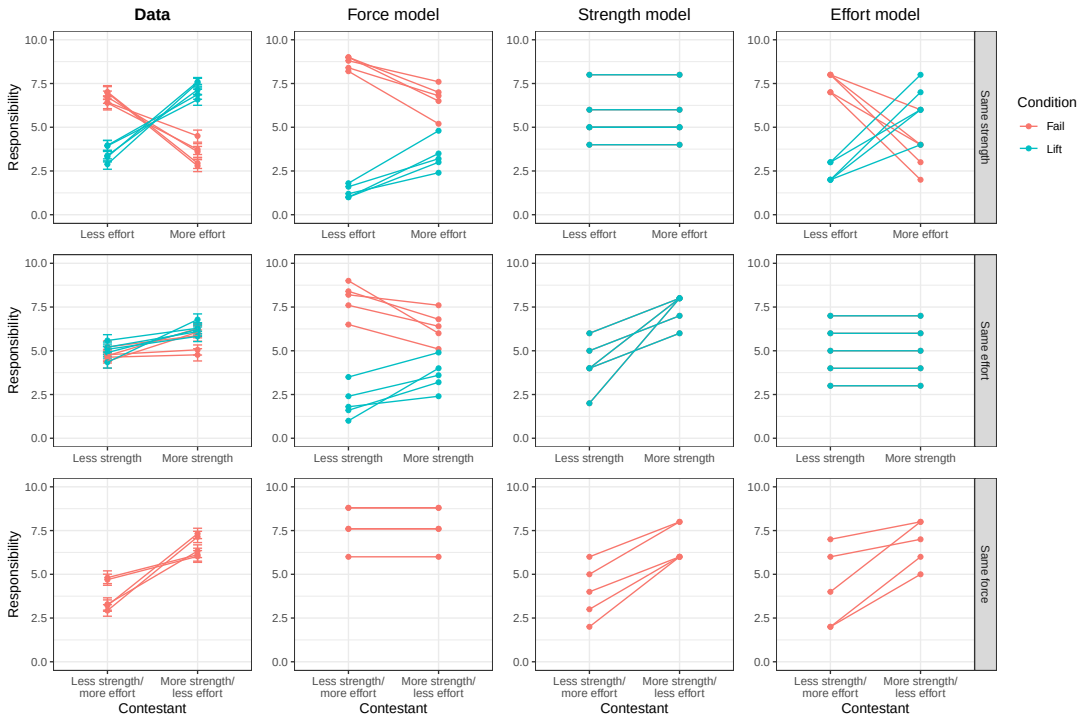


Figure B.10: Data and predictions of the Force model, Strength model, and Effort model in Experiment 2b. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

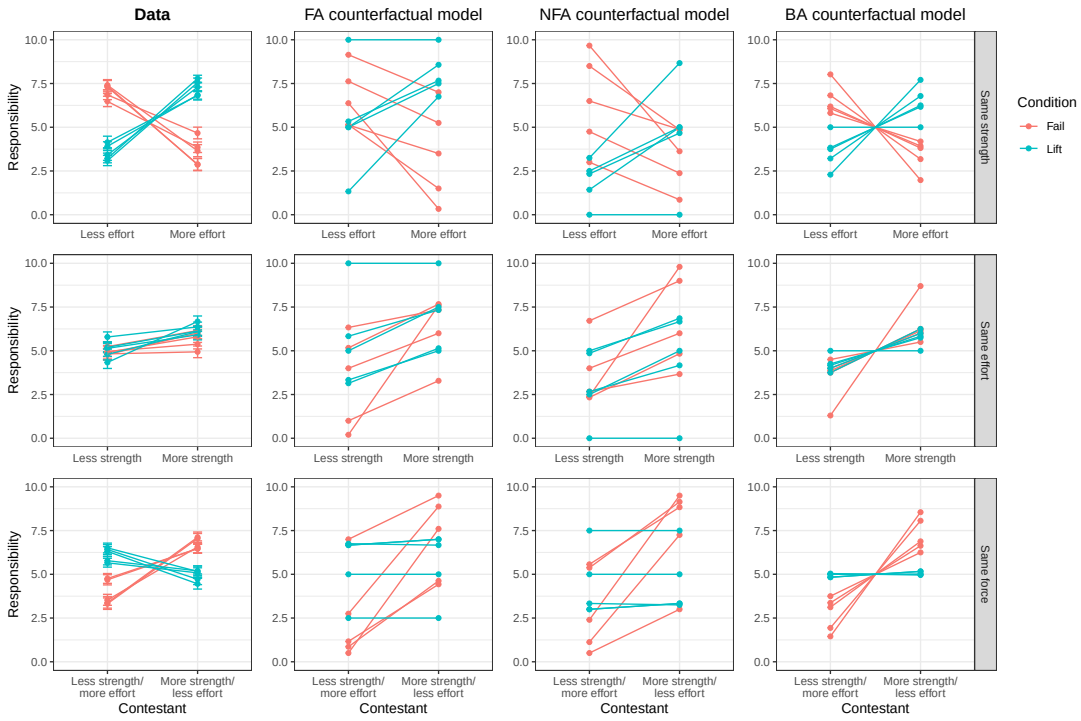


Figure B.11: Data and predictions of the Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models in Experiment 2a. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

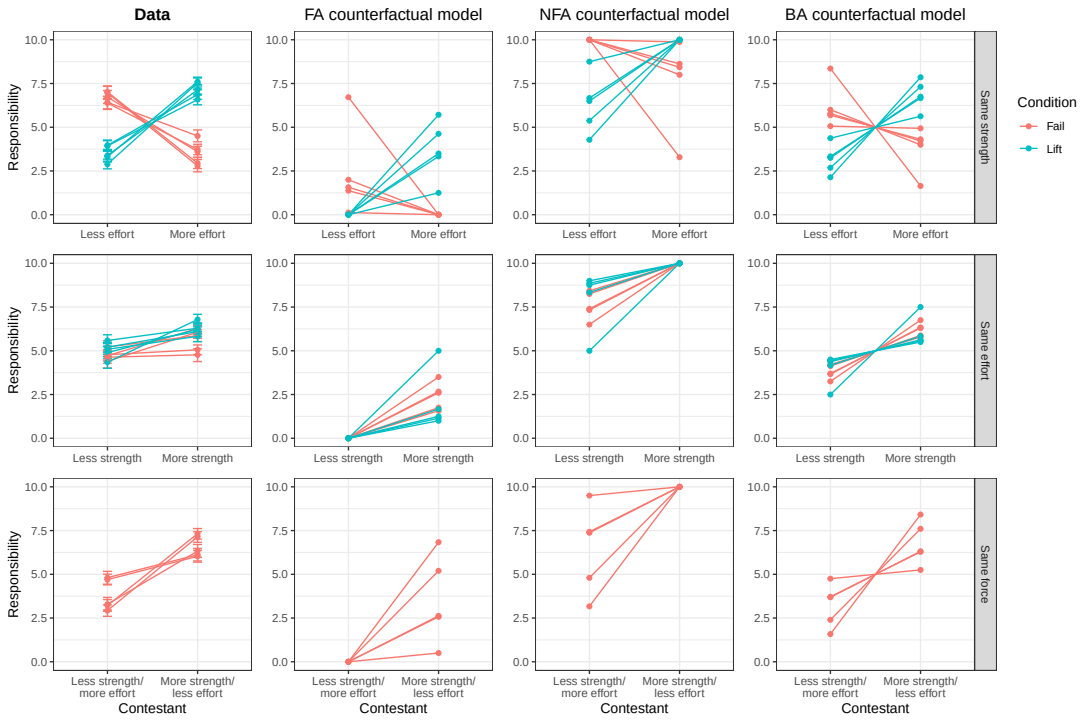


Figure B.12: Data and predictions of the Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models in Experiment 2b. Each line corresponds to a scenario. Error bars in the Data column indicate 95% bootstrapped confidence intervals.

B.7 Comprehension check questions used in the experiments

Correct answers are italicized.

1. What is the structure of the game show?

- It consists of multiple contests between differently-sized groups of contestants.
- *It consists of multiple contests between pairs of contestants.*
- The structure of the game show is unclear.

2. When do contestants change?

- *They change every contest.*
- After every successful lift.
- Contestants are the same across contests.

3. How do we know the weight of the box in each contest?

- The weight of the box will be the same during each contest (equivalent to a strength of 5).
- We will never know the weight of the box.
- *The weight of the box will be revealed to us during each contest.*

4. How much of the reward will each contestant get from a successful lift?

- The stronger contestant will receive a reward.
- *Each will receive a reward.*
- The contestant who exerted more effort will receive a reward.

5. Who would be able to lift a box with a weight of 5?

- *Anyone with a strength of 5 or higher, given that they exert enough effort.*
- Anyone, given that they exert enough effort.
- Anyone with a strength of 5, no matter how much effort they exert.

6. Suppose a contestant has a strength of 6. If they put in a 50% effort, how much force will they produce? *

- 6
- 5
- *3*

*This question appeared only in Experiments 2a and 2b.

B.8 Deviations from pre-registrations

B.8.1 Experiments 1a and 1b

In our pre-registration, we stated that we would convert the responsibility judgments and use the converted responsibility as the response variable. We decided that converting the predictor variables (effort and force) was a better approach than converting the response variables. In this way, we were able to use participants' original responses as the response variable.

We pre-registered a regression model regressing converted responsibility on strength, effort, and the interaction between strength and effort, with random effects of each predictor variable grouped by subject. We did not run this regression because it turned out to be irrelevant to our theory.

B.8.2 Experiment 3

We pre-registered a responsibility attribution task and a bonus allocation task, and in this paper we are reporting the results of the responsibility attribution task.

Our data set included 99 participants, instead of the 100 that we aimed for. We followed the stopping rule for data collection which we pre-registered ("we will stop data collection once 100 participants have submitted their responses on MTurk"). 100 participants submitted their task on MTurk, but we only received data from 99 participants.

In our pre-registration, we stated that we would convert the responsibility judgments and use the converted responsibility as the response variable. We decided that converting the predictor variables (effort and force) was a better approach than converting the response variables. In this way, we were able to use participants' original responses as the response variable.

We pre-registered a regression model regressing converted responsibility on strength, effort, and the interaction between strength and effort, with random effects of each predictor variable grouped

by subject. We did not run this regression because it turned out to be irrelevant to our theory.

As stated in the main text, we excluded 3 scenarios due to errors in design (the box weight was greater than 10 or a contestant's strength was greater than 10, which were impossible scenarios given our setup).

References

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

[2] Allen, K. R., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., and Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

[3] Almaatouq, A., Alsobay, M., Yin, M., and Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences*, 118(36):e2101062118.

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

[5] Atkinson, A. B. et al. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263.

[6] Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

[7] Baer, C. and Odic, D. (2022). Mini managers: Children strategically divide cognitive labor among collaborators, but with a self-serving bias. *Child Development*, 93(2):437–450.

[8] Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally interacting minds. *Science*, 329(5995):1081–1085.

[9] Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10.

[10] Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3):329–349.

[11] Bartel, A. P. (1994). Productivity gains from the implementation of employee training programs. *Industrial Relations: A Journal of Economy and Society*, 33(4):411–425.

[12] Bass, I., Espinoza, C., Bonawitz, E., and Ullman, T. D. (2024). Teaching without thinking: Negative evaluations of rote pedagogy. *Cognitive Science*, 48(6):e13470.

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

[14] Bentham, J. (1970). An introduction to the principles of morals and legislation (1789), ed. by j. *H Burns and HLA Hart, London*.

[15] Bernstein, E., Shore, J., and Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, 115(35):8734–8739.

[16] Bigelow, E., Xiang, Y., Gerstenberg, T., Ullman, T. D., and Gershman, S. J. (2025). Actual or counterfactual? asymmetric responsibility attributions in language models.

[17] Bigman, Y. E. and 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.

[18] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[19] Blake, P., McAuliffe, K., Corbit, J., Callaghan, T., Barry, O., Bowie, A., Kleutsch, L., Kramer, K., Ross, E., Vongsachang, H., et al. (2015). The ontogeny of fairness in seven societies. *Nature*, 528(7581):258–261.

[20] Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 91(1):166–193.

[21] Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330.

[22] Bowles, S. and Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, 65(1):17–28.

[23] Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, 101(2):327–341.

[24] Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477.

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

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

[27] Brownell, C. A., Ramani, G. B., and Zerwas, S. (2006). Becoming a social partner with peers: Cooperation and social understanding in one-and two-year-olds. *Child Development*, 77(4):803–821.

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

[29] Butterfill, S. (2012). Joint action and development. *The Philosophical Quarterly*, 62(246):23–47.

[30] Caldwell, C. A., Renner, E., and Atkinson, M. (2018). Human teaching and cumulative cultural evolution. *Review of Philosophy and Psychology*, 9(4):751–770.

[31] Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

[32] Campano, F. and Salvatore, D. (2006). *Income Distribution: Includes CD*. Oxford University Press.

[33] Candidi, M., Curioni, A., Donnarumma, F., Sacheli, L. M., and Pezzulo, G. (2015). Interactional leader–follower sensorimotor communication strategies during repetitive joint actions. *Journal of the Royal Society Interface*, 12(110).

[34] Caro, T. M. and Hauser, M. D. (1992). Is there teaching in nonhuman animals? *The Quarterly Review of Biology*, 67(2):151–174.

[35] Celniker, J. B., Gregory, A., Koo, H. J., Piff, P. K., Ditto, P. H., and Shariff, A. F. (2023). The moralization of effort. *Journal of Experimental Psychology: General*, 152(1):60–79.

[36] Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

[37] Charpentier, C. J., Iigaya, K., and O'Doherty, J. P. (2020). A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, 106(4):687–699.

[38] Chen, A. M., Palacci, A., Vélez, N., Hawkins, R. D., and Gershman, S. J. (2024). A hierarchical bayesian model of adaptive teaching. *Cognitive Science*, 48(7):e13477.

[39] Cheng, S., Zhao, M., Zhu, J., Zhou, J., Shen, M., and Gao, T. (2022). Intentional commitment through an internalized theory of mind: Acting in the eyes of an imagined observer. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

[40] Chockler, H. and Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115.

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

[42] collaboration, A. (2022). A detailed map of higgs boson interactions by the atlas experiment ten years after the discovery. *Nature*, 607(7917):52–59.

[43] Colman, A. M. (2024). Team reasoning cannot be viewed as a payoff transformation. *Economics & Philosophy*, 40(1):1–11.

[44] Colman, A. M. and Gold, N. (2018). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review*, 25(5):1770–1783.

[45] Cooke, N. J. (2015). Team cognition as interaction. *Current Directions in Psychological Science*, 24(6):415–419.

[46] Cooke, N. J., Gorman, J. C., Myers, C. W., and Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, 37(2):255–285.

[47] Cortes Barragan, R. and Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence. *Proceedings of the National Academy of Sciences*, 111(48):17071–17074.

[48] Curioni, A. (2022). What makes us act together? on the cognitive models supporting humans' decisions for joint action. *Frontiers in Integrative Neuroscience*, 16:900527.

[49] Curioni, A., Voinov, P., Allritz, M., Wolf, T., Call, J., and Knoblich, G. (2022). Human adults prefer to cooperate even when it is costly. *Proceedings of the Royal Society B: Biological Sciences*, 289(1973).

[50] Daw, N. D. and Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369:20130478.

[51] Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9(1):519–540.

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

[53] Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science*, 14(2):175–180.

[54] Dowe, P. (2000). *Physical Causation*. Cambridge University Press.

[55] Edmondson, A. C. (2012). *Teaming: How organizations learn, innovate, and compete in the knowledge economy*. John Wiley & Sons.

[56] Elnaga, A. and Imran, A. (2013). The effect of training on employee performance. *European Journal of Business and Management*, 5(4):137–147.

[57] Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87.

[58] Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

[59] Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

[60] Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., and Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, 52(2):203–224.

[61] Gabora, L. (2013). Cultural evolution as distributed computation. In *Handbook of Human Computation*, pages 447–461. Springer.

[62] Gallucci, M. and Perugini, M. (2000). An experimental test of a game-theoretical model of reciprocity. *Journal of Behavioral Decision Making*, 13(4):367–389.

[63] Gergely, G. and Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–292.

[64] Gershman, S. J. and Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review*, 30(4):1273–1293.

[65] Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349:273–278.

[66] Gerson, S. A. and Woodward, A. L. (2014). Learning from their own actions: The unique effect of producing actions on infants' action understanding. *Child Development*, 85(1):264–277.

[67] Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, 377(1866):20210339.

[68] Gerstenberg, T., Ejova, A., and Lagnado, D. A. (2011). Blame the skilled. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

[69] Gerstenberg, T., Goodman, N., and Lagnado, D. (2012). Noisy newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

[70] Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

[71] Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128:936–975.

[72] Gerstenberg, T. and Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, 115(1):166–171.

[73] Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, 28(12):1731–1744.

[74] Gerstenberg, T. and Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, 216:104842.

[75] Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., and Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, 177:122–141.

[76] Gilbert, M. (1987). Modelling collective belief. *Synthese*, 73(1):185–204.

[77] Gilbert, M. (1990). Walking together: A paradigmatic social phenomenon. *Mid West Studies in Philosophy*, 15(1):1–14.

[78] Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies*, 144(1):167–187.

[79] Girotto, V., Legrenzi, P., and Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78(1-3):111–133.

[80] Goldstone, R. L., Andrade-Lotero, E. J., Hawkins, R. D., and Roberts, M. E. (2024). The emergence of specialized roles within groups. *Topics in Cognitive Science*, 16(2):257–281.

[81] Golman, R., Bhatia, S., and Kane, P. B. (2020). The dual accumulator model of strategic deliberation and decision making. *Psychological Review*, 127:477–504.

[82] Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

[83] Goodman, N. D. and Stuhlmüller, A. (2014). The design and implementation of probabilistic programming languages.

[84] Greene, J. and Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1451):1775.

[85] Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3):364–371.

[86] Grosz, B. J. and Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

[87] Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388.

[88] Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, 25(10):896–910.

[89] Hackel, L. M., Doll, B. B., and Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, 18(9):1233–1235.

[90] Hackman, J. R. (2011). *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers.

[91] Hall, N. (2004). Two concepts of causation. *Causation and Counterfactuals*, pages 225–276.

[92] Ham, H., Zhao, B., Griffiths, T. L., and Vélez, N. (2025). Teaching recombinable motifs through simple examples. *Cognitive Science*, 49(8):e70103.

[93] Hamann, K., Bender, J., and Tomasello, M. (2014). Meritocratic sharing is based on collaboration in 3-year-olds. *Developmental Psychology*, 50(1):121.

[94] Hamann, K., Warneken, F., Greenberg, J. R., and Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, 476(7360):328–331.

[95] Hawkins, R. D., Berdahl, A. M., Pentland, A. S., Tenenbaum, J. B., Goodman, N. D., and Krafft, P. (2023). Flexible social inference facilitates targeted social learning when rewards are not observable. *Nature Human Behaviour*, 7(10):1767–1776.

[96] Heider, F. (1958). *The psychology of interpersonal relations.* John Wiley & Sons Inc.

[97] Henderson, A. M. and Woodward, A. L. (2011). "let's work together": What do infants understand about collaborative goals? *Cognition*, 121(1):12–21.

[98] 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.

[99] Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

[100] Henrich, J. and Boyd, R. (2016). How evolved psychological mechanisms empower cultural group selection. *Behavioral and Brain Sciences*, 39.

[101] Henrich, J. and Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, 72:207–240.

[102] Hertwig, R., Davis, J. N., and Sulloway, F. J. (2002). Parental investment: how an equity motive can produce inequality. *Psychological Bulletin*, 128(5):728.

[103] Hertz, U., Köster, R., Janssen, M. A., and Leibo, J. Z. (2025). Beyond the matrix: Experimental approaches to studying cognitive agents in social-ecological systems. *Cognition*, 254:105993.

[104] Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, 39(4):632–657.

[105] Ho, M. K., Cushman, F., Littman, M. L., and Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*.

[106] Ho, M. K., Saxe, R., and Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971.

[107] Hutchins, E. (1991). The social organization of distributed cognition. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on Socially Shared Cognition*, pages 283–307. American Psychological Association.

[108] Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

[109] Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., and Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, 112(10):3098–3103.

[110] Ibbotson, P., Hauert, C., and Walker, R. (2019). Effort perception is made more accurate with more effort and when cooperating with slackers. *Scientific Reports*, 9(1):17491.

[111] Icard, T. F., Kominsky, J. F., and Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161:80–93.

[112] Jara-Ettinger, J. and Dunham, Y. (2024). The institutional stance. *Behavioral and Brain Sciences*, pages 1–62.

[113] Jara-Ettinger, J. and Gweon, H. (2017). Minimal covariation data support future one-shot inferences about unobservable properties of novel agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39.

[114] Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20:589–604.

[115] Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., and Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, 140:14–23.

[116] Jara-Ettinger, J., Kim, N., Muetener, P., and Schulz, L. (2014). Running to do evil: Costs incurred by perpetrators affect moral judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

[117] Joseph, K. and Kalwani, M. U. (1998). The role of bonus pay in salesforce compensation plans. *Industrial Marketing Management*, 27(2):147–159.

[118] Kahneman, D. and Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2):136–153.

[119] Kanngiesser, P. and Warneken, F. (2012). Young children consider merit when sharing resources with others. *PLOS ONE*, 7.

[120] Kao, J. T., Wu, J. Y., Bergen, L., and Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

[121] Karau, S. J. and Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4):681.

[122] Keil, F. C., Stein, C., Webb, L., Billings, V. D., and Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, 32(2):259–300.

[123] Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J.-C., and Rao, R. P. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, 5(11):eaax8783.

[124] Kishore, S., Rao, R. S., Narasimhan, O., and John, G. (2013). Bonuses versus commissions: A field study. *Journal of Marketing Research*, 50(3):317–333.

[125] Kleiman-Weiner, M., Sosa, F., Thompson, B., van Opheusden, S., Griffiths, T., Gershman, S., and Cushman, F. (2020). Downloading culture. zip: Social learning by program induction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

[126] Knobe, J. and Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7(1):67–83.

[127] Knoblich, G., Butterfill, S., and Sebanz, N. (2011). Psychological research on joint action: theory and data. *Psychology of Learning and Motivation*, 54:59–101.

[128] Koenig, M. A., Clément, F., and Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, 15(10):694–698.

[129] Kozlowski, S. W. and Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7(3):77–124.

[130] Lacey, N. (1988). State punishment: political principles and community values.

[131] Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., and Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, 129:101412.

[132] Lanzetta, J. T. and Hannah, T. (1969). Reinforcing behavior of" naive" trainers. *Journal of Personality and Social Psychology*, 11(3):245.

[133] Larson Jr, J. R. (2013). *In search of synergy in small group performance*. Psychology Press.

[134] Latané, B., Williams, K., and Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37(6):822.

[135] Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

[136] Leonard, J. A., Bennett-Pierre, G., and Gweon, H. (2019). Who is better? preschoolers infer relative competence based on efficiency of process and quality of outcome. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.

[137] Leonard, J. A., Garcia, A., and Schulz, L. E. (2020). How adults' actions, outcomes, and testimony affect preschoolers' persistence. *Child Development*, 91(4):1254–1271.

[138] Leonard, J. A., Lee, Y., and Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, 357(6357):1290–1294.

[139] Leventhal, G. S., Weiss, T., and Long, G. (1969). Equity, reciprocity, and reallocating rewards in the dyad. *Journal of Personality and Social Psychology*, 13(4):300.

[140] Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, 97(4):182–197.

[141] Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

[142] Lin, H., Westbrook, A., Fan, F., and Inzlicht, M. (2021). An experimental manipulation increases the value of effort. *Nature Human Behaviour*.

[143] Liu, S., Cushman, F., Gershman, S., Kool, W., and Spelke, E. S. (2019). Hard choices: Children's understanding of the cost of action selection. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.

[144] Liu, S., Karakose-Akbiyik, S., Outa, J., and Kim, M. J. (2026). How physical information is used to make sense of the psychological world. *Nature Reviews Psychology*, 5(1):59–73.

[145] Liu, S., Ullman, T. D., Tenenbaum, J. B., and Spelke, E. S. (2017a). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041.

[146] Liu, S., Ullman, T. D., Tenenbaum, J. B., and Spelke, E. S. (2017b). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041.

[147] Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4):303–332.

[148] Lopes, L. L. (1994). Psychology and economics: Perspectives on risk, cooperation, and the marketplace. *Annual Review of Psychology*, 45(1):197–227.

[149] Lucas, A. J., Kings, M., Whittle, D., Davey, E., Happé, F., Caldwell, C. A., and Thornton, A. (2020). The value of teaching increases with tool complexity in cumulative cultural evolution. *Proceedings of the Royal Society B*, 287(1939):20201885.

[150] Lucas, C. G. and Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, 122(4):700.

[151] Lucca, K., Horton, R., and Sommerville, J. A. (2020). Infants rationally decide when and how to deploy effort. *Nature Human Behaviour*, 4(4):372–379.

[152] Lutz, D. J. and Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, 73(4):1073–1084.

[153] Magid, R. W., DePascale, M., and Schulz, L. E. (2018a). Four-and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, 2(2):72–85.

[154] Magid, R. W., DePascale, M., and Schulz, L. E. (2018b). Four-and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, 2(2):72–85.

[155] Mascaro, O. and Csibra, G. (2022). Infants expect agents to minimize the collective cost of collaborative actions. *Scientific Reports*, 12(1):17088.

[156] McCarthy, W. P., Hawkins, R. D., Wang, H., Holdaway, C., and Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *arXiv preprint arXiv:2107.00077*.

[157] Messick, D. M. (1993). *Equality as a decision heuristic.* Cambridge University Press.

[158] Michael, J., Sebanz, N., and Knoblich, G. (2016). Observing joint action: Coordination creates commitment. *Cognition*, 157:106–113.

[159] Mieczkowski, E., Mon-Williams, R., Bramley, N., Lucas, C. G., Velez, N., and Griffiths, T. L. (2025a). Predicting multi-agent specialization via task parallelizability. *arXiv preprint arXiv:2503.15703*.

[160] Mieczkowski, E., Turner, C., Vélez, N., and Griffiths, T. L. (2025b). People evaluate idle collaborators based on their impact on task efficiency. *Cognition*, 264:106200.

[161] Milkovich, G. T. and Wigdor, A. K. (1991). *Pay for performance: Evaluating performance appraisal and merit pay*. National Academy Press.

[162] Mill, J. S. (2016). Utilitarianism. In *Seven Masterpieces of Philosophy*, pages 329–375. Routledge.

[163] Miller, C. E. and Komorita, S. S. (1995). Reward allocation in task-performing groups. *Journal of Personality and Social Psychology*, 69(1):80.

[164] Miton, H. and Jackson, J. C. (2025). Complex technology requires cultural innovations for distributing cognition. *Trends in Cognitive Sciences*.

[165] Morton, R. W., Colenso-Semple, L., and Phillips, S. M. (2019). Training for strength and hypertrophy: an evidence-based approach. *Current Opinion in Physiology*, 10:90–95.

[166] Nagel, J. and Waldman, M. (2012). Force dynamics as a basis for moral intuitions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

[167] Nash Jr, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

[168] Nasvytis, L., Gershman, S. J., and Cushman, F. (2025). Belief neglect: A heuristic in social reasoning. *PsyArxiv preprint*.

[169] Nda, M. M. and Fard, R. Y. (2013). The impact of employee training and development on employee productivity. *Global Journal of Commerce and Management Perspective*, 2(6):91–93.

[170] Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

[171] Oliehoek, F. A. and Amato, C. (2016). *A Concise Introduction to Decentralized POMDPs*. Springer.

[172] O'mahony, S. and Ferraro, F. (2007). The emergence of governance in an open source community. *Academy of Management Journal*, 50(5):1079–1106.

[173] Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

[174] Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

[175] Payne, J. W., Bettman, J. R., and Luce, M. F. (1996). When time is money: Decision behavior under opportunity-cost time pressure. *Organizational Behavior and Human Decision Processes*, 66(2):131–152.

[176] Pesquita, A., Whitwell, R. L., and Enns, J. T. (2018). Predictive joint-action model: A hierarchical predictive approach to human cooperation. *Psychonomic Bulletin & Review*, 25(5):1751–1769.

[177] Pollack, M. E. (1990). Plans as complex mental attitudes. *Intentions in Communication*, 77(104):277–282.

[178] Powell, M. J. et al. (2009). The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26:26–46.

[179] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

[180] Quillien, T. and Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.

[181] Raihani, N. J. and Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society Open Science*, 3(11):160510.

[182] Ramos, A. (2022). *A Comparison of Fixed Pay, Piece-Rate Pay, and Bonus Pay when Performers Receive Tiered Goals*. Western Michigan University.

[183] Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

[184] Regidor, E. (2004). Measures of health inequalities: part 1. *Journal of Epidemiology and Community Health*, 58(10):858.

[185] Rest, S., Nierenberg, R., Weiner, B., and Heckhausen, H. (1973). Further evidence concerning the effects of perceptions of effort and ability on achievement evaluation. *Journal of Personality and Social Psychology*.

[186] Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, 84:971–985.

[187] Rollwage, M., Pannach, F., Stinson, C., Toelch, U., Kagan, I., and Pooresmaeili, A. (2020). Judgments of effort exerted by others are influenced by received rewards. *Scientific Reports*, 10(1):1868.

[188] Sadras, V. and Bongiovanni, R. (2004). Use of lorenz curves and gini coefficients to assess yield inequality within paddocks. *Field Crops Research*, 90(2-3):303–310.

[189] Sanna, L. J. and 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.

[190] Sarin, A. and Cushman, F. (2024). One thought too few: An adaptive rationale for punishing negligence. *Psychological Review*, 131(3):812.

[191] Schäfer, M., Haun, D. B., and Tomasello, M. (2023). Children's consideration of collaboration and merit when making sharing decisions in private. *Journal of Experimental Child Psychology*, 228:105609.

[192] Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, 114(3):327–358.

[193] Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

[194] Searle, J. R. (1990). Collective intentions and actions john r. searle. *Intentions in Communication*, 401.

[195] Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76.

[196] Sebanz, N. and Knoblich, G. (2021). Progress in joint-action research. *Current Directions in Psychological Science*, 30(2):138–143.

[197] Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition*, 88(3):B11–B21.

[198] Sen, A., Sen, M. A., Foster, J. E., Amartya, S., Foster, J. E., et al. (1997). *On economic inequality*. Oxford University Press.

[199] Shafto, P., Eaves, B., Navarro, D. J., and Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, 15(3):436–447.

[200] Shafto, P., Goodman, N. D., and Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89.

[201] Shafto, P., Wang, J., and Wang, P. (2021). Cooperative communication as belief transport. *Trends in Cognitive Sciences*, 25(10):826–828.

[202] Shultz, T. R., Schleifer, M., and Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 13(3):238.

[203] Shultz, T. R., Wright, K., and Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, pages 177–184.

[204] Skerry, A. E., Carey, S. E., and Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proceedings of the National Academy of Sciences*, 110(46):18728–18733.

[205] Smith, A. (1937). *The wealth of nations [1776]*. New York: Modern Library.

[206] Smith, A. (2010). *The Wealth of Nations: An inquiry into the nature and causes of the Wealth of Nations*. Harriman House Limited.

[207] Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

[208] Sobel, D. M. and Kushnir, T. (2013). Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, 120(4):779.

[209] Sommerville, J. A., Woodward, A. L., and Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, 96(1):B1–B11.

[210] Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., and Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, 217:104890.

[211] Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge (acting with technology)*. MIT Press.

[212] Strohminger, N. and Jordan, M. R. (2022). Corporate insecthood. *Cognition*, 224:105068.

[213] Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, 10(1):69–89.

[214] Sumers, T. R., Ho, M. K., Hawkins, R. D., and Griffiths, T. L. (2023). Show or tell? exploring when (and why) teaching with language outperforms demonstration. *Cognition*, 232:105326.

[215] Tabibnia, G., Satpute, A. B., and Lieberman, M. D. (2008). The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, 19(4):339–347.

[216] Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12(1):49–100.

[217] Tang, N., Gong, S., Zhao, M., Gu, C., Zhou, J., Shen, M., and Gao, T. (2022). Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

[218] Tessler, M. H. and Goodman, N. D. (2019). The language of generalization. *Psychological Review*, 126(3):395.

[219] Thomas, K. A., DeScioli, P., Haque, O. S., and Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology*, 107(4):657.

[220] Thomas, V., Wang, Y., and Fan, X. (2001). *Measuring education inequality: Gini coefficients of education*, volume 2525. World Bank Publications.

[221] Tian, F., Gershman, S. J., and Xiang, Y. (2026). Training collaborators for effective division of labor. *PsyArxiv preprint*.

[222] Tomasello, M. and Carpenter, M. (2007). Shared intentionality. *Developmental Science*, 10(1):121–125.

[223] Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.

[224] Tomasello, M. and Hamann, K. (2012). The 37th sir frederick bartlett lecture: Collaboration in young children. *Quarterly Journal of Experimental Psychology*, 65(1):1–12.

[225] Török, G., Pomiechowska, B., Csibra, G., and Sebanz, N. (2019). Rationality in joint action: Maximizing coefficiency in coordination. *Psychological Science*, 30(6):930–941.

[226] Török, G., Stanciu, O., Sebanz, N., and Csibra, G. (2021). Computing joint action costs: co-actors minimize the aggregate individual costs in an action sequence. *Open Mind*, pages 1–13.

[227] Tuomela, R. (1992). Group beliefs. *Synthese*, 91(3):285–318.

[228] Uhlmann, E. L., Pizarro, D. A., and Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, 10(1):72–81.

[229] Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., and Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, 22.

[230] Van Herpen, M., Van Praag, M., and Cools, K. (2005). The effects of performance measurement and compensation on motivation: An empirical study. *De economist*, 153:303–329.

[231] VanderBorght, M. and Jaswal, V. K. (2009). Who knows best? preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development: An International Journal of Research and Practice*, 18(1):61–71.

[232] Vaughan, G. M., Tajfel, H., and Williams, J. (1981). Bias in reward allocation in an intergroup and an interpersonal context. *Social Psychology Quarterly*, pages 37–42.

[233] Vélez, N., Chen, A. M., Burke, T., Cushman, F. A., and Gershman, S. J. (2023a). Teachers recruit mentalizing regions to represent learners' beliefs. *Proceedings of the National Academy of Sciences*, 120(22):e2215015120.

[234] Vélez, N., Christian, B., Hardy, M., Thompson, B. D., and Griffiths, T. L. (2023b). How do humans overcome individual computational limitations by working together? *Cognitive Science*, 47(1):e13232.

[235] Vélez, N. and Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in Cognitive Science*, 11(2):299–315.

[236] Vélez, N. and Gweon, H. (2020). Preschoolers use minimal statistical information about social groups to infer the preferences and group membership of individuals. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

[237] Vélez, N. and Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, 38:110–115.

[238] Vélez, N., Wu, C. M., Gershman, S. J., and Schulz, E. (2024). The rise and fall of technological development in virtual communities.

[239] Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010a). A minimal architecture for joint action. *Neural Networks*, 23(8-9):998–1003.

[240] Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010b). A minimal architecture for joint action. *Neural Networks*, 23(8-9):998–1003.

[241] Vizmathy, L., Begus, K., Knoblich, G., Gergely, G., and Curioni, A. (2024). Better together: 14-month-old infants expect agents to cooperate. *Open Mind*, 8:1–16.

[242] Von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

[243] Waldmann, M. R. and 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.

[244] Wang, P., Wang, J., Paranamana, P., and Shafto, P. (2020). A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33:17582–17593.

[245] Warneken, F., Steinwender, J., Hamann, K., and Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. *Cognitive Development*, 31:48–58.

[246] Warneken, F. and Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765):1301–1303.

[247] Waytz, A. and Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science*, 23(1):77–85.

[248] Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of Group Behavior*, pages 185–208. Springer.

[249] Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of Educational Research*, 42(2):203–215.

[250] Weiner, B. (1993). On sin versus sickness: A theory of perceived responsibility and social motivation. *American Psychologist*, 48(9):957–965.

[251] Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press.

[252] Weiner, B. and Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, 15(1):1.

[253] Williams, K. D. and Karau, S. J. (1991). Social loafing and social compensation: The effects of expectations of co-worker performance. *Journal of Personality and Social Psychology*, 61(4):570.

[254] Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136(1):82–111.

[255] Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34.

[256] Woodward, J. (2011). Mechanisms revisited. *Synthese*, 183:409–427.

[257] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688.

[258] Wu, C. M., Vélez, N., Cushman, F. A., Dezza, I. C., Schulz, E., and Wu, C. M. (2022). Representational exchange in human social learning. *The Drive for Knowledge*, 169.

[259] Wu, S. A. and Gerstenberg, T. (2024). If not me, then who? responsibility and replacement. *Cognition*, 242:105646.

[260] Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., and Kleiman-Weiner, M. (2021a). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432.

[261] Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., and Kleiman-Weiner, M. (2021b). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432.

[262] Xiang, Y., Bigelow, E., Gerstenberg, T., Ullman, T. D., and Gershman, S. J. (2025a). Language models assign responsibility based on actual rather than counterfactual contributions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

[263] Xiang, Y. and Gershman, S. (2025). Modeling intrinsic motivation as reflective planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

[264] Xiang, Y., Gershman, S. J., and Gerstenberg, T. (2026). A signaling theory of self-handicapping. *Cognition*, 266:106288.

[265] Xiang, Y., Landy, J., Cushman, F., Vélez, N., and Gershman, S. J. (2025b). People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, 116:104699.

[266] Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., and Gershman, S. J. (2023a). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, 241:105609.

[267] Xiang, Y., Vélez, N., and 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.

[268] Xiang, Y., Vélez, N., and Gershman, S. J. (2024). Optimizing competence in the service of collaboration. *Cognitive Psychology*, 150:101653.

[269] Yin, J., Xu, H., Ding, X., Liang, J., Shui, R., and Shen, M. (2016). Social constraints from an observer's perspective: Coordinated actions make an agent's position more predictable. *Cognition*, 151:10–17.

[270] Yoon, E. J., Tessler, M. H., Goodman, N. D., and Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, 4:71–87.

[271] Zultan, R., Gerstenberg, T., and Lagnado, D. A. (2012). Finding fault: causality and counterfactuals in group attributions. *Cognition*, 125(3):429–440.