# If not me, then who?<br>Responsibility and replacement

Sarah A. Wu
Stanford University

Tobias Gerstenberg[*]
Stanford University

Abstract

How do people hold others responsible? Responsibility judgments are affected not only by what actually happened, but also by what could have happened if things had turned out differently. Here, we look at how replaceability – the ease with which a person could have been replaced by someone else – affects responsibility. We develop the counterfactual replacement model which runs simulations of alternative scenarios to determine the probability that the outcome would have been different if the person of interest had been replaced. The model predicts that a person is held more responsible when it would have been more difficult to replace them. To test the model's predictions, we design a paradigm that quantitatively varies replaceability by manipulating the number of replacements as well as the probability with which each replacement would have been available. Across three experiments featuring increasingly complex scenarios, we show that the model explains participants' responsibility judgments well in both social and physical settings, and better than alternative models that rely only on features of what actually happened.

*Keywords:* responsibility; counterfactuals; social cognition; mental simulation; causality.

[*] Corresponding author: Tobias Gerstenberg, Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, Email: gerstenberg@stanford.edu. All the data, study materials, and analysis code are available here: `https://github.com/cicl-stanford/responsibility_replacement`

## Introduction

In the heist drama *Ocean's 8* – an all-female spin-off of *Ocean's Eleven* – main characters Debbie and Lou are recruiting talents to join them in pulling off a massive robbery. Lou introduces possible candidates to Debbie, who is often skeptical at first. After meeting Nine Ball, a computer hacker, Lou insists that "she's one of the best hackers on the East Coast." While observing Constance, a pickpocket, Lou gives Debbie a different reassurance – that they have other choices too because "the turnover in pickpockets is huge". Ultimately, both Nine Ball and Constance manage to impress Debbie and join the team, which succeeds in pulling off the heist. The movie ends with everyone splitting the loot evenly and silently parting ways.

All eight characters played a unique and essential role in the mission, put in their best effort, and accomplished what was asked of them. Although they all received an equal share of the reward, one might wonder whether they were equally responsible for the success. Perhaps Nine Ball's contribution was more important for the success because she accomplished something that fewer people would have been able to do? If Constance had not been there, then Debbie and Lou could have easily found another pickpocket to replace her given the high turnover. On the other hand, if Nine Ball had not been there, they would have struggled to find another hacker with skills as remarkable as hers. For that reason, it could be argued that Nine Ball was more responsible for the success because her contribution was less easily *replaceable*.

In this paper, we explore what role replaceability plays in how people hold others responsible. We look at situations in which multiple causes contributed to an outcome, and develop a computational model that explains responsibility attributions by considering how the situation would have unfolded if a particular contribution needed to be replaced. The rest of the paper is organized as follows. We first review prior work on how people make responsibility attributions and reason about counterfactual replacement. Then, we describe our model and test its predictions in three experiments. We conclude by discussing the key contributions of our work as well as some limitations that need to be addressed in future research.

### Responsibility and contribution

Many factors influence how people hold others responsible. Some involve an agent's character or mental states, such as their beliefs and intentions. For example, we generally hold others more responsible when they intended for the outcome to happen (e.g. Alicke, 2000; Lagnado & Channon, 2008; Malle, Guglielmo, & Monroe, 2014; Shaver, 1985). Other factors pertain to the causal role the agent played in bringing about the outcome (Gerstenberg et al., 2018; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021). Agents are generally held more responsible if their action played a pivotal role in bringing about the outcome (Gerstenberg & Lagnado, 2010; Lagnado, Gerstenberg, & Zultan, 2013; Zultan, Gerstenberg, & Lagnado, 2012). Here, we look at how individuals are held responsible for group outcomes, focusing on the causal role that each person's contribution played.

When multiple causes affected an outcome, there are several ways to conceptualize what contribution each made. First, contributions can differ in *value* (e.g. Caruso, Epley,

& Bazerman, 2006). For example, one teammate may have scored more points than another and thereby may be viewed as more responsible for the winning game (all else being equal). Second, contributions can differ in how how much of a *difference* they made to the outcome (e.g. Chockler & Halpern, 2004; Lagnado et al., 2013). For example, intuitively a citizen's vote is more responsible for a politician's election success when the outcome is close as opposed to when it's a landslide win. This intuition cannot be explained in terms of differing value since each vote counts the same. Contributions can also differ in how easily they could have been *replaced*. In *Ocean's 8*, Nine Ball and Constance's contributions cannot easily be compared because they each had unique jobs. Furthermore, both agents were pivotal as the heist would not have succeeded without them. However, arguably the pickpocket Constance's contribution was more easily replaceable than that of the hacker Nine Ball. If replaceability affects responsibility judgments, then Nine Ball may be viewed as more responsible than Constance for the team's success. In the following sections, we will review prior work on responsibility attribution in groups falling under each of these conceptualizations of contribution: value, difference-making, and replaceability.

**Responsibility and value.** One way that people may allocate responsibility in groups is in proportion to the amount of some units put into achieving the outcome, such as points scored, time spent, or effort exerted (Gerstenberg & Lagnado, 2010, 2012; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018; Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021; Xiang, Vélez, & Gershman, 2022). This is especially intuitive for collaborative efforts like playing a team sport or writing a manuscript together. When people are asked to assess their own responsibility in such cases, they tend to overestimate their personal contributions and underestimate others', producing an egocentric bias or "over-claiming" effect (Caruso et al., 2006; Forsyth, Zyzniewski, & Giammanco, 2002; Schroeder, Caruso, & Epley, 2016). Encouraging people to consider the individual contributions of others increases the responsibility allocated to them (Halevy, Maoz, Vani, & Reit, 2022; Savitsky, Van Boven, Epley, & Wight, 2005). Conversely, when people see others "free-riding" on group benefits, they reduce their own contributions, partly because it violates the social norm that shared responsibility comes from shared contributions (Kerr & Bruun, 1983). These effects highlight the intuitive mapping between contributed value and proportioned responsibility.

**Responsibility and difference-making.** The notion of value falls short in situations where contributions are incommensurable or don't combine additively toward the group outcome. In some cases, multiple agents can all be fully responsible for the same outcome (Kaiserman, 2021; Lagnado et al., 2013). In other cases, multiple agents can contribute the same value, but still be held responsible to different degrees. For example, while different countries in the United Nations may have the same number of votes, their voting power may differ based on the voting coalitions they tend to form (Felsenthal & Machover, 2004).

Chockler and Halpern (2004) define responsibility in situations like this using the notion of *pivotality.* In their model, the closer a person's contribution was to making a difference to the outcome, the more responsible they are. Consider a committee of eleven members that voted 10-1 for some policy A and 6-5 for another policy B. Intuitively, each of the six members who voted for policy B is more responsible for the marginal win there, than each of the ten members that voted for the clear win of policy A (see also Langenhoff et al., 2021; Livengood, 2011). All committee members contributed the same value — a single vote — towards the total count for both policies. However, each of the six majority voters for policy B was pivotal because had any of them voted differently, the policy would not have won. In contrast, the ten majority voters for policy A are each further away from being pivotal in the sense that, for each of them, four other members would have needed to vote against policy A in order to create a situation in which that voter would have become pivotal. Prior work has shown that individuals whose actions were (closer to being) pivotal, are held more responsible for the outcome (Gerstenberg & Lagnado, 2010; Gerstenberg et al., 2018; Lagnado et al., 2013; Zultan et al., 2012).

The extent to which individuals are held responsible also depends on how critical their contributions were perceived to be for a positive group outcome (Gerstenberg, Lagnado, & Zultan, 2023; Lagnado et al., 2013; Langenhoff et al., 2021). While *pivotality* captures how close a person's contribution was to making a difference *after* the outcome has happened, *criticality* captures how important a person's contribution is *before* any actions have taken place. For instance, in a bystander situation, everyone is pivotal because any one person could have intervened to change the outcome, regardless of how many people were present. Yet, there is a diffusion of responsibility such that the more (equally pivotal) bystanders are present, the less responsible each one feels (Darley & Latané, 1968). This can be explained by the fact that the more bystanders are present, the less critical each person becomes for the outcome due to the disjunctive nature of the situation. So, when multiple causes contribute the same value to the outcome (e.g. one vote or a helping hand), the extent to which their vote was critical and pivotal affects how responsible they are perceived.

Some other accounts have linked responsibility judgments to people's beliefs about how much a given event changed the probability of the outcome happening (Brewer, 1977; Fincham & Jaspars, 1983; Gerstenberg & Lagnado, 2012; Parker, Paul, & Reinholtz, 2020; Spellman, 1997). Accordingly, responsibility increases with the perceived likelihood that the outcome would happen as a consequence of the action. For example, people may regard a deciding goal that was scored in the last minute as more responsible for the team's success than a goal that was scored early in the game (Henne, Kulesza, Perez, & Houcek, 2021).

What all these accounts have in common is that they link responsibility judgments to a consideration of how much of a difference the action made. Intuitively, however, it not only matters how much of a difference someone made to the outcome, but also whether it's conceivable that they could have acted differently in the first place. If it was impossible for a person to have taken a different action, then we should not hold them responsible even if the outcome had been different had they taken that (impossible) action (Kominsky & Phillips, 2019; Malle et al., 2014; Weiner, 1993; Wells & Gavanski, 1989). Petrocelli, Percy, Sherman, and Tormala (2011) capture this intuition in their *counterfactual potency* model which predicts that responsibility judgments are related to the product of two quantities: if-likelihood and then-likelihood. Consider the following counterfactual statement: "IF only Mr. Jones had been driving more slowly, THEN he wouldn't have hit the pedestrian." This counterfactual is potent to the extent that the if-likelihood is high (i.e. it is easy to imagine that Mr. Jones could have driven more slowly), and the then-likelihood is high (i.e. it is plausible that the pedestrian wouldn't have been hit in that case). The product of these two quantities determines how potent a counterfactual is, which then predicts responsibility according to the model. So, for example, if Mr. Jones consistently speeds while driving, then the if-likelihood would be low, rendering the counterfactual impotent. Similarly, if it was unlikely that Mr. Jones' driving more slowly would have prevented the pedestrian from being hit, then the then-likelihood would be low and potency low as well.

**Responsibility and replaceability.** Multiple agents contributed to the successful heist in *Ocean's 8*. While each agent's contribution was necessary to make it happen, some contributions intuitively mattered more than others and are thus deserving of more responsibility. Here, we propose a third way of thinking about what difference a contribution made to the outcome: namely, how easily it could have been replaced. The easier it would have been to replace someone's contribution, the less responsible that person is held for the outcome.

Prior studies on responsibility in groups have alluded to the notion of replacement. Responsibility is affected by how a person's contribution compares to expectations about how someone would or should have acted in that situation. Exceeding expectations results in more responsibility when it reveals something positive about the person's character (Gerstenberg et al., 2018; Langenhoff et al., 2021). Such expectations may come from prior knowledge about the person, from simulating what oneself would do (Simpson, Alicke, Gordon, & Rose, 2020), or from norms in different domains. For example, in the law, jurors are sometimes asked to evaluate the defendant against what a "reasonable person" would have done in the same situation (Schaffer, 2010; Tobia, 2018). In baseball, the Wins Above Replacement (WAR) metric measures a player's value in terms of how many wins they contribute to their team compared with a possible replacement-level player (Gerstenberg et al., 2018; Lagnado & Gerstenberg, 2015). All of these standards rely on a comparison between the agent who actually contributed to the outcome, and a hypothetical agent who could have replaced them in the same situation.

Expectations about individuals in groups may also be based on their roles. Different roles elicit different responsibility judgments even when the group members make essentially equivalent contributions (Awad et al., 2020; Forsyth et al., 2002; Sanders et al., 1996). One possible explanation is that different replacement standards are applicable for different roles. For instance, in situations where one agent made decisions and another implemented them, people hold the decider more responsible than the implementer, possibly because they view the implementer as more easily replaceable (Gantman, Sternisko, Gollwitzer, Oettingen, & Van Bavel, 2020). If the implementer had refused, then the decider could have recruited someone else to carry out their intent.

People also often use replacement logic to deny responsibility for immoral behavior by reasoning that 'if I don't do it, someone else will" (Falk, Neuber, & Szech, 2020; Falk & Szech, 2013), or to absolve themselves in common goods dilemmas along the lines of "it doesn't really matter what I do" (Glover & Scott-Taggart, 1975; Green, 1991; Kerr, 1996). The larger the group is, the more potential replacements there are, and generally the less responsible people feel. For example, in Falk and Szech's (2013) experiment, more participants were willing to kill a mouse for a fixed amount of money when the decision was made as the result of a market trade compared to an individual decision. In markets, traders can reason "if I don't buy or sell, someone else will" and thereby downplay personal responsibility for the negative consequences of the trade. The more traders are present in the market, the more likely someone else will buy or sell instead, and thus the less responsible each person feels for the consequences of their actions.

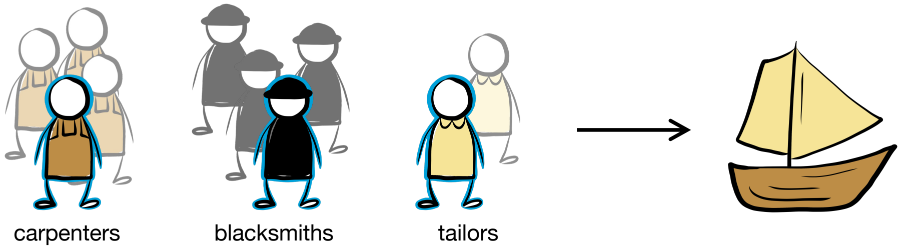

*Figure 1*. Illustration of an example trial in the *agent condition*. One craftsperson of each type (highlighted in blue) helped build the ship. Here, there were three other carpenters, three other blacksmiths, and one other tailor who could have been potential replacements. Between trials, we varied the number of possible replacements of each type.

## Overview of experimental paradigm

In this paper, we explore how replaceability affects responsibility attributions. We develop a paradigm that quantitatively manipulates information about replaceability. Imagine a fictional village with three types of craftspeople who build ships together: carpenters, blacksmiths, and tailors (see Figure 1). Each ship is made of wood, metal, and fabric, which requires the expertise of one craftsperson of each type. Any particular person might not be able to help, but as long as there is (at least) one craftsperson of each type willing to help, a ship will be successfully built. In this example, the village has four carpenters, four blacksmiths, and two tailors, and the ship was built. How responsible is each of the three helping craftspeople for the positive outcome?

Our model predicts that the easier it would have been to replace someone who contributed, the less responsible that person is judged. Accordingly, despite all three craftspeople making equal contributions, the carpenter and the blacksmith are less responsible than the tailor because there were more carpenters and blacksmiths that could have filled in those roles. In contrast, if it weren't for that particular tailor, then the village would have had to rely on the only other tailor who might not have been available either.

We test the predictions of our model in three experiments. In Experiment 1, we show that responsibility judgments are sensitive to the number of possible replacements. In Experiment 2, we show that responsibility judgments are also sensitive to how likely a possible replacement would have been available. In Experiment 3, we manipulate how likely the contributor would have needed to be replaced to begin with. In each experiment, we test people's responsibility judgments in social and physical contexts.

## Counterfactual Replacement Model (CRM)

The *Counterfactual Replacement Model* (CRM) assigns responsibility to individuals for group efforts. The CRM predicts that a person will be held more responsible for the outcome the lower the probability was that a successful counterfactual replacement could have been made. The model predicts that when a replacement would have almost inevitably made the outcome come about anyway, like in crowded markets (Falk & Szech, 2013), people hold a person less responsible.

There are many ways in which a candidate replacement can play out. For example, in

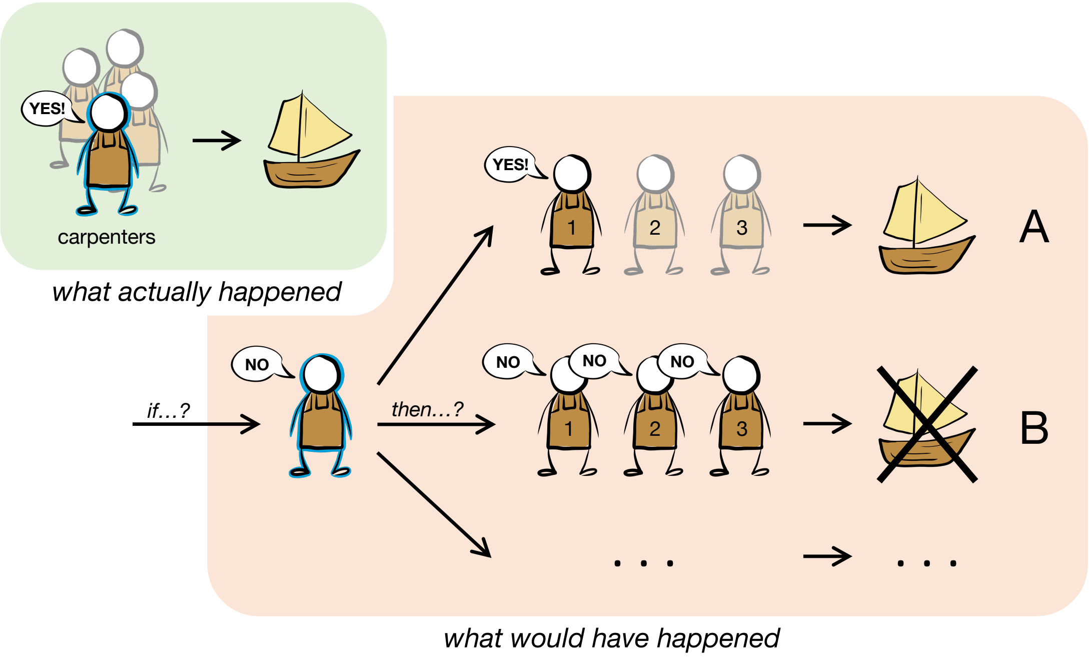

*Figure 2*. Schematic diagram of the model. The model predicts the responsibility attributed to the carpenter highlighted in blue for the successful ship by considering what would have happened if that carpenter had said “no”. Two possible counterfactual scenarios are shown. In scenario A, the first of the three other carpenters was available, so the ship would still have been built. In scenario B, none of the other carpenters were available as a replacement, so the ship would not have been built. The model computes responsibility by enumerating and computing the probability of a successful replacement.

a sports context, we may consider what would have happened if a player on the court had been replaced with a player from the bench. And in a legal context, we may consider what would have happened if a “reasonable person” had found themselves in the same situation as the defendant. In such instances, simulating what would have happened in the relevant counterfactual situation can be challenging.

Here, we develop the CRM for relatively simple settings like the one shown in Figure 1. For example, to determine the extent to which the carpenter is responsible, the CRM simulates what would have happened if the carpenter had refused to help build the ship. Figure 2 illustrates a schematic of that process. If the carpenter had said “no” then the other carpenters would have been asked one by one if they were able to help instead. If another carpenter had said “yes” to helping (scenario A), then the ship would still have been built. If all of the other carpenters had said “no”, then ship would not have been built (scenario B). By relying on a generative model of the situation, the CRM can explicitly enumerate and compute how likely each possible counterfactual scenario would have resulted in a success or a failure.

For all of the other $n$ carpenters in the village, let $p_i$ be the probability that carpenter $i$ says “yes” to helping. Scenario A, in which replacement $i = 1$ had said “yes”, would have happened with probability $p_1$ and resulted in a successful ship. Scenario B, in which all

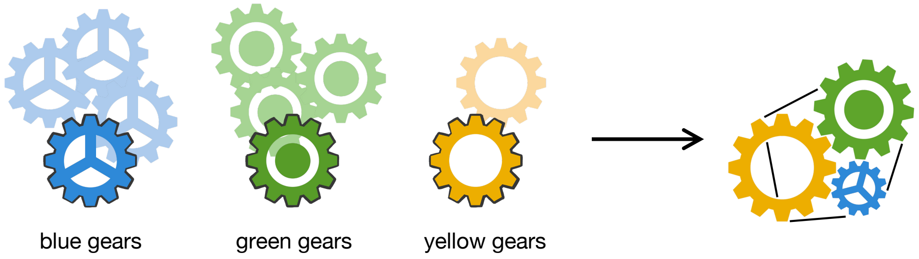

*Figure 3*. Illustration of an example trial in the *object condition.* One gear of each type (highlighted in black) helped form the machine. Here, there were three other blue gears, three other green gears, and one other yellow gear that could have been potential replacements. Between trials, we varied the number of possible replacement gears of each type.

three potential replacements had said "no", would have happened with probability $(1 - p_1) \times (1 - p_2) \times (1 - p_3)$ and resulted in a failed ship. The outcome would have failed if and only if none of the replacements had been available, as in scenario B. More generally, if we use $p_i$ to denote the probability that counterfactual replacement $i$ would have succeeded, then we can compute the probability of a successful counterfactual replacement as

$$\text{replaceability} = 1 - \prod_{i=1}^{n} 1 - p_i. \quad (1)$$

The CRM predicts that the higher the value of this term, the lower the responsibility attributed to the person that would have been replaced. Replaceability increases with increasing $n$ and increasing values of $p_i$. The more potential replacements there were (higher values of $n$) and the more likely those replacements were to say "yes" (higher values of $p_i$), the more likely a successful counterfactual replacement would have been made. Conversely, the easier it would have been to replace someone, the less responsible that person is for the group outcome. Importantly, responsibility judgments only depend on $n$ and $p_i$ for each individual, and not on any of the other contributors in the situation. For example, when considering the responsibility of the carpenter, we assume that the blacksmith and tailor still said "yes" and only imagine what would have happened if the particular carpenter had said "no". In the following experiments, we test the CRM by manipulating $n$ and $p_i$ and measuring responsibility judgments.

## Experiment 1: Number of replacements

Experiment 1 investigates what effect the number of possible replacements $n$ has on responsibility judgments. We had participants judge how responsible each craftsperson was for the ship in scenes such as Figure 1 and varied the number of possible replacements for each person while keeping the outcome the same. In line with Equation 1, we predicted that the more replacements there were for a person, the less responsibility would be attributed to that person.

We also tested whether the CRM applies to responsibility judgments about objects, or whether agents are treated differently from objects. Half the participants learned about the craftspeople building ships, and the other half were introduced to a parallel scenario involving three types of gears (blue, green, and yellow) forming a machine together (see Figure 3). Similar to the ships, each machine requires exactly one gear of each type to work properly. However, the gears can sometimes be broken, in which case other gears of the same type can be used instead. Participants in this condition saw scenes in which the machine was a success and were asked to judge how responsible each gear was for the success. Just like in the agent condition, we manipulated the number of possible replacement gears of each type.

## Methods

All materials including data, experiments, and analyses scripts are available at: `https://github.com/cicl-stanford/responsibility_replacement`. The experiment was programmed in jsPsych (de Leeuw, 2015) and pre-registered (agent condition: `https://osf.io/jnuay`; object condition: `https://osf.io/w2eh6`).[1]

**Participants.** The task was posted as an online study on Prolific, a crowd-sourcing research platform. 101 participants were recruited and compensated at a rate of $11/hour. One was excluded for failing an attention check (described in the next section), leaving a final sample size of $N = 100$ (*age*: M = 25, SD = 6; *gender*: 34 female, 63 male, 1 non-binary, 2 undisclosed; *race*: 64 White, 7 Black, 7 Asian, 3 Multiracial, 19 undisclosed). Participants were randomly assigned to the *agent* or *object* condition with $n = 50$ in each.

**Procedure & design.** Participants were first guided through instructions with two examples and then answered three comprehension questions to make sure they understood the setting. They were only able to proceed to the main task if they answered all three questions correctly, otherwise they were redirected to the beginning of the instructions. During the main task, they did two practice trials followed by 20 test trials in a randomized order.

In each trial, participants were shown the three contributors and the number of possible replacements for each one in a display similar to Figure 1. They were told that the outcome was successful and asked to judge how responsible they thought each craftsperson was for the ship, or how responsible each gear was for the machine, depending on the condition. Participants responded using three continuous sliders from “not at all” (0) to “very much” (100).

We emphasized that in every trial, the three contributors played an equal role in bringing about the successful outcome. Our only manipulation was the number of possible replacements for each one in each scene. We included all possible combinations of replacements ranging from zero to three (see Table A1 in the Appendix for details). For instance, Figure 1 shows a scene in which two contributors each have three possible replacements and the third contributor has one. We randomized the permutation of the three numbers across trials so that overall there were not more carpenters or yellow gears, for example. We also included a trial in which all three contributors had zero replacements, which was used as an attention check. Participants were excluded if their highest and lowest ratings differed by more than 30 on this trial. After the last trial, participants had the option to share demographic information and comments about the experiment. The average time to complete the experiment was 9.8 minutes (SD = 5.6).

[1] The experiments reported in this paper are part of a larger project that includes additional pre-registered studies.

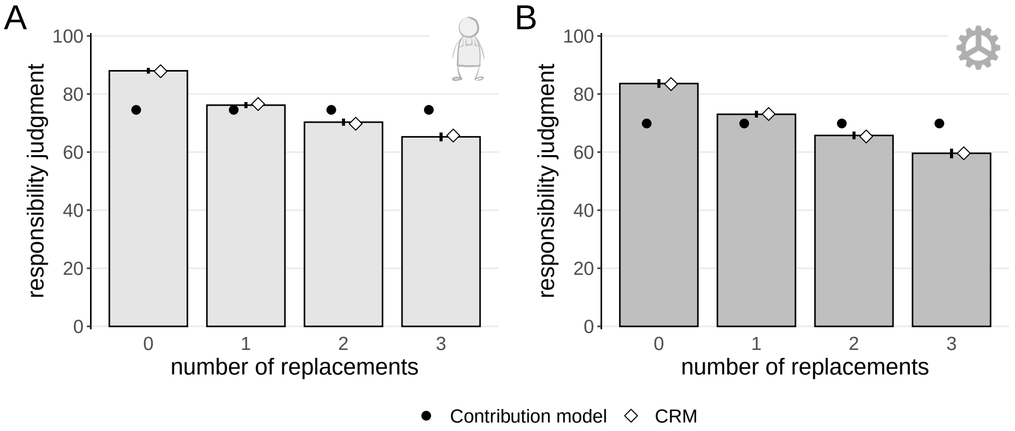

*Figure 4.* Mean responsibility judgments for the (A) agent and (B) object conditions as a function of the number of replacements in Experiment 1. The black and white symbols show model predictions. Error bars are bootstrapped 95% confidence intervals.

## Results

Figure 4 shows participants' mean responsibility judgments as a function of the number of possible replacements. The more replacements there were for a particular contribution, the less responsible people tended to hold it for the outcome, regardless of whether it was an agent or an object. We discuss the results from each condition in turn.

Table 1

*Experiment 1 model comparison. 'Intercept' and 'Replaceability' show the posterior means of each predictor along with 95% highest density intervals (HDIs). The contribution model only included an intercept as a predictor. r = Pearson correlation coefficient and RMSE = root mean squared error. "Δelpd" shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models, along with the associated standard error. Lower numbers indicate worse performance (Vehtari, Gelman, & Gabry, 2017). "n best" is the number of participants whose judgments were best predicted by each model. We adopt the convention of calling an effect* notable *if the 95% HDI of the estimated parameter in the Bayesian model excludes 0. The results show that replaceability is a notable predictor of participants' responsibility judgments in both conditions.*

| Model | Intercept | Replaceability | $r$ | RMSE | Δelpd (se) | $n$ best |
|---|---|---|---|---|---|---|
| *Agent condition* | | | | | | |
| CRM | 87.93 [83.83, 92.04] | −28.35 [−38.10, −18.76] | 0.99 | 1.40 | 0 (0) | 35 |
| Contribution | 74.32 [68.55, 80.03] | | | 8.24 | −1591.1 (74.2) | 15 |
| *Object condition* | | | | | | |
| CRM | 85.72 [75.71, 91.71] | −41.86 [−64.34, −19.90] | 0.96 | 2.39 | 0 (0) | 44 |
| Contribution | 69.85 [64.25, 75.74] | | | 8.96 | −1777.8 (80.2) | 6 |

**Agent condition.** The filled circles in Figure 4 show the predictions of the CRM. We fit two different Bayesian mixed effects models to participants' responsibility judgments. One is the CRM, which uses replaceability as a predictor. The other is a "contribution model", which only includes a fixed intercept. The contribution model predicts that each craftsperson should be held equally responsible because they each contributed the same to the outcome. Both models include random intercepts for each participant and the CRM also includes random slopes. All Bayesian models reported in this paper were written in Stan (Carpenter et al., 2017) and specified with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019).

To compute the probability of replacement in the CRM, we assumed a uniform probability $p$ that a potential replacement craftsperson would have helped, and found the value of $p$ that minimizes the squared error between model predictions and mean judgments (see Figure A1 for details). Then, given $n$ possible replacements, Equation 1 becomes

$$\text{replaceability} = 1 - (1 - p)^n. \tag{2}$$

Participants' judgments in the agent condition were well-captured by the CRM with a correlation of $r = 0.99$ and RMSE = 1.40, far better than the contribution model. The replaceability predictor was notable (see Table 1). To evaluate the CRM against the contribution model, we ran an approximate leave-one-out cross-validation comparison. We also fitted the models to individual participants and used the same cross-validation procedure to evaluate which model fit each participant's responses best. Table 1 shows that the CRM accounts best for the overall data and the majority of individual participants (35 out of 50).

**Object condition.** Mean responsibility judgments in the object condition were similar to those in the agent condition. They were well-captured by the CRM with a correlation of $r = 0.96$ and RMSE = 2.64, and the replaceability predictor was notable. Table 1 shows that, in this condition too, the CRM fares better on cross-validation on the overall data and best explains a majority of 44 out of 50 individual participants' judgments.

## Discussion

The results of Experiment 1 show that even when each person's contribution toward the outcome was the same, their responsibility differed. The more potential replacements a person had, the less responsible that person was judged to be. Prior work showed that the number of contributors and the way their contributions affect the outcome, influence responsibility judgments (e.g. Lagnado et al., 2013). Here we show that even when the number of actual contributors is held constant, and when each contributor affects the outcome in the same way, participants' still differentiate between them in their responsibility judgments. For responsibility, it not only matters what one did, but also how easily one's contribution could have been replaced by someone else.

Overall, we found that responsibility judgments were accurately predicted by replaceability. On the individual level, we found that a majority of participants' responses were best explained by the CRM, while only a minority was best explained by the contribution model. Participants' comments about what factors influenced their responses also revealed these individual differences. For some people, responsibility is only about the contribution itself (e.g. "They all had the same importance. The ship needs all three professions to be built therefore they all share an equal part in the ship building success, regardless of how many people were available."). This group was best fit by the contribution model, which predicts uniform judgments throughout. But for most people, it also matters how easily someone else could have stepped in to achieve the same outcome (e.g. "The more gears of the same color [the] village had, the less responsible the one of the same color was, because in case one fails there's another to replace it.").

While many participants explicitly mentioned replacement in their comments, it's possible that some of those best fit by the CRM used a different reasoning strategy instead. Because we fit a uniform probability that each replacement would have been available and only varied the number of replacements, it is difficult to tease apart the predictions of the CRM from a simpler model that explains responsibility judgments only in terms of group size. For example, a simple diffusion of responsibility model (Darley & Latané, 1968), which says that contributions decrease as the number of individuals involved increases, would predict the same negative relationship between responsibility and number of replacements without being sensitive to the causal structure of the situation.[2] To provide a more direct test of the CRM, we manipulated both the number and availability of replacements in Experiment 2.

## Experiment 2: Availability of replacements

In Experiment 1, we varied the number of replacements $n$ and found that it influenced responsibility judgments. Agents and objects were seen as less responsible for the outcome the more replacements they had. If participants are really reasoning about counterfactual replacements, however, then they should be sensitive not to the number of replacements per se, but rather to factors indicative of replaceability more generally. Replaceability increases with the number of replacements in the absence of any other information, but depends more directly on the probability that a replacement is available (see Equation 1). For instance, a carpenter with an unavailable replacement is less replaceable than one with a readily available replacement. In a counterfactual scenario, the unavailable replacement would have been less likely to actually step in to help build the ship. The CRM predicts that the carpenter with the unavailable replacement is more responsible. In this experiment, we test whether the availability of replacements influences responsibility judgments.

### Methods

The experiment was programmed in jsPsych (de Leeuw, 2015) and pre-registered (agent condition: `https://osf.io/j7vw6`; object condition: `https://osf.io/bdf95`).

**Participants.** The experiment was posted on Prolific. $N = 100$ participants (*age*: M = 25, SD = 6; *gender*: 58 male, 40 female, 2 non-binary; *race*: 58 White, 7 Black, 5 Asian, 3 Multiracial, 2 American Indian/Alaska Native, 25 undisclosed), excluding any from Experiment 1, were recruited and compensated at a rate of $11/hour. They were randomly assigned to the *agent* and *object* conditions with $n = 50$ in each.

[2] Note that in order for a diffusion of responsibility model to apply here, it would have to make the assumption that the potential replacements were *involved* in bringing about the outcome (Forsyth et al., 2002).

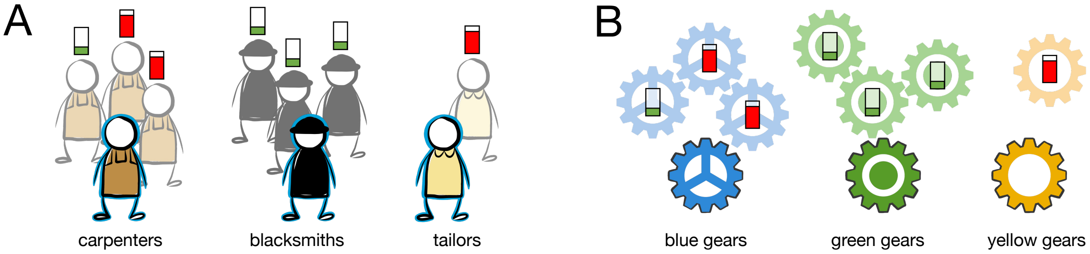

*Figure 5*. Example trials in Experiment 2 for the (A) agent and (B) object conditions. A fuller red bar indicates that the member is more likely to be busy or broken, and hence has low availability. An emptier green bar indicates that it is less likely to be busy or broken, and hence has high availability.

**Procedure & design.** The procedure and design were the same as that of Experiment 1, except that we additionally introduced the availability of each replacement. Figure 5 shows an example of what a trial looked like. In the agent condition, each craftsperson could be more or less busy, which indicated their probability of being available to help build the ship. In the object condition, each gear could be more or less brittle, which indicated its probability of being broken if used in the machine. In each trial, participants were shown the availability of all replacements in the scene, but not the three that actually contributed to the outcome.

We designed 15 possible sets of replacements where the number of replacements ranged from 0 to 4 and the availability of each one was either low or high (see Table B1). For example, one possible set (namely, set 5 in Figure 6) consists of two replacements, one with high availability and one with low availability. Each trial featured three different sets of replacements, one for each contributor in the scene. We designed 20 trials and ensured that each set of replacements appeared in at least two different trials. The sets were distributed among the trials so that there were no more than 12 total agents or objects (including the three contributors) in any one trial, in order to avoid overwhelming the display. For example, there was no trial in which the carpenter, blacksmith, and tailor all had four replacements each (as that would have been 15 total agents). Like in Experiment 1, we included a trial in which all three contributors have zero replacements, which was used as an attention check. Participants were excluded if their highest and lowest ratings differed by more than 30 on this trial. All participants passed the attention check in this experiment. Participants took an average of 12.3 minutes (SD = 6.5) to complete the experiment.

## Results

Figure 6 shows participants' mean responsibility judgments across all possible sets of replacements that we tested. They are sorted in order of increasing number and availability. The filled symbols show model predictions. For both conditions, we fit three different Bayesian mixed effects models to participants' responsibility judgments. One is the $\text{CRM}_{\text{uniform}}$, which assumes a uniform probability of success $p$ for any replacement, as in Experiment 1. This model computes the replaceability predictor using Equation 2. Another model is the $\text{CRM}_{\text{full}}$, which assumes two different probabilities $p_{\text{low}}$ and $p_{\text{high}}$ for

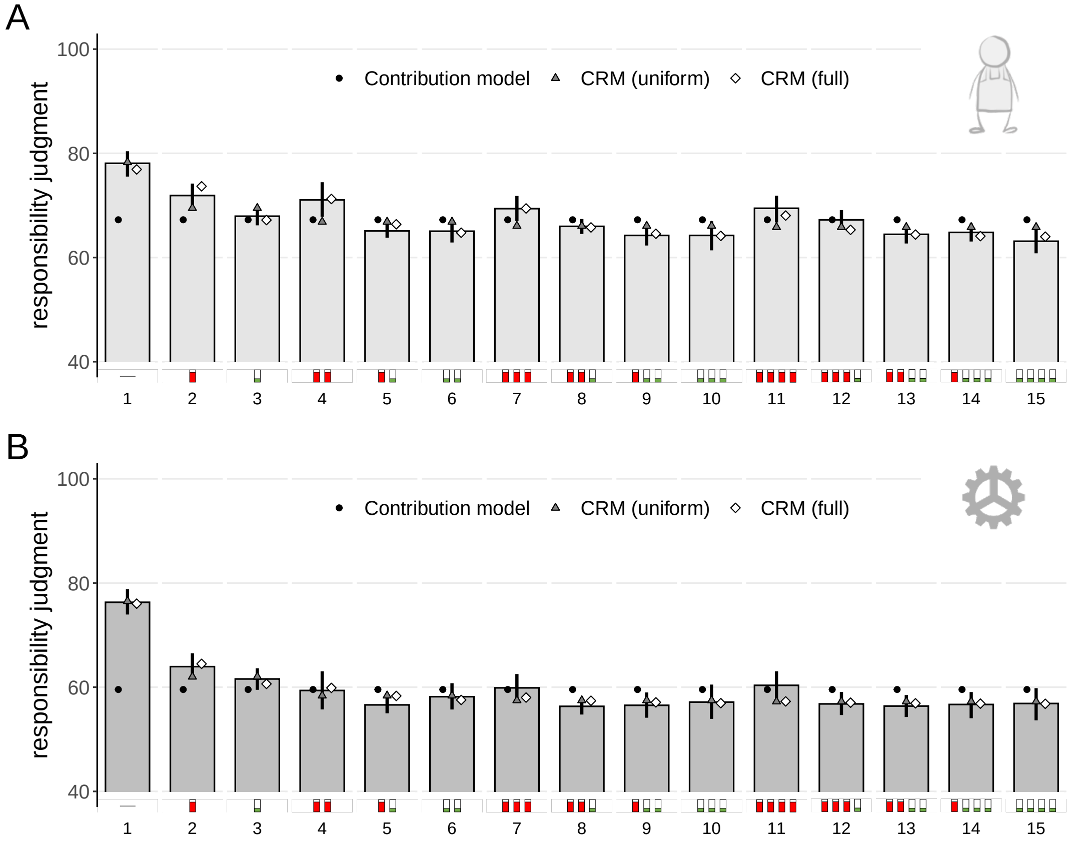

*Figure 6*. Mean responsibility judgments for the (A) agent and (B) object conditions for each set of replacements in Experiment 2. Each contributor had up to four possible replacements, each of which had either low or high availability. The sets are ordered by increasing number and availability of replacements. The different shaded symbols represent model predictions. Error bars are bootstrapped 95% confidence intervals. Note that the y-axis is truncated: participants judged responsibility on a scale from 0 to 100.

replacements with either low or high availability, respectively. The full model computes the replaceability predictor as

$$\text{replaceability} = 1 - (1 - p_{\text{low}})^{n_{\text{low}}}(1 - p_{\text{high}})^{n_{\text{high}}} \tag{3}$$

where $n_{\text{low}}$ is the number of replacements having low availability and $n_{\text{high}}$ is the number with high availability. The parameters $p$, $p_{\text{low}}$, and $p_{\text{high}}$ were fit to minimize the squared error between the respective model predictions and mean judgments in each condition. We ran a grid search over values between 0 and 1 with the only constraint being that $p_{\text{low}} < p_{\text{high}}$ (see Figures B1 and B2 for parameter search details). Both versions of the CRM included random slopes for each participant. Finally, we also fit the contribution model which only includes an intercept to capture the fact that each craftsperson or gear contributed the same amount to the outcome. All three models included random intercepts for each participant.

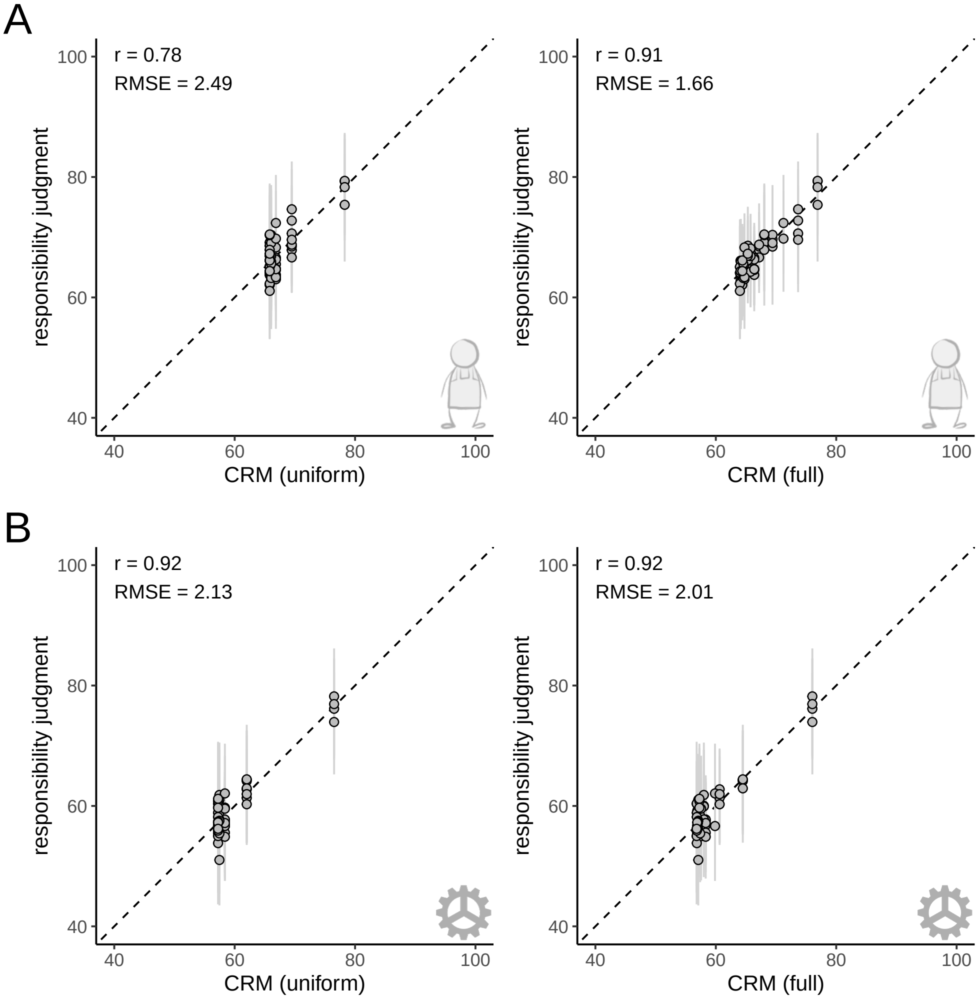

*Figure 7*. Scatter plots showing the relationship between mean responsibility judgments and the uniform and full version of the CRM in Experiment 2, in the (A) agent and (B) object conditions. $r$ = Pearson correlation coefficient, RMSE = root mean squared error. Error bars are bootstrapped 95% confidence intervals. Note that the axes are truncated: participants judged responsibility on a scale from 0 to 100.

The results in Figure 6 illustrate the relationship between responsibility and replaceability parameters $n$ and $p$ as predicted by the CRM. The more replacements there were for a particular contribution, the less responsible participants tended to hold it, thus replicating what we found in Experiment 1. However, for a fixed number of replacements, the less available they were individually, the more responsible participants rated that contribution. For example, participants judged a craftsperson with four replacements who all had high availability (mean responsibility 63.3, 95% CI [58.9, 67.3]) to be less responsible than a craftsperson whose four replacements all had low availability (69.5 [65.1, 77.38]). We discuss the results from each condition in turn.

**Agent condition.** Mean responsibility judgments in the agent condition were well captured by the $\mathrm{CRM}_{\mathrm{full}}$ with a correlation of $r = 0.91$ and RMSE = 1.66, which is better than the $\mathrm{CRM}_{\mathrm{uniform}}$ and the contribution model. The best-fitting availability values were $p = 0.7$ for the $\mathrm{CRM}_{\mathrm{uniform}}$, and $p_{\mathrm{low}} = 0.25$ and $p_{\mathrm{high}} = 0.75$ for the $\mathrm{CRM}_{\mathrm{full}}$. Figure 7 shows model predictions compared to mean judgments across all trials. Table 2 compares all three models. The uniform model captures participants' judgments somewhat ($r = 0.78$, RMSE = 2.49), but does not perform well in cross-validation and only best explains 4 out of 50 individual participants. The full model accounts best for participants' judgments overall and also best explains a majority of 35 individual participants. The replaceability predictor was notable in the full model.

**Object condition.** Responsibility judgments in the object condition followed the same pattern as those in the agent condition, but were overall lower. The best-fitting availability values were $p = 0.75$ for the $\mathrm{CRM}_{\mathrm{uniform}}$, and $p_{\mathrm{low}} = 0.6$ and $p_{\mathrm{high}} = 0.8$ for the $\mathrm{CRM}_{\mathrm{full}}$. Here, the replaceability predictor in the full model was also notable. While both the $\mathrm{CRM}_{\mathrm{uniform}}$ and $\mathrm{CRM}_{\mathrm{full}}$ make predictions that correlate highly with participants' judgments ($r = 0.92$), the full model outperforms the other two models in cross-validation and best explains the most individual participants.

## Discussion

In this experiment, we tested a more comprehensive version of the CRM. We manipulated not only the number of replacements but also the probability that each replacement would have been available. The CRM predicts that availability influences responsibility

Table 2
*Results of model comparison for Experiment 2. 'Intercept' and 'Replaceability' show the posterior means or each predictor along with 95% highest density intervals (HDIs). The contribution model only included an intercept as predictor, while the CRM models additionally computed replaceability by assuming either a uniform p (Equation 3) or varying p (Equation 3). r = Pearson correlation coefficient and RMSE = root mean squared error. "Δelpd" shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models, along with the associated standard error. Lower numbers indicate worse performance. "n best" is the number of participants whose judgments were best predicted by each model. The results show that replaceability is a notable predictor of participants' responsibility judgments in both conditions.*

| Model | Intercept | Replaceability | $r$ | RMSE | Δelpd (se) | $n$ best |
|---|---|---|---|---|---|---|
| *Agent condition* | | | | | | |
| $\mathrm{CRM}_{\mathrm{full}}$ | 76.91 [66.32, 87.33] | −12.96 [−24.23, −1.44] | 0.91 | 1.66 | 0 (0) | 35 |
| $\mathrm{CRM}_{\mathrm{uniform}}$ | 78.27 [69.72, 86.76] | −12.56 [−19.99, −5.09] | 0.78 | 2.49 | −411.7 (35.9) | 4 |
| Contribution | 67.54 [62.80, 72.35] | | | 4.11 | −586.4 (40.2) | 11 |
| *Object condition* | | | | | | |
| $\mathrm{CRM}_{\mathrm{full}}$ | 76.01 [66.18, 85.59] | −19.17 [−30.55, −7.88] | 0.92 | 2.01 | 0 (0) | 28 |
| $\mathrm{CRM}_{\mathrm{uniform}}$ | 76.52 [67.55, 85.54] | −19.35 [−29.84, −8.98] | 0.92 | 2.13 | −61.8 (6.8) | 12 |
| Contribution | 59.49, [54.41, 64.56] | | | 4.11 | −307.3 (24.8) | 10 |

because it affects the probability of successful replacement – the addition of many replacements means very little if they are all unavailable, for example. The results show that participants' responsibility judgments were sensitive to both number of replacements and individual availability and best explained by a full version of the CRM that considers both of them. The better performance of the full CRM over one that assumes uniform replaceability demonstrates that responsibility judgments cannot be explained by a simple heuristic that only considers the total number of agents or objects present in the scene. Nor can participants' judgments be explained by a model that only considers the contributions themselves, since those were constant across all trials.

When we looked at individual participants' judgments, we found considerable variation. Like in Experiment 1, there were two main groups of response patterns, which was also reflected in participants' free-response comments about what factors influenced their judgments. Most participants explicitly mentioned the number and availability of the replacements (e.g. "If a tradesman's colleagues are all very busy and he agreed to help build the ship I deemed him more responsible for the success (as if he didn't step up, the others may have refused to help).") and this corresponded roughly with those best fit by either of the CRM models. Some participants, however, focused only on what actually happened (e.g. "Success, to my understanding, is defined as making the machine work by having (at least) one of each three different gears in working condition. All were in working condition, so all (gears) were very much equally responsible for success.") and this minority was best fit by the baseline model.

Overall, we found that the more likely a replacement was available for a particular contribution – which increased the more replacements there were and the more individually available each one was – the less responsible participants tended to hold that contributor for the outcome. The difference between low and high availability seemed to matter more for agents than objects. This may be due to participants having more uncertainty about agents. The state of a machine part, perceived as generally reliable, may not convey much information compared to the state of a person, who can exhibit a vast range of possible behaviors. We also found overall somewhat lower judgments in the object condition compared to the agent condition. One possibility for this could be that in the object condition, part of the responsibility goes to the engineers who designed the gears rather than the gears themselves.

We deliberately did not specify the prior availability of the actual contributor because we didn't want information about the contribution itself to influence judgments. However, participants could still have made inferences based on the availability of the replacements. They could have reasoned that, for instance, if all the replacement carpenters had low availability, then perhaps the carpenter who actually helped had high availability and was able to do so precisely because of that. On the other hand, perhaps the common low availability of all the other carpenters suggests that carpentry is very demanding in general and thus the carpenter who actually helped did so *despite* having low availability. These would have had opposite influences on responsibility judgments, assuming that the prior availability of the person who contributed actually matters. In Experiment 3, we explored how the prior availability of the contributor affects responsibility judgments.

## Experiment 3: Availability of contributor

Experiment 3 investigates how the prior availability of the contributor affects responsibility judgments. The CRM predicts responsibility by considering how a counterfactual situation in which a particular contribution had not been made would have unfolded, but it doesn't consider features of the contributing cause itself, or how crucially a replacement might have been needed in the first place. For instance, although there was high turnover for pickpockets in *Ocean's 8*, if Constance had been very eager or very reluctant to join the team, then the turnover rate would have mattered to different extents. Responsibility in general is affected by characteristics of the person themselves, such as their capacity, which influences how feasible different actions were for them (e.g. Malle et al., 2014).

The prior availability of the contributor maps onto the if-likelihood in the counterfactual potency model (Petrocelli et al., 2011). Consider the counterfactual statement: "If another hacker had been recruited instead of Nine Ball, then the heist would have failed." The potency of this counterfactual depends on the if-likelihood (how easy it is to imagine that another hacker could have been recruited), and the then-likelihood (how plausible it is that the heist would have failed in that case). If there was never any doubt that Nine Ball would be the hacker on the team, then the if-likelihood would be low, rendering the counterfactual impotent. Similarly, if there were many other highly-skilled hackers around that would have also done a successful job, then the then-likelihood would be low and potency low as well. Accordingly, Nine Ball would be attributed less responsibility for the successful outcome in either of these cases.More generally, the more available a contributor was, the less likely a counterfactual replacement might have been needed. In turn, the lower the if-likelihood, the less potent the counterfactual scenario in which a replacement was made and the outcome failed, and thus the less responsible the contributor would be held.

However, it is also possible for lower if-likelihood to actually correspond to *more* responsibility. Consider the difference between a killer who acts intentionally with no doubts and one who wavers back and forth before committing the act. The counterfactual scenario in which the intentional killer had not acted has low if-likelihood because it seems implausible for them to not act. In contrast, the counterfactual in which the hesitant killer had not acted has high if-likelihood because it is easy to imagine them changing their mind to not act. Counterintuitively, potency predicts that the intentional killer would be less responsible than the hesitant one. Thus, it's unclear how the prior availability of the contributor actually affects their responsibility for the outcome. In Experiment 3, we built on the previous two experiments by additionally manipulating whether each contributor in each trial had prior low or high availability.

### Methods

The experiment was programmed in jsPsych (de Leeuw, 2015) and pre-registered (agent condition: `https://osf.io/gxjs6`; object condition: `https://osf.io/6svnt`).

**Participants.** Participants were Stanford undergraduates who were granted 0.5 credit hours for completing the experiment online. 102 students were recruited. Two were excluded for submitting multiple times, leaving a final sample size of $N = 100$ (*age*: M = 20, SD = 1; *gender*: 43 male, 56 female, 1 undisclosed; *race*: 36 White, 7 Black, 44 Asian, 1 American Indian/Alaska Native, 6 Multiracial, 6 undisclosed). They were randomly split between the *agent* and *object* conditions with $n = 50$ in each.

**Procedure & design.** The setup and design of the experiment followed that of Experiments 1 and 2. In each trial, participants were shown the availability of all craftspeople or gears, including the ones that actually helped and all of their potential replacements. To prevent each scene from appearing too visually overwhelming, we used two contributors in each trial instead of three. Furthermore, to isolate the influence of the craftsperson or gear that actually helped versus the replacements on responsibility judgments, each trial featured a situation in which the two contributors had the same number of replacements and either differed in their own availability, or in the availability of their possible replacements. For instance, in trial 1 in the agent condition, both the carpenter and tailor who helped have low availability, and there is one other carpenter and tailor in the village. The other carpenter has low availability while the other tailor has high availability. Participants were asked about both groups in every trial. We designed 20 different situations in which the the prior availability of the contributor was either low or high, the number of replacements ranged from 0 to 3, and the availability of each replacement was either low or high (see Table C1 for details). For example, one possible situation (namely, situation 8 in Figure 8) consists of a contributor with high prior availability and two low availability replacements. We distributed these situations across 19 different trials. Each trial contrasted two situations with the same number of replacements – one for each contribution – so each situation appeared in two different trials (with the exception of the case with zero replacements). Participants took an average of 7 minutes (SD = 2.8)[3].

## Results

Figure 8 shows participants' mean responsibility judgments across the 20 different situations we tested. They are ordered by increasing availability of the contributor, number of possible replacements, and availability of the replacements. For both conditions, we fit three Bayesian mixed effects models to participants' responsibility judgments. The first model includes an intercept and replaceability as a fixed effect (calculated using Equation 3), as well as an additional fixed effect of the prior availability of the contributor, $p_{\text{contributor}}$. This was equal to either $p_{\text{low}}$ or $p_{\text{high}}$. We call this model the CRM* because it computes and uses the probability of replacement, like the CRM, but it also takes into account the availability of the contributor (which may affect the responsibility judgment positively or negatively). The second model is a counterfactual potency (CP) model which includes an intercept and potency as a predictor. With respect to the counterfactual, "IF the contributor had been unavailable, THEN the outcome would have failed," if-likelihood is how plausibly the contributor might have been busy or broken (i.e. the complement of $p_{\text{contributor}}$), and then-likelihood is how likely no replacement would have been available (i.e. the complement of replaceability). Thus, we compute potency for each trial as

$$\begin{aligned}\text{potency} &= \text{if-likelihood} \times \text{then-likelihood}\\ &= \big(1 - p_{\text{contributor}}\big) \times \big((1 - p_{\text{low}})^{n_{\text{low}}}(1 - p_{\text{high}})^{n_{\text{high}}}\big). \qquad (4)\end{aligned}$$

[3] For reporting this time, we excluded one outlier participant who took 10.5 hours to complete the experiment (may have e.g. left the study open in their browser overnight before submitting), but included their data otherwise.

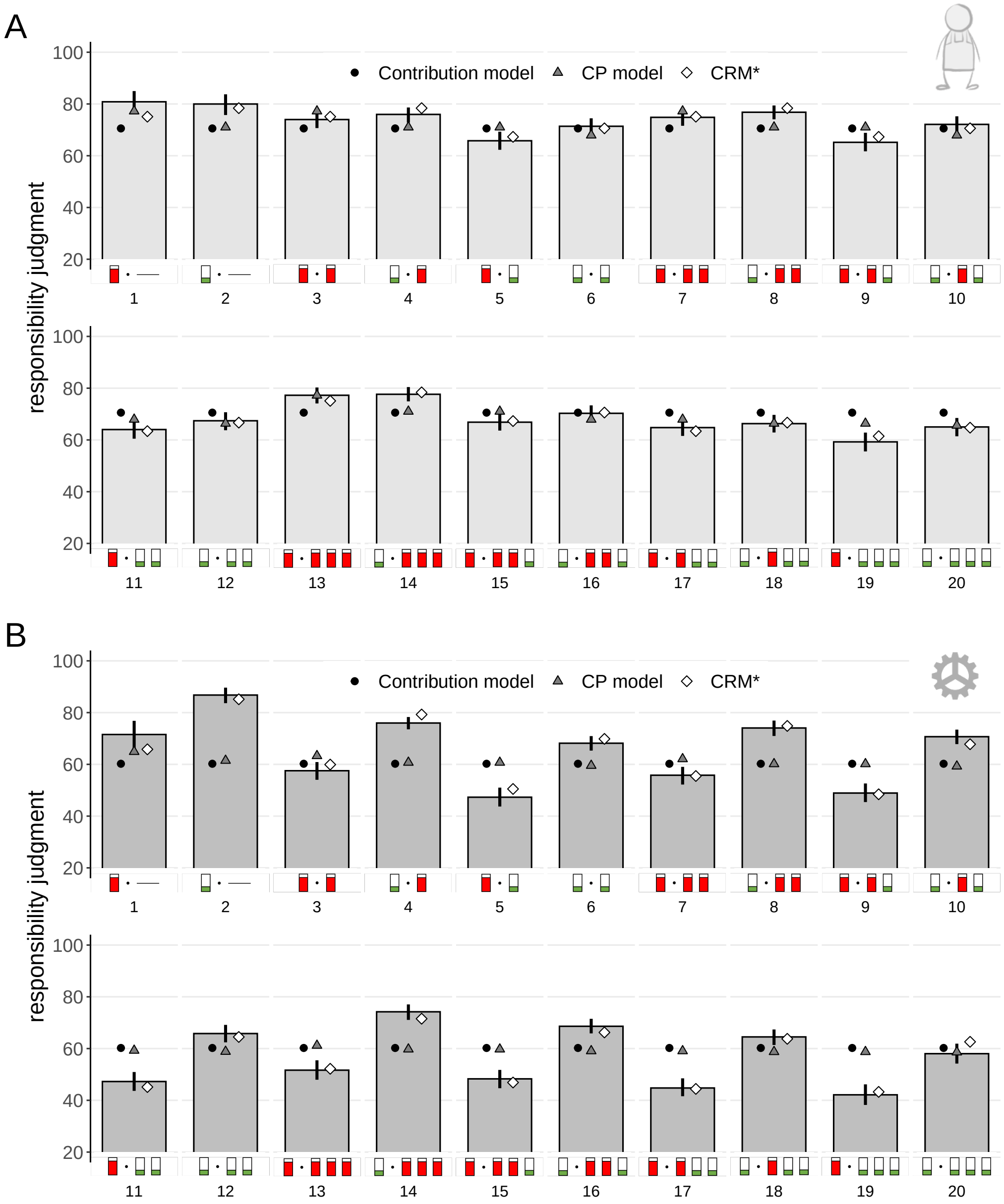

*Figure 8*. Mean responsibility judgments in the (A) agent and (B) object conditions in Experiment 3. Each situation on the x-axis is formatted as "contributor · replacements". The actual contributor had low or high prior availability and up to three possible replacements, each of which also had low or high availability. The situations are numbered by increasing availability and number of replacements. The different shaded symbols represent model predictions. Error bars are bootstrapped 95% confidence intervals. Note that the y-axis is truncated: participants judged responsibility on a scale from 0 to 100.

The parameters $p_{\text{low}}$ and $p_{\text{high}}$ were fit to minimize squared error between model predictions and participants' judgments in each condition. We ran a grid search over values between 0

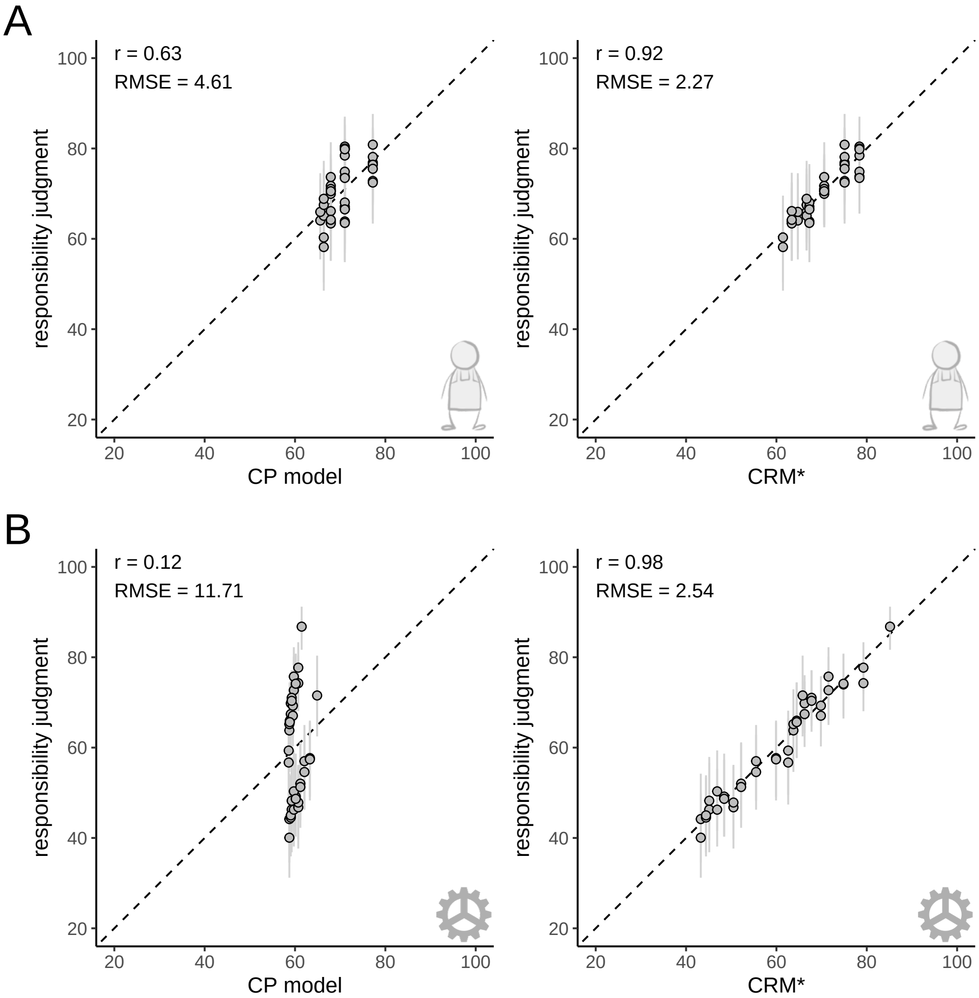

*Figure 9*. Scatter plots showing the relationship between mean responsibility judgments and model predictions in Experiment 3, in the (A) agent and (B) object conditions. $r$ = Pearson correlation coefficient, RMSE = root mean squared error. Error bars are bootstrapped 95% confidence intervals. Note that the y-axis is truncated: participants judged responsibility on a scale from 0 to 100.

and 1 with the only constraint being that $p_{\text{low}} < p_{\text{high}}$ (see Figure C1 for parameter search details). Finally, we also fit a third model that only included an intercept to represent the contribution model. All three Bayesian mixed effects models had random intercepts for each participant, and both the CRM* and CP model also included random slopes.

We found two main trends across both conditions from the results in Figure 8. First, the more replacements there were for a particular contribution and the more available those replacements were, the less responsible participants tended to hold that contribution. This replicates the results from Experiments 1 and 2. The second trend is that, for a fixed set of replacements, people tended to hold the contributor more responsible if they had high prior availability. These differences are especially noticeable in the object condition. We discuss the results from each condition in turn.

**Agent condition.** The filled symbols in Figure 8 indicate model predictions. Participants' judgments in the agent condition were well-captured by the CRM* with a correlation of $r = 0.92$ and RMSE = 2.27, outperforming the CP and contribution models (see Figure 9). The best-fitting availability values were $p_{\text{low}} = 0$ and $p_{\text{high}} = 0.5$.[4] In the CRM*, replaceability was a notable predictor, but not the availability of the contributor, as the 95% HDI on this predictor includes zero (see Table 3). However, the mean of the posterior for this predictor is positive, indicating that contributors who were more available received *more* responsibility, against the predictions of the CP model. Table 3 also summarizes a model comparison based on leave-one-out cross validation as well as individual participant best fit. The results show that the CP model captures participants' judgments somewhat ($r = 0.63$, RMSE = 4.61), but fares poorly in cross-validation and only best explains 9 out of 50 individuals. The CRM*, on the other hand, best accounts for the overall data. It performs best on the cross-validation despite having one more parameter than the CP model, and two more than the contribution model. The CRM* also best captures a majority of 28 out of 50 individuals.

**Object condition.** The results in the object condition were similar to those in the agent condition. Participants' responsibility judgments were well-captured by the CRM* with a correlation of $r = 0.98$ and RMSE = 2.54, far better than the CP and contribution

Table 3
*Experiment 3 model comparison. ‘Intercept‘, ‘Contributor‘, and ‘Replaceability‘, and ‘Potency‘ show the posterior means of each predictor along with 95% highest density intervals (HDIs). The contribution model only included an intercept as a predictor, while the CRM* also included the availability of the contributor and replaceability (Equation 3) and the CP model also included potency (Equation 4). r = Pearson correlation coefficient and RMSE = root mean squared error. “Δelpd” shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models, along with the associated standard error. Lower numbers indicate worse performance. “n best” is the number of participants whose judgments were best predicted by each model.*

| Model | Intercept | Contributor | Replaceability | $r$ | RMSE | Δelpd (se) | $n$ best |
|---|---|---|---|---|---|---|---|
| *Agent condition* | | | | | | | |
| CRM* | 74.91 [67.96, 81.84] | 6.58 [−8.43, 22.72] | −15.59 [−22.04, −9.09] | 0.92 | 2.27 | 0 (0) | 28 |
| CP | 64.82 [57.02, 72.70] | Potency: 12.40 [3.65, 21.32] | | 0.63 | 4.61 | −579.9 (55.9) | 9 |
| Contribution | 70.55 [64.61, 76.60] | | | | 8.24 | −793.0 (61.8) | 13 |
| *Object condition* | | | | | | | |
| CRM* | 53.71 [42.48, 64.39] | 48.37 [33.10, 64.03] | −23.55 [−32.97, −14.03] | 0.98 | 2.54 | 0 (0) | 37 |
| CP | 58.47 [51.50, 65.47] | Potency: 8.56 [−6.33, 23.19] | | 0.12 | 11.71 | −611.3 (40.5) | 2 |
| Contribution | 60.23 [54.24, 66.23] | | | | 8.96 | −710.6 (42.5) | 11 |

[4] The fact that the best-fitting value for $p_{\text{low}}$ is 0 implies that the number of low availability replacements does not affect responsibility. For example, participants provided very similar responsibility judgments in situations 3, 7, and 13 in Figure 8 and the model captures this. While 0 turned out to be the best-fitting value for this experiment, the loss gradient for this parameter is smooth as Figure C1 shows. So even if the parameter took on a slightly higher value, the model would still be able to capture participants' judgments well.

models. The best-fitting availability values were $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.65$. The contributor predictor was notably positive in the CRM* (see Table 3). Like in the agent condition, this suggests that the availability of the contributor affects responsibility judgments in the opposite direction of what the CP model would predict. In the CP model, potency was not notable. The results of the cross-validation show that the CRM* best explains judgments overall, despite having the most parameters, and also best captures a majority of 37 out of 50 individual participants.

### Discussion

In this experiment, we found that responsibility judgments were well-predicted by a combination of both the probability of counterfactual replacement and the prior availability of the contributor. The more likely a replacement would have been successful, the less responsible participants tended to hold the contributor. This finding is consistent with the results from Experiments 1 and 2. Additionally, the more available the contributor was, the *more* responsible participants judged the contributor to be. This pattern is captured by the CRM*, but the opposite of what the CP model predicts. The key difference between the CRM* and the CP model is that the former assumes additive effects of the contributor and replaceability predictors, while the latter assumes a multiplicative effect. Counterfactual potency (Petrocelli et al., 2011) suggests that if-likelihood and then-likelihood influence responsibility in the same direction. In other words, the probability of replacement should be particularly important when a replacement is likely to be needed in the first place. But we found the opposite effect here, which resulted in the CRM* outperforming the CP model in both conditions.

The effect of the contributor on responsibility judgments was stronger in the object condition. Looking at the subset of participants who were best fit by the CRM*, we found that more participants assigned a positive weight to the contributor in the object condition compared to the agent condition. There was a wider range of posterior means for the contributor predictor in the agent condition – some participants placed little weight on this term, while others had strongly positive or strongly negative weights. This variation likely led to the positive, but not notable, effect in the agent condition.

Why did we find that a contributor was held *more* responsible when it was less likely that they needed to be replaced? One possibility is that participants used the information about the contributor's prior availability to make additional inferences about their contribution. For example, they might have inferred that busier craftspeople put less effort into their actions and thus deserved less responsibility. We will return to this point in the general discussion.

## General discussion

From determining who is at fault after a regrettable company decision, to naming the Most Valuable Player in a sports team, how people assign responsibility to individuals in groups is a complex question with important implications for our everyday lives. In this paper, we developed the Counterfactual Replacement Model (CRM), a computational model that explains responsibility judgments in terms of how easily a person's contribution could have been replaced. The CRM considers how a group situation would have turned out had a particular contribution not been made, while holding everything else that happened constant. It computes how likely the contribution could have been replaced, and predicts that the more likely a successful replacement could have been made, the less responsible the actual contribution was for the outcome. To test the model, we designed an experimental setting where we manipulated two parameters: the number of possible replacements, and the probability that each replacement would have been available to contribute instead. We also test an extension of the model in which we manipulated a third parameter, the prior availability of the actual contribution.

We tested the CRM across three experiments. In Experiment 1, we varied the number of replacements. In Experiment 2, we varied both the number of replacements and their availability. In Experiment 3, we tested an extension of the CRM that considers both the replaceability and the prior availability of the contributor. Across all three experiments, the CRM accurately captured participants' judgments in both a social domain, where the contributions were made by agents, and a physical domain, where the contributions were made by components of a mechanistic device.

The CRM outperforms alternative models that only consider the actual contributions, and its extension in Experiment 3 outperforms a model based on counterfactual potency (CP, Petrocelli et al., 2011). The CRM focuses on the role that counterfactual reasoning about particular causes and their replacements play in responsibility judgments, but it does not make any claims about the effect of the contributor itself. In contrast, the potency model predicts responsibility judgments to be a multiplicative combination of replaceability and the prior availability of the contributor. The CP model doesn't capture participants' responsibility judgments in Experiment 3. While the effect of replaceability was in the predicted direction, the effect of the contributor was in the opposite direction. Participants judged contributors to be more responsible when they were *more* available to contribute in the first place and hence less likely to require a counterfactual replacement.

These effects may have arisen from inferences participants made about the contributors based on their availability. For example, busyness might be suggestive of a craftsperson's skills (i.e. good craftspeople might be in greater demand), but might also be indicative of how much effort they are capable of putting into the job (i.e. a busy craftsperson might have less time to work on the job). Much work has shown that various factors about an agent can affect responsibility attributions in general, including consideration of the agent's mental states like their intentions (Cushman, 2008; Lagnado & Channon, 2008; Lombrozo, 2010) and reasons (Cushman, 2008), their skills and capacities (Gerstenberg, Ejova, & Lagnado, 2011; Malle et al., 2014; Weiner & Kukla, 1970), and their character more generally (Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Uhlmann, Pizarro, & Diermeier, 2015; Zhao & Kushnir, 2022). The various inferences that are licensed based on the information about the contributor's prior availability were reflected in the individual model fits – some participants assigned more responsibility to busy contributors, some assigned less responsibility, and for some it made no difference.

In the following sections, we discuss several aspects of our model in more detail and propose directions for future work.

### Prior availability and normality

In Experiment 3, we manipulated the normality of the actual contribution by specifying whether the contributor had low or high prior availability. Participants tended to attribute more responsibility to contributors with high prior availability, especially in the object condition. However, much prior work has found that people often attribute greater responsibility and causality to abnormal events (e.g. Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard, Kominsky, & Knobe, 2017; Knobe, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). How does availability relate to normality in our paradigm? On one hand, having high availability implies being previously engaged in fewer helping tasks, which would make a particular instance of helping more abnormal for that contributor. On the other hand, for a contributor with high prior availability, helping in the moment is more normal for them as they are more likely to say "yes" compared to a contributor with low prior availability.

If we assume that high availability means that it's abnormal for the person to help, then our findings are consistent with prior research showing that people tend to attribute more causality to abnormal factors compared to normal ones in conjunctive causal structures in which each person's contribution is necessary for the outcome to come about (Gerstenberg & Icard, 2020; Icard et al., 2017; Kirfel & Lagnado, 2021; Kominsky et al., 2015). However, this interpretation relies on the assumption that current availability is diagnostic for past helping behavior. In reality, there are many reasons a craftsperson (or gear) can have low or high availability besides how many ships or machines they have already contributed to.

It seems more natural to instead interpret contributors with high availability as the more normal cause. In that case, our results in Experiment 3 run counter to what prior research has found. One possible explanation is that participants' judgments were affected by additional inferences they made about the contribution based on availability. For example, some participants inferred quality from availability (e.g. "I rated green gears higher since durability means it most likely will support the machine longer."). Others made inferences about how much effort was put in (e.g. "If someone was busier, they likely had less time and energy to commit to the ship-building process, and thus were less of a contribution to the final product."). If busyness or brittleness supports inferences about the contribution, then this would explain the inverse relationship between the availability of the contributor and responsibility observed here. In the case of agents specifically, busyness could be used to make a skill inference. Some participants might have inferred that busier craftspeople had more skill and thus should be *more* responsible (e.g. "Business likely increased output and productivity.").

### Counterfactual selection

Any counterfactual simulation model must specify what counterfactuals to consider. Sometimes, it is most natural to think about what would have happened if a particular contribution had been replaced, as the CRM does, such as when a particular role must be filled in a heist or a sports game. In other contexts, however, it may be more natural to consider what would have happened if the person under consideration had acted differently, rather than what another person would have done (Lagnado & Gerstenberg, 2015). These two different ways of selecting counterfactuals have parallels in the law (Lagnado

& Gerstenberg, 2017). For example, to establish negligence, one must establish (among other things) that the defendant *breached* a duty of care and that the defendant's breach actually *caused* the negative outcome. To prove breach, jurors are often asked to consider how a "reasonable person" would have acted in the same situation (Simpson et al., 2020; Tobia, 2018; Uhlmann et al., 2015; Uhlmann & Zhu, 2013; Uhlmann, Zhu, & Tannenbaum, 2013). To establish proximate causation, one must show that the negative outcome would not have happened "but for" the defendant's actions (Summers, 2018). The "reasonable person" test requires reasoning about counterfactual replacement whereas the "but for" test involves reasoning about alternative counterfactual actions.

The CRM performs a combination of both types of counterfactual tests. First, it considers a "but for" test in which the craftsperson or gear who contributed had been busy or broken, and then it simulates the replacement process that would have followed. While the two tests are used for different purposes in the law, it's an interesting psychological question of what counterfactuals most naturally come to people's minds (Byrne, 2005; Gerstenberg & Stephan, 2021; Kominsky & Phillips, 2019). What counterfactuals someone considers depends on their causal knowledge of the situation, and on what's being evaluated. When considering whether a person's action was responsible for an outcome, we can imagine them not acting or taking a different action instead. But when we hold a person as a whole responsible instead, we may be more likely to compare them to other people who could have been in the same situation. In our experiments, the agents all just took a single action, so these two different ways of generating counterfactuals don't come apart in our setting. Future work should look into situations in which agents perform multiple actions and investigate how people assign responsibility in these cases.

### Counterfactual simulation

Simulating counterfactuals requires a generative causal model of the domain together with a specification of what counterfactual interventions can be performed. The counterfactual simulation model (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021) uses a physics engine to simulate counterfactual physical events in which an object had been removed from the scene. Here, we treat the group tasks in our paradigm as essentially a sequence of coin flips. The CRM considers the counterfactual situation in which one of the coins (representing one of the craftspeople or gears) had been set to a different value (from "yes" to "no"), and simulates flipping the remaining coins (representing the possible replacements) to determine whether the outcome would have been different. The current work contributes towards a more general framework of assigning responsibility that is grounded in causal models of the situation, which allow for the evaluation of relevant counterfactuals (Gerstenberg et al., 2021; Pearl, 2000). Future work can build on these models to incorporate the many additional factors that are known to influence responsibility judgments.

In our setting, the CRM assumes a deterministic relationship between a successful counterfactual replacement and a successful outcome. That is, if at least one other carpenter had said "yes" to helping, then the ship would have been built. In a more complex situation where, for instance, the quality of the ship also matters, it would matter not only how likely a replacement carpenter could have been found, but also whether and how the replacement would have acted. Would they have done a better job, or a worse job?

In real-world situations, people's causal knowledge is often much richer than what the CRM captures so far. For example, we may ask whether a basketball team would still have won the game if one of the players had needed to be substituted out. This depends not only on the likelihood of finding a suitable bench player to substitute in – which we focused on here – but also on precisely how the game would have played out with the substitute. Complex counterfactual simulations may require people to abstract certain elements of their causal models or to rely on heuristics. To predict responsibility judgments in a situation like a basketball gameplay, a model would need to be able to generate sequences of counterfactual states over time, rather than a binary counterfactual outcome. In future work, we will explore people's responsibility judgments in situations that involve multiple actions over time.

### Limitations and future directions

In this work, we only tested successful group outcomes. What about failures? Past work shows that there are asymmetries between praise for positive outcomes and blame for negative ones. For example, Guglielmo and Malle (2019) have shown that blame judgments for individual actions tend to be more extreme than praise judgments. In group settings, people also tend to adopt a praise-many, blame-fewer strategy where they hold more members of the group responsible for positive outcomes compared to negative outcomes (Schein, Jackson, Frasca, & Gray, 2020). Future work should investigate how replaceability may matter differently for group successes versus group failures.

Another limitation of this work is that only two different values of the availability parameter were concretely tested, one for low and one for high availability. In principle, however, every replacement could have a distinct probability of success. We fitted the model parameters to participants' judgments in each experimental condition. The CRM then computes replaceability explicitly for each scenario given the fitted values in each trial. Prior research and anecdotal evidence shows that people are generally quite poor at estimating probabilities in such conjunctive scenarios, however. For instance, consider $n = 3$ with $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.75$. If all three replacements had low availability, then the probability of successful replacement would be $1 - (1 - 0.25)^3 = 0.58$. If all three replacements had high availability, then the probability would be $1 - (1 - 0.75)^3 = 0.98$. But people tend to believe that the second value is only slightly higher than the first, failing to recognize how rapidly conjunctive probability drops off (Bar-Hillel, 1973; Nilsson, Rieskamp, & Jenny, 2013). Although participants were sensitive to the difference between these two scenarios, they probably didn't have an accurate sense of the actual values. Thus, follow-up work might empirically test for and use people's subjective estimates of probability of replacement in each trial directly, instead of computing replaceability.

Finally, future work should study more closely what inferences participants make based on a contributor's prior availability. Future experiments can manipulate availability in different ways that go beyond "busy" agents and "brittle" objects. For the CRM, only the probability with which a successful replacement could have been found matters, but the inferences that people draw about a person's contribution cleary matter, too.

## Conclusion

In this paper, we developed and tested a computational model that predicts how responsible a particular cause is for a group outcome by considering how easily that cause could have been replaced. The model captures participants' judgments in increasingly complex situations, where multiple factors jointly determine the replaceability of a particular contribution. This work brings us one step closer towards a comprehensive computational account of how responsibility is attributed to individuals in groups.

## Acknowledgments

This work was supported by the NSF GRFP and a grant from the Stanford Institute for Research in the Social Sciences to S.A.W., and a grant from the Stanford Human-Centered Artificial Intelligence Institute to T.G. All experiments were approved by Stanford's Institutional Review Board.

# References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., . . . Rahwan, I. (2020, February). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*(2), 134–143. (Number: 2 Publisher: Nature Publishing Group) doi: 10.1038/s41562-019-0762-8

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, *9*(3), 396–406.

Brewer, M. B. (1977). An information-processing approach to attribution of responsibility. *Journal of Experimental Social Psychology*, *13*(1), 58–69.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality.* MIT Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Caruso, E. M., Epley, N., & Bazerman, M. H. (2006). The costs and benefits of undoing egocentric responsibility assessments in groups. *Journal of Personality and Social Psychology*, *91*(5), 857.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Falk, A., Neuber, T., & Szech, N. (2020). Diffusion of Being Pivotal and Immoral Outcomes. *The Review of Economic Studies*, *87*(5), 2205–2229.

Falk, A., & Szech, N. (2013). Morals and markets. *Science*, *340*(6133), 707–711.

Felsenthal, D., & Machover, M. (2004). A priori voting power: what is it all about? *Political Studies Review*, *2*(1), 1–23.

Fincham, F. D., & Jaspars, J. M. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology*, *22*(2), 145–161.

Forsyth, D. R., Zyzniewski, L. E., & Giammanco, C. A. (2002). Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, *28*(1), 54–65.

Gantman, A. P., Sternisko, A., Gollwitzer, P. M., Oettingen, G., & Van Bavel, J. J. (2020). Allocating moral responsibility to multiple agents. *Journal of Experimental Social Psychology*, *91*, 104027.

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 720–725). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(6), 936–975.

Gerstenberg, T., & Icard, T. F. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

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

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review*, *19*(4), 729–736.

Gerstenberg, T., Lagnado, D. A., & Zultan, R. (2023). Making a difference: Criticality in groups. *PsyArXiv*. Retrieved from `https://psyarxiv.com/xwm3g/`

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122-141.

Glover, J., & Scott-Taggart, M. (1975). It makes no difference whether or not i do it. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *49*, 171–209.

Green, R. M. (1991). When Is "Everyone's Doing It" a Moral Justification? *Business Ethics Quarterly*, 75–93.

Guglielmo, S., & Malle, B. F. (2019, March). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLOS ONE*, *14*(3), e0213544. (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0213544

Halevy, N., Maoz, I., Vani, P., & Reit, E. S. (2022). Where the Blame Lies: Unpacking Groups Into Their Constituent Subgroups Shifts Judgments of Blame in Intergroup Conflict. *Psychological Science*, *33*(1), 76–89.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *11*, 587–612.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. doi: 10.1016/j.cognition.2017.01.010

Kaiserman, A. (2021). Responsibility and the 'Pie Fallacy'. *Philosophical Studies*, *178*(11), 3597–3616.

Kerr, N. L. (1996). "Does my contribution really matter?": Efficacy in social dilemmas. *European Review of Social Psychology*, *7*(1), 209–240.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*(1), 78–94.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Knobe, J. (2009). Folk judgments of causation. *Studies In History and Philosophy of Science Part A*, *40*(2), 238–242.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, *43* (11), e12792. Retrieved from http://dx.doi.org/10.1111/cogs.12792 doi: 10.1111/cogs.12792

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Koskuba, K., Gerstenberg, T., Gordon, H., Lagnado, D. A., & Schlottmann, A. (2018). What's fair? how children assign reward to members of teams with differing causal structures. *Cognition*, *177*, 234-248.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108* (3), 754–770.

Lagnado, D. A., & Gerstenberg, T. (2015). A difference-making framework for intuitive judgments of responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 3, pp. 213–241). Oxford University Press.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

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

Livengood, J. (2011). Actual causation and simple voting scenarios. *Noûs*, 1–33.

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

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25* (2), 147-186.

Nilsson, H., Rieskamp, J., & Jenny, M. A. (2013). Exploring the overestimation of conjunctive probabilities. *Frontiers in Psychology*, *4* .

Parker, J. R., Paul, I., & Reinholtz, N. (2020, Mar). Perceived momentum influences responsibility judgments. *Journal of Experimental Psychology: General*, *149* (3), 482–489.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, England: Cambridge University Press.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, *100* (1), 30–46.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Sanders, J., Lee Hamilton, V., Denisovsky, G., Kato, N., Kawai, M., Kozyreva, P., ... Tokoro, K. (1996). Distributing Responsibility for Wrongdoing Inside Corporate Hierarchies: Public Judgments in Three Societies. *Law & Social Inquiry*, *21* (4), 815–855. doi: 10.1111/j.1747-4469.1996.tb00098.x

Savitsky, K., Van Boven, L., Epley, N., & Wight, W. M. (2005). The unpacking effect in allocations of responsibility for group tasks. *Journal of Experimental Social Psychology*, *41* (5), 447–457.

Schaffer, J. (2010). Contrastive causation in the law. *Legal Theory*, *16* (04), 259–297.

Schein, C., Jackson, J. C., Frasca, T., & Gray, K. (2020). Praise-many, blame-fewer: A common (and successful) strategy for attributing responsibility in groups. *Journal of Experimental Psychology: General*, *149*(5), 855. (Publisher: US: American Psychological Association) doi: 10.1037/xge0000683

Schroeder, J., Caruso, E. M., & Epley, N. (2016). Many hands make overlooked work: Over-claiming of responsibility increases with group size. *Journal of Experimental Psychology: Applied*, *22*(2), 238–246.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* Springer-Verlag, New York.

Simpson, A., Alicke, M. D., Gordon, E., & Rose, D. (2020). The reasonably prudent person, or me? *Journal of Applied Social Psychology*, *50*(5), 313–323.

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

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348.

Summers, A. (2018). Common-sense causation in the law. *Oxford Journal of Legal Studies*, *38*(4), 793–821.

Tobia, K. P. (2018). How people judge what is reasonable. *Alabama Law Review*, *70*, 293–359.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Uhlmann, E. L., & Zhu, L. L. (2013). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science*, *5*(3), 279–285.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432.

Weiner, B. (1993). A Theory of Perceived Responsibility and Social Motivation. *American Psychologist*, 9.

Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, *15*(1), 1–20.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *56*(2), 161–169.

Xiang, Y., Vélez, N., & Gershman, S. J. (2022). Collaborative decision making is grounded in representations of other people's competence and effort. *PsyArXiv*. Retrieved from `https://psyarxiv.com/gcnrq/`

Zhao, X., & Kushnir, T. (2022). When it's not easy to do the right thing: Developmental changes in understanding cost drive evaluations of moral praiseworthiness. *Developmental science*, e13257.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, *125*(3), 429–440.

# Appendix

## Appendix A. Additional information about Experiment 1

Table A1

*Number of replacements for each trial in Experiment 1. These were randomly permuted among the three contributors in each trial. For instance, trial 15 represents a scene in which two contributors each have three possible replacements and the third contributor has one. In the agent condition, the tailor happened to have one replacement in this trial (see Figure 1). In the object condition, the yellow gear happened to have one replacement in this trial (see Figure 3).*

| Trial | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Agent condition* | | | | | | | | | | | | | | | | | | | |
| Carpenters | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 3 | 1 | 1 | 3 | 2 | 2 | 3 | 2 | 3 | 3 | 3 |
| Blacksmiths | 1 | 0 | 3 | 0 | 1 | 1 | 2 | 3 | 3 | 1 | 2 | 1 | 2 | 3 | 3 | 2 | 2 | 2 | 3 |
| Tailors | 0 | 2 | 0 | 1 | 2 | 3 | 2 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 |
| *Object condition* | | | | | | | | | | | | | | | | | | | |
| Yellow gears | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 3 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 3 | 3 |
| Green gears | 0 | 0 | 3 | 0 | 0 | 3 | 2 | 2 | 0 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 3 |
| Blue gears | 1 | 2 | 0 | 1 | 1 | 1 | 2 | 3 | 3 | 1 | 1 | 3 | 2 | 1 | 3 | 2 | 2 | 2 | 3 |

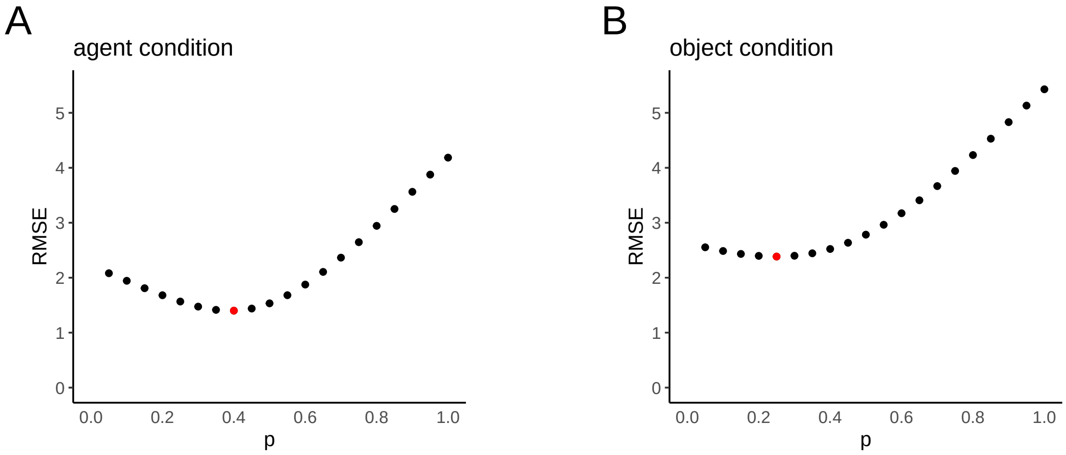

*Figure A1*. Results of fitting a uniform probability of success parameter in Experiment 1, in the (A) agent and (B) object conditions. The best-fitting values, indicated in red, minimize the squared error between model predictions and participants' judgments. The best-fitting parameter value was $p = 0.4$ in the agent condition, and $p = 0.25$ in the object condition. This means that, for example, any craftsperson would have a 0.4 chance of helping build the ship.

## Appendix B. Additional information about Experiment 2

Table B1

*Information about the replacements in each trial in Experiment 2. Like in Experiment 1, each set of replacements was randomly permuted among the three contributors in each trial. For instance, trial 1 features a contributor who had no replacements, which happened to be the carpenter in the agent condition and the yellow gear in the object condition.*

| Trial | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Agent condition* | | | | | | | | | | | | | | | | | | | | | |
| Carpenters | $n_{\text{low}}$ | 0 | 0 | 0 | 1 | 4 | 0 | 1 | 0 | 1 | 1 | 3 | 0 | 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| | $n_{\text{high}}$ | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 3 | 0 | 3 | 0 | 3 | 1 | 4 | 2 | 4 | 1 | 2 | 2 | 3 |
| Blacksmiths | $n_{\text{low}}$ | 4 | 0 | 2 | 3 | 0 | 3 | 0 | 4 | 2 | 3 | 0 | 2 | 1 | 2 | 2 | 1 | 3 | 2 | 3 | 1 |
| | $n_{\text{high}}$ | 0 | 4 | 1 | 1 | 1 | 1 | 4 | 0 | 1 | 0 | 1 | 1 | 3 | 1 | 2 | 1 | 0 | 1 | 0 | 1 |
| Tailors | $n_{\text{low}}$ | 1 | 3 | 4 | 0 | 2 | 1 | 2 | 0 | 2 | 1 | 1 | 1 | 0 | 0 | 2 | 1 | 2 | 1 | 1 | 2 |
| | $n_{\text{high}}$ | 3 | 1 | 0 | 0 | 2 | 3 | 2 | 1 | 2 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 1 | 1 | 1 | 0 |
| *Object condition* | | | | | | | | | | | | | | | | | | | | | |
| Yellow gears | $n_{\text{low}}$ | 0 | 0 | 2 | 3 | 4 | 1 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 0 |
| | $n_{\text{high}}$ | 0 | 0 | 1 | 1 | 0 | 3 | 4 | 1 | 1 | 0 | 1 | 0 | 2 | 2 | 2 | 4 | 1 | 1 | 0 | 3 |
| Green gears | $n_{\text{low}}$ | 1 | 0 | 4 | 1 | 0 | 0 | 2 | 0 | 2 | 1 | 3 | 2 | 3 | 2 | 2 | 1 | 2 | 1 | 1 | 1 |
| | $n_{\text{high}}$ | 3 | 4 | 0 | 2 | 1 | 1 | 2 | 3 | 2 | 3 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 2 | 1 | 1 |
| Blue gears | $n_{\text{low}}$ | 4 | 3 | 0 | 0 | 2 | 3 | 1 | 4 | 1 | 3 | 1 | 0 | 1 | 0 | 2 | 1 | 3 | 2 | 0 | 2 |
| | $n_{\text{high}}$ | 0 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 3 | 3 | 4 | 2 | 1 | 0 | 1 | 2 | 0 |

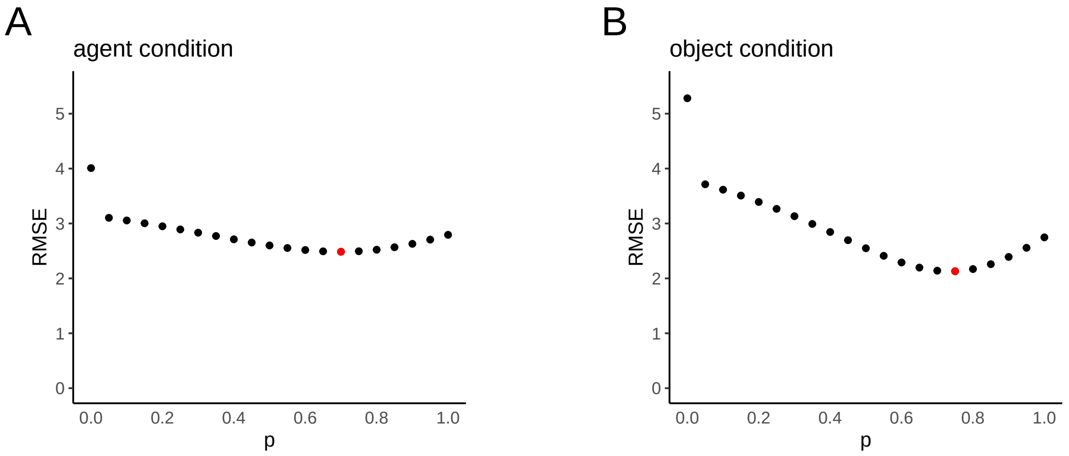

*Figure B1*. Results of fitting a uniform probability of success parameter in Experiment 2, in the (A) agent and (B) object conditions. The best-fitting values, indicated in red, minimize the squared error between model predictions and participants' judgments. This value was $p = 0.7$ in the agent condition and $p = 0.75$ in the object condition. This means that, for example, any craftsperson would have a 0.7 chance of helping build the ship.

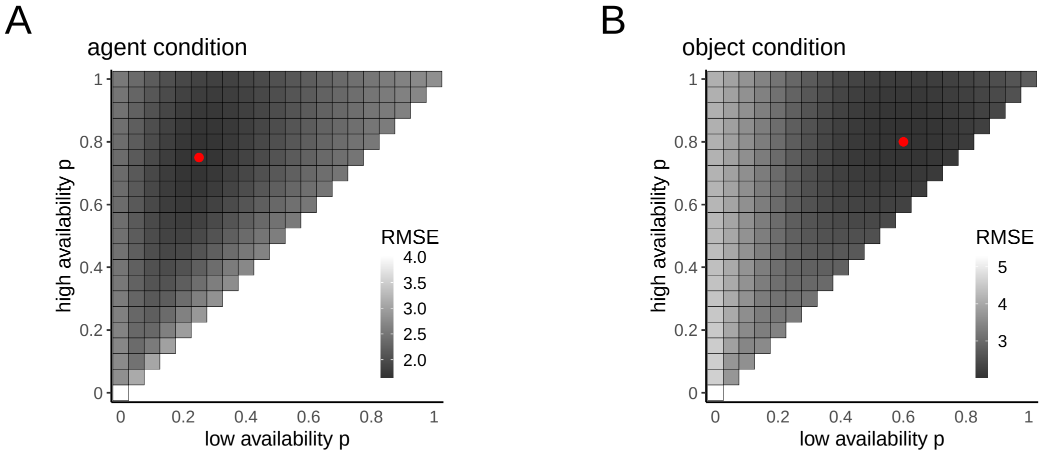

*Figure B2*. Results of a grid search over the two probability of success parameters for replacements with low and high availability in Experiment 2, in the (A) agent and (B) object conditions. The best-fitting parameters, indicated with the red dots, minimize the squared error between model predictions and participants' judgments. They were $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.75$ in the agent condition, and $p_{\text{low}} = 0.6$ and $p_{\text{high}} = 0.8$ in the object condition. This means that, for example, any craftsperson with high availability would have a 0.75 chance of helping build the ship.

## Appendix C. Experiment 3 additional information

Table C1

*Information about the replacements and the prior availability of the contributor (H = high, L = low) for each trial in Experiment 3. Like in Experiments 1 and 2, each situation was randomly assigned between the two contributors in each trial.*

| Trial | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Agent condition* | | | | | | | | | | | | | | | | | | | | |
| Carpenters | Contr. | H | L | H | H | H | H | L | L | H | H | H | H | L | L | H | H | H | H | H |
| | $n_{\text{low}}$ | 0 | 1 | 0 | 1 | 0 | 2 | 1 | 1 | 0 | 2 | 0 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 0 |
| | $n_{\text{high}}$ | 1 | 0 | 1 | 0 | 2 | 0 | 1 | 1 | 2 | 0 | 3 | 0 | 1 | 2 | 3 | 2 | 1 | 0 | 0 |
| Tailors | Contr. | H | L | L | L | H | H | L | L | L | L | H | H | L | L | L | L | L | L | L |
| | $n_{\text{low}}$ | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 2 | 0 | 2 | 2 | 1 | 0 | 3 | 0 | 1 | 2 | 3 | 0 |
| | $n_{\text{high}}$ | 0 | 1 | 1 | 0 | 1 | 1 | 2 | 0 | 2 | 0 | 1 | 2 | 3 | 0 | 3 | 2 | 1 | 0 | 0 |
| *Object condition* | | | | | | | | | | | | | | | | | | | | |
| Yellow gears | Contr. | H | L | H | H | H | H | L | L | L | H | H | H | L | L | H | L | L | L | L |
| | $n_{\text{low}}$ | 1 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 2 | 2 | 1 | 0 | 1 | 0 | 1 | 2 | 3 | 0 |
| | $n_{\text{high}}$ | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 2 | 0 | 1 | 2 | 3 | 2 | 3 | 2 | 1 | 0 | 0 |
| Blue gears | Contr. | H | L | L | L | H | H | L | L | H | L | H | H | L | L | L | H | H | H | H |
| | $n_{\text{low}}$ | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 3 | 2 | 3 | 0 | 1 | 2 | 3 | 0 |
| | $n_{\text{high}}$ | 1 | 0 | 1 | 0 | 2 | 1 | 2 | 0 | 2 | 0 | 3 | 0 | 1 | 0 | 3 | 2 | 1 | 0 | 0 |

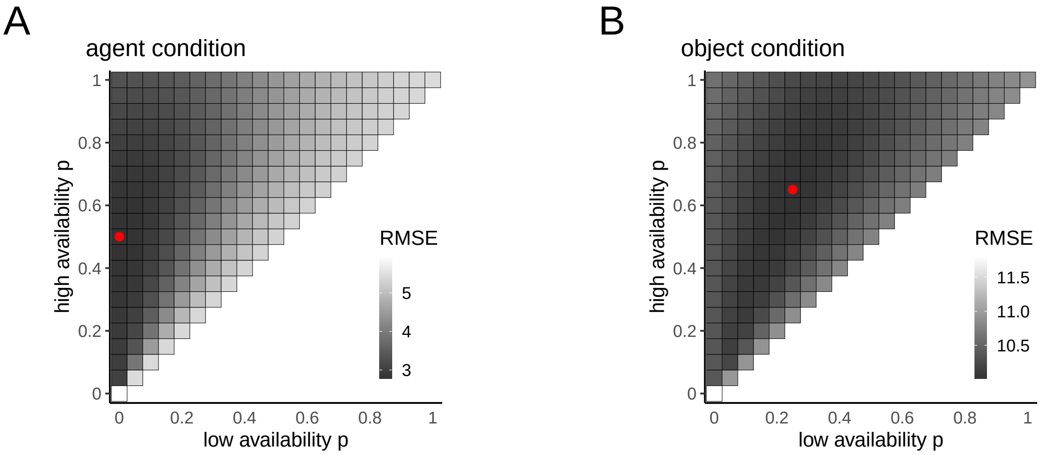

*Figure C1.* Results of a grid search over the two probability of success parameters for replacements with low and high availability in Experiment 3, in the (A) agent and (B) object conditions. The best-fitting parameters, indicated with the red dots, minimize the squared error between model predictions and participants' judgments. They were $p_{\text{low}} = 0$ and $p_{\text{high}} = 0.5$ in the agent condition, and $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.65$ in the object condition. This means that, for example, any craftsperson with high availability would have a 0.5 chance of saying helping build the ship.