If not me, then who?
Responsibility and replacement

Sarah A. Wu
Stanford University

Tobias Gerstenberg
Stanford University

Abstract

How do people hold each other responsible? In this paper, we show that responsibility in group settings is influenced by the notion of replaceability. The easier it would have been to replace a contributing cause, the less responsible that cause will be held for the actual outcome. We present a computational model that predicts responsibility judgments by considering how likely a counterfactual replacement could have produced the same outcome. To test the model, we developed a paradigm that allows us to quantitatively manipulate information about replaceability—the number of possible replacements and the probability that each one would have succeeded counterfactually—while keeping the actual events and outcome fixed. Across three experiments featuring increasingly complex scenarios, we show that the model explains participants' responsibility judgments well in both social and physical settings.

Keywords: responsibility; counterfactual reasoning; computational modeling

Corresponding author: sarahawu@stanford.edu

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 part 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, then 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 the role that 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 had a particular contribution not been made. 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. We then describe our model and test the predictions of the model in three experiments. We conclude by discussing the key contributions of our work as well as some limitations that may be addressed by future research.

Responsibility and contribution

Responsibility is linked to an understanding of contribution. When multiple causes affected an outcome, there are several ways to conceptualize the contribution that each cause made. First, contributions can differ in *value* (e.g. Caruso, Epley, & Bazerman, 2006). For example, one co-author may have written more pages of a manuscript than another. When the manuscript gets accepted, the person who wrote the larger portion (all else being equal) should be more responsible for the success. Second, contributions can differ in how close they were to *making a difference* (e.g. Chockler & Halpern, 2004; Lagnado, Gerstenberg, & Zultan, 2013). 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 holds greater responsibility for the marginal win there, than each of the ten members that voted for the landslide win of policy A. This intuition cannot be explained in terms of differing value since all committee members contributed the same value (a single vote) towards the total count for both policies. Finally, contributions can differ in how easily they could have been *replaced*. In *Ocean's 8*, Nine Ball and Constance's contributions are incomparable in terms of value because they fulfilled unique roles, and they both made a key difference to the outcome as the heist could not have succeeded without either person. What is notable is that Constance's contribution was easily replaceable by many other pickpockets, while Nine Ball's contribution was much harder to replace. By that reasoning, Nine Ball is more responsible for the team's success. In the following, we will review prior work on responsibility attributions falling under each of these conceptualizations of contribution: value, difference-making, and replaceability.

Responsibility and value. One way that people allocate responsibility in groups is in proportion to the amount of some units put into achieving the outcome, such as points scored or time spent. This is especially intuitive for collaborative efforts such as 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; 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" 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 highlighting the intuitive mapping between contributed value and proportioned responsibility.

Responsibility may be affected not only by the (perceived) value of contribution, but also by how that value compares to expectation. Exceeding expectations results in more responsibility, but only when the actions reveal something potentially positive about the person (Gerstenberg et al., 2018; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021). For example, people assign more credit to a soccer goalie who saves a more unexpected shot, as that is diagnostic of the goalie's skill, but not to a game show contestant who happens to win in a harder game of pure chance. Expectations may come from prior knowledge about the specific person and group situation, from different norms, or from oneself (Simpson, Alicke, Gordon, & Rose, 2020). For example, in the law, one comparison standard is what the "reasonable person" would have done (Tobia, 2018). In sports, expectations are sometimes based on the average league player. The Wins Above Replacement (WAR) metric commonly used to rank baseball players measures a player's value in terms of how many wins he contributes to his team over any possible replacement-level player in his position (Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, & Tenenbaum, 2014).

Responsibility and difference-making. The notion of value falls short in situations where contributions are incommensurable or not straightforwardly "additive" towards the group outcome. In some cases, multiple agents can all be fully responsible for the same outcome (Kaiserman, 2021). In others, multiple agents can contribute the same value, but still be held responsible to different degrees. For example, each committee member from earlier contributed one vote, which has the same value as any other member's vote. However, the majority voters for a marginal win (policy B) seem more responsible for the outcome than the majority voters for a landslide win (policy A).

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. In the above example, the six majority voters for policy B are pivotal because had any of them voted against the policy, it would not have won (i.e. a 5-6 loss). In contrast, the ten majority voters for policy A are farther 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 then become pivotal.

Another notion that complements pivotality and captures additional variation in responsibility is *criticality.* The more important someone is perceived to be for the outcome by virtue of their role or the structure of the situation, the more responsible they are (Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Langenhoff et al., 2021; Zultan, Gerstenberg, & Lagnado, 2012). For instance, in a bystander situation, everyone is pivotal because any one person can intervene to change the outcome, regardless of how many people are 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. When multiple causes contribute the same value, pivotality and criticality captures how responsibility can still vary: the contributions can structurally make more or less of a difference to the outcome.

Some models of responsibility describe difference-making in terms of how much an action or event contributed towards the probability of the outcome, rather than the outcome itself. The more someone's action increased the perceived likelihood that the outcome would happen, the more responsibility is attributed to that person (Brewer, 1977; Fincham & Jaspars, 1983; Parker, Paul, & Reinholtz, 2020; Spellman, 1997). Intuitively, 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. Accordingly, the easier it is to imagine that a person could have acted differently, the more responsible they are (Malle, Guglielmo, & Monroe, 2014; Weiner, 1993; Wells & Gavanski, 1989). Petrocelli, Percy, Sherman, and Tormala (2011) capture this in their *counterfactual potency* framework, which predicts responsibility from the product of two independent quantities: if-likelihood and then-likelihood. Consider the following counterfactual statement: "IF another hacker had been recruited instead of Nine Ball, THEN the heist would have failed." This counterfactual is potent to the extent that the if-likelihood is high (i.e. it is easy to imagine that another hacker could have been recruited), and independently to the extent that the then-likelihood is high (i.e. it is plausible that the heist would have failed 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 there never was 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.

Responsibility and replaceability. What about situations like the heist in *Ocean's 8*, in which multiple causes contributed to the outcome, but there is no suitable measure of value and no structural asymmetries? In this paper, we propose a third way of thinking about differences in contribution, which is how easily the contribution could have

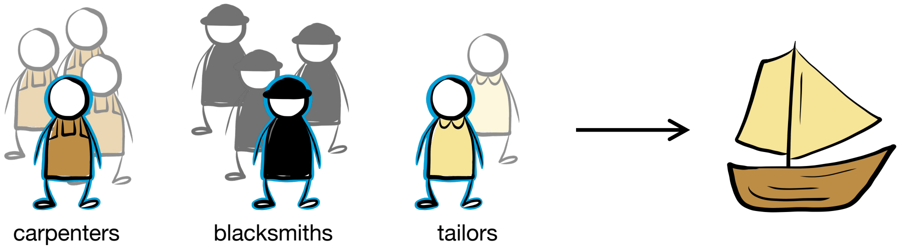

Figure 1. One craftsperson of each type (highlighted in blue) helped build the ship. Here, there were three other carpenters, three other blacksmiths, and one other tailor who could have been potential replacements. Between trials, we varied the number of possible replacements of each type.

been replaced. The easier it would have been to replace someone's contribution, the less responsible that person is held.

Prior studies on responsibility in groups have alluded to the notion of replacement. In situations where one agent made decisions and another implemented them, people hold the decider as more responsible than the implementer, possibly because they view the implementer as more 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; Kerr, 1996). The larger the group is, the more potential replacements there are, and generally the stronger these effects are observed to be. 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. The explanation is that, 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 present in the market, the more likely someone else will buy or sell instead, and thus the less responsibility felt.

Overview of experimental paradigm

In this paper, we explore how replaceability affects responsibility attributions. We develop a paradigm that allows us to quantitatively manipulate information about replaceability and model its influence on responsibility judgments. 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 at the moment, but as long as there is (at least) one craftsperson of each type who can, then a ship will be successfully built. In this example, the village has four carpenters, four blacksmiths, and two tailors, and the ship was a success. We would like to understand how responsible each of the three craftspeople who helped are for the outcome.

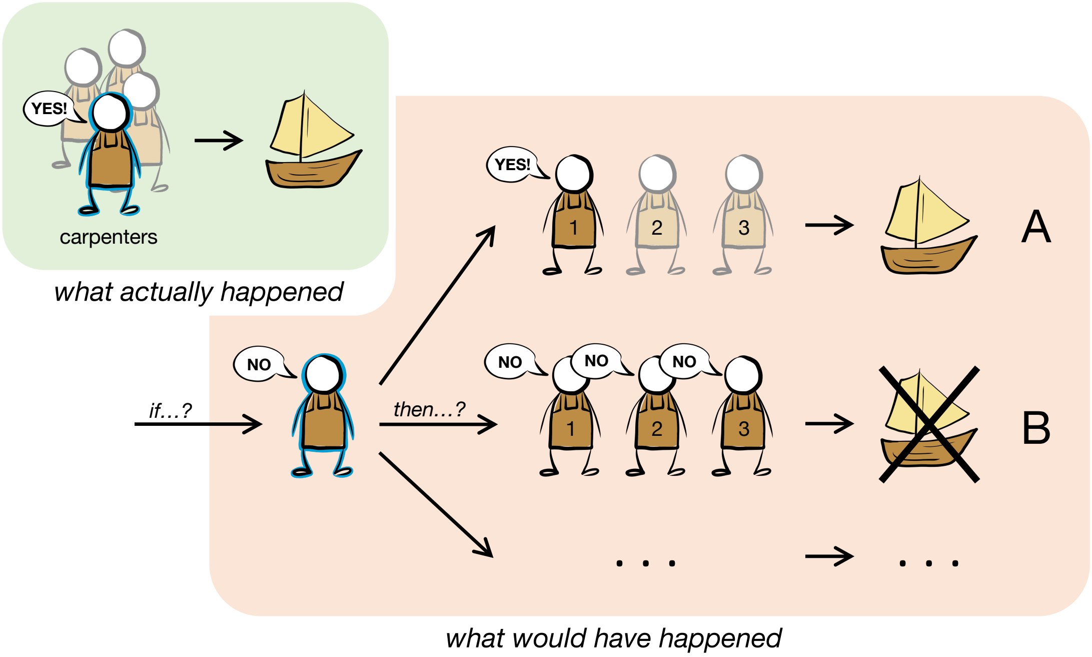

Figure 2. Schematic diagram of the model. The model predicts the responsibility attributed to the carpenter highlighted in red for the successful ship by considering what would have happened if that carpenter had said "no". Two possible counterfactual scenarios are shown. In scenario A, the first of the three other carpenters is available, so the ship would still have been a success. In scenario B, none of the other carpenters are available as a replacement, so the ship would have been a fail. The model computes responsibility by enumerating and computing the probability of a successful replacement.

Our model predicts that the easier it would have been to replace someone who contributed, the less responsible that person is held. Despite all three craftspeople making equally critical contributions, the carpenter and the blacksmith seem 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) to build the ship. Similarly, in *Ocean's 8*, Nine Ball seems more responsible for the successful heist than Constance because she had fewer counterfactual replacements.

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 our model can predict responsibility from replaceability more generally. In Experiment 3, we test more complex scenarios and identify specific factors that affect replaceability and, in turn, responsibility. In each experiment, we additionally test how people's responsibility judgments may differ between social and physical contexts.

Counterfactual Replacement Model (CRM)

We present the *Counterfactual Replacement Model* (CRM) of how people assign responsibility to individuals for group efforts in situations where contributions differ by how easily they could have been replaced. The CRM predicts that a person will be held more responsible for the outcome the lower the probability was that a counterfactual replacement could have been made. When it seems that a replacement would have almost inevitably made the outcome come about anyway, like in crowded markets, people can largely deny personal responsibility as predicted by the CRM. On the other extreme, people often attribute great responsibility to individuals just by virtue of them being rare or special in the sense of being irreplaceable. We will illustrate a concrete implementation of the model with an example. Consider the carpenters from Figure 1. The CRM predicts how responsible the carpenter who actually helped is for the successful ship by simulating what would have happened if that carpenter had been busy and said "no" to helping. 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 a success. If they had also said "no", then the next person would have been asked, until all the possible replacements were exhausted (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.

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 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 been a failure 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 successful replacement to be:

$$\text{probability of counterfactual replacement} = 1 - \prod_{i=1}^{n} 1 - p_i. \tag{1}$$

The CRM predicts that the higher the value of this term, the lower the responsibility attributed to the cause that would have been replaced. This term 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" (high values of p_i), then the more likely a successful counterfactual replacement would have been made. Accordingly, the easier it would have been to replace someone, the less responsible that person is for the group outcome. In the following experiments, we concretely test the CRM by manipulating n and p_i and measuring responsibility judgments.

Experiment 1: Number of replacements

The goal of Experiment 1 was to investigate the effect of the number of possible replacements n on responsibility. 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 always the same. We predicted that the more replacements there were, the lower the responsibility attributed according to Equation 1.

We also tested how people's reasoning about the responsibility and replaceability of physical objects may differ from that of agents. 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. 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. Our key manipulation was the number of possible replacement gears of each type.

Methods

Participants. The task was preregistered and posted as an online study on Prolific, a crowd-sourcing research platform. 101 participants (*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) were recruited and compensated at a rate of $11/hour. One was excluded for failing an attention check, leaving a final sample size of $N = 100$. 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, one of which was used as our exclusion criterion (trial X in Table A1 in the Appendix). The order of the test trials was randomized for each participant except that the attention check trial was always inserted as the 10$^{\text{th}}$ trial.

In each trial, participants were shown the three contributions and the number of possible replacements for each one. 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). After the last trial, they were given the option to share demographic information and comments about the experiment.

In every trial, the three contributions 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 three numbers from zero to three (see Table A1 for details). For instance, Figure 1 shows a situation in which two contributors each have three possible replacements and the third contributor has one, although the exact ordering of the three groups here (carpenters, blacksmiths, and tailors) is arbitrary. We randomized the permutation of the three numbers across trials so there was no sense that there were systematically more carpenters or more yellow gears, for example. The average time to complete the experiment was 9.8 minutes (SD = 5.6).

Model fitting. We fit two Bayesian mixed effects models to predict participants' responsibility judgments, one using the probability of replacement as a predictor (the CRM)

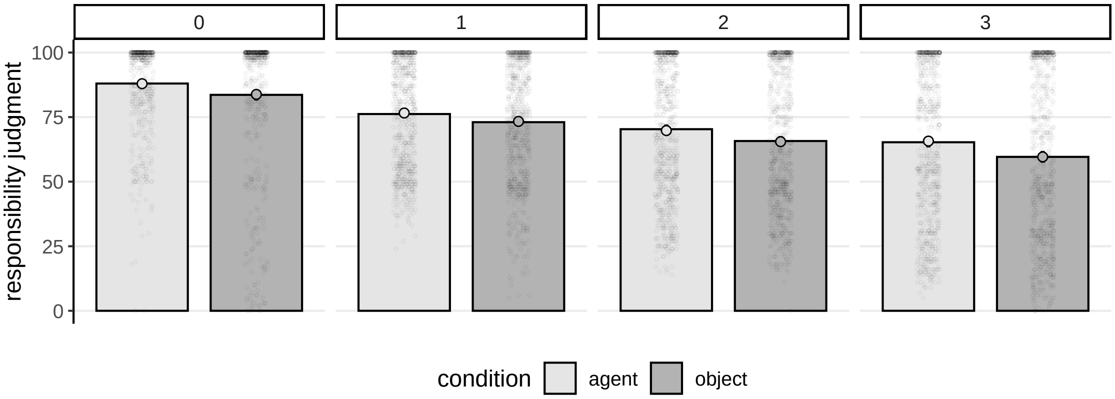

Figure 3. Mean responsibility judgments for the agent (light gray, left bar) and object (dark gray, right bar) as a function of the number of replacements in Experiment 1. Small points indicate individual judgments and the circle represents the predictions of the CRM. Error bars are bootstrapped 95% confidence intervals.

and one with only an intercept (the "baseline" model). Both models also include random intercepts and the CRM has an additional random slope for each participant. 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, we assumed a uniform probability p that each possible replacement would have succeeded, and fit p to the data to minimize the squared error between model predictions and participants' judgments in each condition. Then, given n possible replacements, Equation 1 becomes

$$\text{probability of counterfactual replacement} = 1 - (1 - p)^n, \tag{2}$$

which was then used as the predictor in the CRM.

Results

Figure 3 shows participants' mean responsibility judgments as a function of the number of possible replacements. The more replacements there were for a particular contribu-

Table 1
Posterior means and 95% highest density intervals for the intercept and probability of replacement predictor in the CRM fit in Experiment 1. The model is specified by `judgment ~ 1 + probability of replacement + (1 + probability of replacement | participant)` *with Equation 2. We adopt the convention of calling an effect significant if the 95% highest density interval (HDI) of the estimated parameter in the Bayesian model excludes 0. The results show that the probability of replacement is a significant predictor of participants' responsibility judgments in both conditions.*

condition	intercept	probability of replacement
agent	87.93 [83.83, 92.04]	-28.35 [-38.10, -18.76]
object	85.72 [75.71, 91.71]	-41.86 [-64.34, -19.90]

tion, 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.

Agent condition. Participants' responsibility judgments in this condition were well-captured by the CRM with a correlation of $r = 0.99$ and RMSE = 1.40, far better than the baseline model. The best-fitting value for the uniformly assumed probability that a potential replacement craftsperson would have helped was $p = 0.4$ (see Figure A1 for details). The probability of replacement predictor, computed using this p and Equation 2, was significant (see Table 1). To evaluate the CRM against the baseline model, we ran an approximate leave-one-out cross-validation comparison as well as measured how many individual participants in each condition were best fit by either model. Table 2 summarizes this comparison. The results show that the CRM both accounts best for the overall data despite having additional parameters, and also best captures a majority of 35 out of 50 individual response patterns.

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 = 2.39. The best-fitting value for the probability that a possible replacement gear would have succeeded was $p = 0.25$ (see Figure A1), and the probability of replacement predictor here was also significant (see Table 1). Table 2 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 indicate that participants' responsibility judgments were well explained by the probability of counterfactual replacement, despite the actual contributions and outcome being the same in every trial. The CRM predicts that the more likely a contribution could have been successfully replaced, the lower the responsibility attributed to it. If we assume a uniform probability that a replacement would have succeeded, then the CRM predicts the negative relationship between responsibility judgments and the number of possible replacements observed here. A model of responsibility based on how much value each craftsperson or gear added, or how much of a difference each one made, cannot

Table 2

Results of model comparison in Experiment 1. The baseline model is specified by `judgment ~ 1` *and the CRM by* `judgment ~ 1 + probability of replacement`*. 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.*

condition	model	r	RMSE	Δelpd (se)	n best
agent	CRM	0.99	1.40	0 (0)	35
	baseline	—	8.24	-1591.2 (74.2)	15
object	CRM	0.96	2.39	0 (0)	44
	baseline	—	8.96	-1777.7 (80.2)	6

explain these results because each one's contribution was equally important and pivotal. We also controlled for systematic biases between the three types of craftspeople by randomly permuting each trial, so the results cannot be explained by beliefs that, for example, ships require the most work from carpenters and thus the carpenter was always the most responsible.

On the aggregate level, we found that participants tended to hold a particular agent or object responsible to the extent that they had fewer replacements. On the individual level, we found two distinct clusters of responses. A majority of participants tended to attribute less responsibility to more replaceable contributions, thus driving mean judgments, but a minority were not sensitive to the number of replacements at all. Participants' comments about what factors influenced their reasoning 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 baseline 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."). The judgments of this group was better predicted by additionally considering the replacements.

While many participants explicitly mentioned replacement in their comments, it's possible that some of those best fit by the CRM actually used a different reasoning strategy. They might have used the general number of craftspeople or gears present to make an inference about some feature of the specific member that actually helped, which then affected their judgments, and this mapping may be wholly independent of counterfactual replacements. Because we fit a uniform p and only varied n, it is difficult to tease the predictions of the CRM apart from simpler models that explain responsibility only in terms of the number of replacements in a more heuristic manner. Thus, in Experiment 2, we manipulated both p and n to provide a more direct test of the CRM and rule out alternative explanations.

Experiment 2: Availability of replacements

In Experiment 1, we varied the number of replacements n and found that it significantly influenced responsibility judgments. If participants are really reasoning about counterfactual replacements, however, then they should be sensitive not to the number of replacements per se, but to the general *likelihood of finding* a replacement. This increases with the number of replacements in the absence of any other information, but also depends on other factors, such as the probability p that a particular replacement is actually available, as outlined in Equation 1. For instance, people should judge a carpenter with an unavailable replacement to be more responsible than one with a readily available replacement, even though both have exactly one replacement, because in a counterfactual scenario the unavailable replacement would have been unlikely to actually step in to help build the ship. More available contributors have higher values of p and less available ones have lower values of p. In this experiment, we incorporated exactly such a notion of availability.

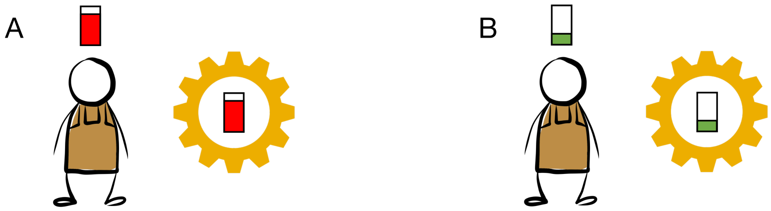

Figure 4. Availability indicators shown in Experiment 2. (A) A craftsperson or gear with a fuller red bar is more likely to be busy or broken (respectively), and hence has *low* availability. (B) A craftsperson or gear with a more empty green bar is less likely to be busy or broken (respectively), and hence has *high* availability.

Methods

Participants. The experiment was preregistered and posted on Prolific. $N = 100$ participants (*age*: M = 25, SD = 6; *gender*: 40 female, 58 male, 2 non-binary; *race*: 58 White, 7 Black, 5 Asian, 2 American Indian/Alaska Native, 2 Multiracial, 25 undisclosed) 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.

Procedure & design. The procedure and design were the same as that of Experiment 1, except that we additionally introduced the availability of each replacement (see Figure 4). 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 corresponded to its probability of not being broken and thus available to be 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.

Across 20 trials, we varied the number of replacements (0 through 4) and the availability of each one (low or high, see Table B1 for details). There are 15 such combinations. Each trial featured three combinations—one for each contribution—and each combination appeared in at least two different trials. The combinations were distributed among the trials so that there were no more than 12 total agents or objects in any one trial, in order to avoid cluttering the display (so for example, there was no trial in which the carpenter, blacksmith, and tailor each had four replacements as that would have been too visually overwhelming). Participants took an average of 12.3 minutes (SD = 6.5) to complete the experiment.

Model fitting. We fit two versions of the CRM to predict participants' responsibility judgments, one assuming a uniform probability of success p for any replacement as in Experiment 1 (the CRM$_{\text{uniform}}$), and the other assuming two different probabilities p_{low} and p_{high} for replacements with either low or high availability respectively (the CRM$_{\text{full}}$). In the uniform model, the predictor was computed using Equation 2, similarly as in Experiment 1. In the full model, it was computed as

$$\text{probability of counterfactual replacement} = 1 - (1 - p_{\text{low}})^{n_{\text{low}}}(1 - p_{\text{high}})^{n_{\text{high}}} \qquad (3)$$

where n_{low} is the number of replacements having low availability and n_{high} is the number with high availability. The parameters p, p_{low}, and p_{high} were all optimized to minimize the

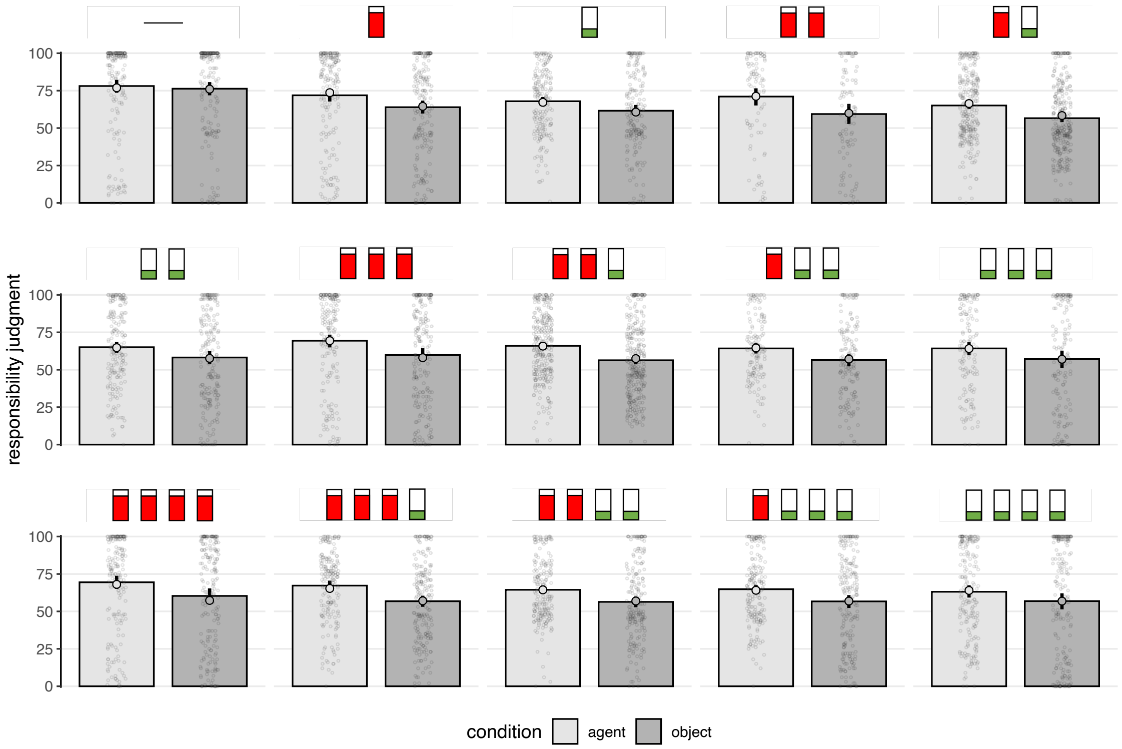

Figure 5. Mean responsibility judgments for the agent (light gray, left bar) and object (dark gray, right bar) conditions for each set of possible replacements in Experiment 2. There could be up to four possible replacements, each of which had either high or low availability. The sets are ordered by increasing probability of replacement ($\propto$ increasing number and availability). Small points indicate individual judgments and the circle represents the predictions of the CRM$_{\text{full}}$. Error bars are bootstrapped 95% confidence intervals.

squared error between the respective model predictions and participants' judgments in each condition. In addition to the two CRMs, we also fit a third Bayesian mixed effects model that only included an intercept (the "baseline" model) for reference.

Results

Figure 5 shows participants' mean responsibility judgments across all combinations of possible replacements that we tested. They are sorted in order of increasing number and availability, corresponding to increasing probability of counterfactual replacement and decreasing responsibility. The results 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 results from Experiment 1. However, even for a fixed number of replacements, the more available they were individually, the more responsible participants rated that contribution. For example, participants judged a craftsperson with four replacements who

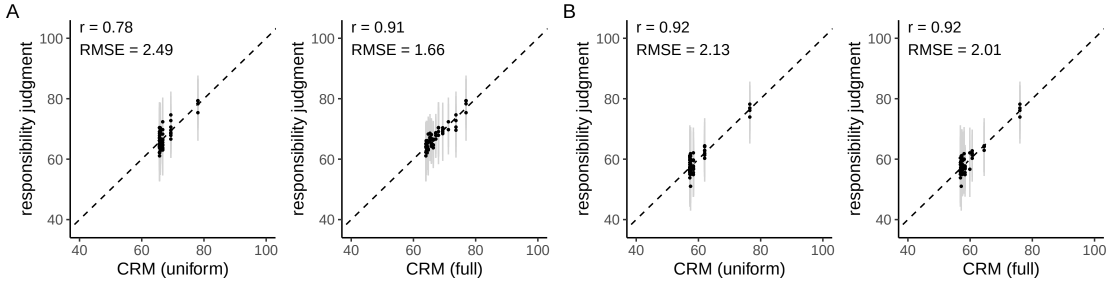

Figure 6. Scatter plots showing the relationship between mean responsibility judgments and the uniform and full version of the CRM in Experiment 2, in the (A) agent and (B) object conditions. r = Pearson correlation coefficient, RMSE = root mean squared error. Error bars are bootstrapped 95% confidence intervals.

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 (mean 69.5, 95% CI [65.1, 773.8]). We discuss the results from each condition in turn.

Agent condition. Participants' judgments in the agent condition were well captured by the CRM_{full} with a correlation of $r = 0.91$ and RMSE = 1.66, better than the $\text{CRM}_{\text{uniform}}$ and baseline models. The best-fitting values for the parameters representing how likely a replacement craftsperson would have helped were $p = 0.7$ used in the $\text{CRM}_{\text{uniform}}$, and $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.75$ used in the CRM_{full} (see Figures B1 and B2 for parameter search details). The probability of replacement predictor in the full model was significant (see Table 3). Figure 6 illustrates model predictions compared to mean judgments across all trials, and Table 4 summarizes a model comparison based on leave-one-out cross-validation as well as individual participant best fit. 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 individuals. The full model accounts best for participants' judgments overall, both quantitatively and qualitatively (see Figure 6), and also captures a majority of 35 individual response patterns.

Object condition. The results in the object condition were similar to those in the agent condition. The best-fitting values for the parameters defining how likely a replacement

Table 3
Posterior means and 95% highest density intervals for the intercept and probability of replacement predictor in the CRM_{full} in Experiment 2. The model is specified by `responsibility ~ 1 + probability of replacement + (1 + probability of replacement | participant)` *with Equation 3. We adopt the convention of calling an effect significant if the 95% highest density interval (HDI) of the estimated parameter in the Bayesian model excludes 0. The results show that the probability of replacement is a significant predictor of participants' responsibility judgments in both conditions.*

condition	intercept	probability of replacement
agent	76.91 [66.32, 87.33]	-12.96 [-24.23, -1.44]
object	76.01 [66.18, 85.59]	-19.17 [-30.55, -7.88]

gear would have succeeded were $p = 0.75$ used in the CRM$_{\text{uniform}}$, and $p_{\text{low}} = 0.6$ and $p_{\text{high}} = 0.8$ used in the CRM$_{\text{full}}$ (see Figures B1 and B2 for parameter search details). Here, the probability of replacement predictor in the full model was also significant (see Table 3). Table 2 summarizes and compares all three model fits in this condition. While both the CRM$_{\text{uniform}}$ and CRM$_{\text{full}}$ make predictions that correlate highly with participants' judgments ($r = 0.92$), the full model outperforms the other two models on cross-validation. It also best explains 28 out of 50 individual participants, compared to the uniform model which best explains 12 participants and the baseline model which best explains 10.

Discussion

In this experiment, we tested a more comprehensive version of the CRM. We extended Experiment 1 by varying the probability that a particular replacement would have succeeded (p_{low} and p_{high}), in addition to the number of replacements (n). A potential replacement's individual ability dictates how likely they could have actually succeeded in a counterfactual scenario, which then influences how responsible the actual contributor may be held. The results show that participants' judgments were sensitive to both varying availability and number, and that they were best explained by a full version of the CRM that allows for flexibly fitting both of these components. Importantly, the better performance of the full CRM over one that makes uniform assumptions about availability demonstrates that the model cannot be reduced to a simple dependence of responsibility on the total number of agents or objects present in the scene. Nor can these results be explained by a model that only considers the contributions themselves or the outcome, 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 re-

Table 4
Results of model comparison in Experiment 2. The baseline is specified by `judgment ~ 1` *and the CRM models by* `judgment ~ 1 + probability of replacement`*, where probability of replacement assumes 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.*

condition	model	r	RMSE	Δelpd (se)	n best
agent	CRM$_{\text{full}}$	0.91	1.66	0 (0)	35
	CRM$_{\text{uniform}}$	0.78	2.49	-411.6 (35.9)	4
	baseline	—	4.11	-586.6 (40.2)	11
object	CRM$_{\text{full}}$	0.92	2.01	0 (0)	28
	CRM$_{\text{uniform}}$	0.92	2.13	-62.0 (6.8)	12
	baseline	—	4.11	-307.6 (24.8)	10

placements (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 it was to find an available replacement for a particular contribution—which increased independently the more replacements there were and the more individually available each one was—the less responsible participants tended to hold that contribution for the outcome. 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 have still 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, like the replacements. This would have had an opposite influence on responsibility judgments. In Experiment 3, we tried to tease these possibilities apart.

Experiment 3: Prior availability of contributor

The goal of Experiment 3 was to investigate how the prior availability of the contributor might affect 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 influencing features of the contribution 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 can also be affected by characteristics of the person themselves, such as how feasible their actions were for them (e.g. Malle et al., 2014).

The prior availability of the contributor maps onto if-likelihood in the counterfactual potency model (Petrocelli et al., 2011). The more available a contributor was, the less likely a counterfactual replacement might have been needed, and thus the less potent the counterfactual scenario in which a replacement was made and the outcome failed. Accordingly, the potency model predicts that the contributor would be held less responsible. 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’s implausible that they would have done so, whereas the counterfactual in which the hesitant killer had not acted has high if-likelihood because it is easy to imagine them doing so. Potency predicts that the intentional killer would therefore be less responsible than the hesitant one, but intuitively the intentional killer seems more to blame. 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

Participants. Participants were Stanford undergraduates who were granted 0.5 credit hours for completing the experiment online. 102 students were recruited (*age*: M = 20, SD = 1; *gender*: 56 female, 43 male, 1 undisclosed; *race*: 34 White, 7 Black, 44 Asian, 1 American Indian/Alaska Native, 6 Multiracial, 8 undisclosed). Two were excluded for submitting multiple times, leaving a final sample size of $N = 100$. 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 and gears, including the ones who actually helped and all of their potential replacements. To prevent each scene from appearing too visually overwhelming, we used two contributions in each trial instead of three. Furthermore, to isolate the influence of the contribution versus the replacements on responsibility judgments, in each trial we paired combinations in which the two groups always had the same number of replacements and differed only in the availability of the contributor, or only in the availability of the replacements (see Table C1 for details). For instance, in trial 3 in the agent condition, both the carpenter and tailor who helped have prior 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 19 trials varying the availability of the contributor (low or high), the number of replacements (0 through 3), and the availability of each replacement (low or high). There are 20 such combinations. Each trial contrasted two combinations with the same number of replacements—one for each contribution—so each combination appeared in two different trials (with the exception of the case with zero replacements). Participants took an average of 7 minutes (SD = 2.8)[1] to complete the experiment.

Model fitting. We fit three Bayesian mixed effects models to predict participants responsibility judgments. The first model (the CRM) included an intercept and two fixed effects: the prior probability of the contribution— $p_{\text{contributor}}$—which was either p_{low} or p_{high} depending on it's availability, and the probability of replacement calculated using Equation 3. The second model (the "CP" model) included an intercept and counterfactual 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 replacements would have been available instead (i.e. the complement of the probability of replacement). Thus, potency in this paradigm is

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

[1] For reporting this time, we excluded one outlier participant who took 10.5 hours to complete the experiment (but included their data otherwise).

The parameters p_{low} and p_{high} were fit to minimize squared error between model predictions and participants' judgments in each condition. Finally, we also fit a third model that only included an intercept (the "baseline" model) for reference.

Results

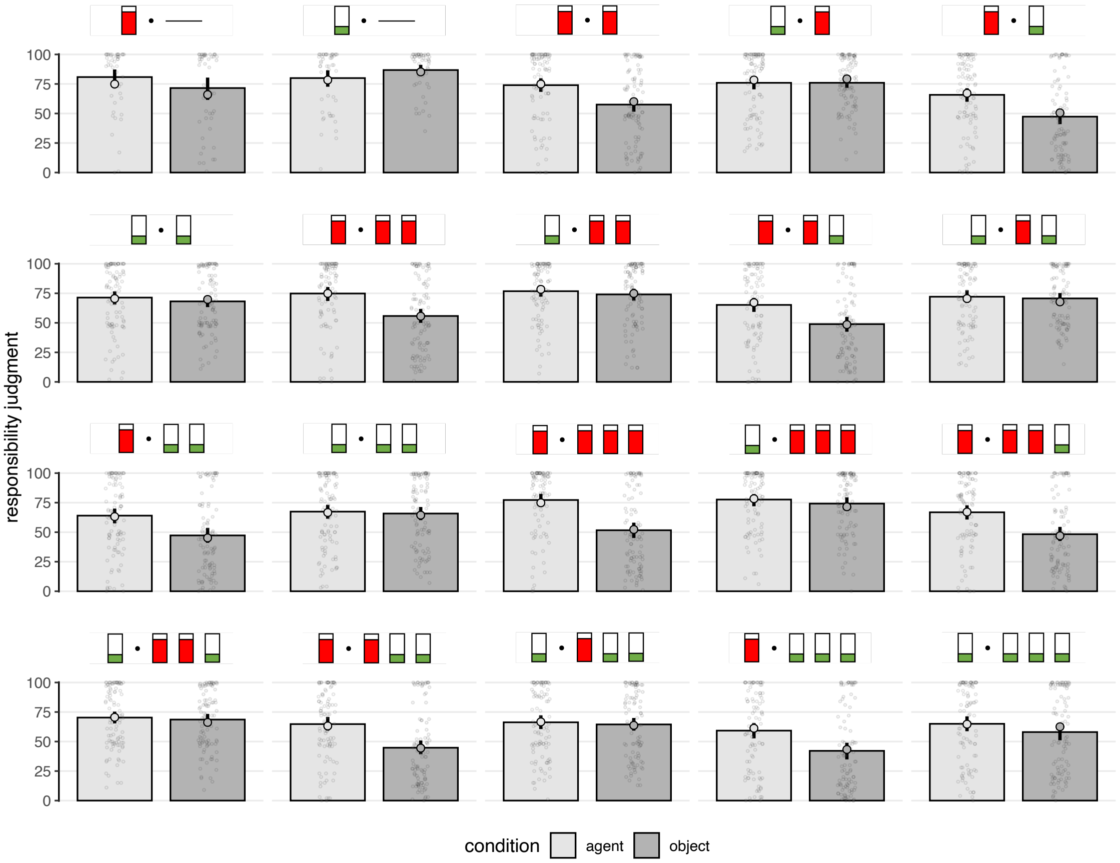

Figure 7. Mean responsibility judgments in the agent (light gray, left bar) and object (dark gray, right bar) conditions in Experiment 3. The actual contributor in each group had either low or high availability (indicated on the left of the dot) and there were up to three possible replacements, each of which also had low or high availability (shown on the right of the dot). The sets are ordered by increasing availability and number of replacements. Small points indicate individual judgments and the circle represents the predictions of the CRM-based model. Error bars are bootstrapped 95% confidence intervals.

Figure 7 shows participants' mean responsibility judgments across the 20 different combinations we tested. They are ordered by increasing availability of the contributor, number of possible replacements, and availability of the replacements. We found two main trends across both conditions. First, the more replacements there were for a particular contribution and the more available those replacements were, the less responsible participants

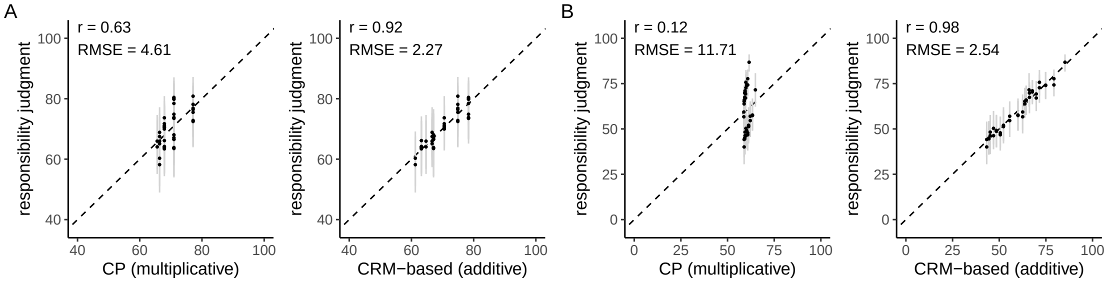

Figure 8. Scatter plots showing the relationship between mean responsibility judgments and the CP and CRM-based models in Experiment 3, in the (A) agent and (B) object conditions. r = Pearson correlation coefficient, RMSE = root mean squared error. Error bars are bootstrapped 95% confidence intervals.

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. Participants' judgments in the agent condition were well-captured by the CRM-based model with a correlation of $r = 0.92$ and RMSE = 2.27, outperforming the CP and baseline models. The best-fitting values for the parameters representing how likely a replacement craftsperson would have helped were $p_{\text{low}} = 0$ and $p_{\text{high}} = 0.5$ (see Figure C1 for parameter search details). In the CRM-based model, the probability of replacement was a significant predictor, but not the availability of the contributor, as the 95% highest density interval includes zero (see Table 5). Figure 8 illustrates model predictions compared to mean judgments across all trials, and Table 6 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-based model, on the other hand, best accounts for the overall data despite having more parameters, and also best captures a majority of 28 out of 50 individuals.

Table 5

Posterior means and 95% highest density intervals for the fixed effects in the CRM fit in Experiment 3. The model is specified by `judgment ~ 1 + contributor + probability of replacement + (1 + contributor + probability of replacement | participant)`*, where contributor was either p_{low} or p_{high}, and probability of replacement was computed with Equation 3. We adopt the convention of calling an effect significant if the 95% highest density interval (HDI) of the estimated parameter in the Bayesian model excludes 0. The results show that all the effects were significant except for the contributor in the agent condition.*

condition	intercept	contributor	probability of replacement
agent	74.91 [67.96, 81.84]	6.88 [-8.42, 22.41]	-15.62 [-22.09, -9.18]
object	53.95 [43.53, 64.68]	48.14 [32.67, 63.10]	-23.71 [-33.76, -14.13]

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-based model with a correlation of $r = 0.98$ and RMSE = 2.54, far better than the CP and baseline models. The best-fitting values for the parameters defining the probability that a replacement gear would have succeeded were $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.65$ (see Figure C1 for parameter search details). Here, both the availability of the contributor and the probability of replacement were significant predictors in the CRM-based model (see Table 5). The results of the model comparison in Table 6 show that the CRM-based model best explains judgments overall, despite having the most parameters, and also best captures a majority of 37 out of 50 individual participants. The CP model best explains 2 participants and the baseline model best explains 11 participants.

Discussion

In this experiment, extending the same paradigm, we found that responsibility judgments were well-predicted by a combination of both the prior availability of the contributor and the probability of counterfactual replacement. The more likely the contribution would have been made in the first place, the more responsible participants tended to hold it. Independently, the more likely a replacement would have been successful, the less responsible participants tended to hold them. In the agent condition, the contributor predictor in the CRM-based model was positive but not significantly so. A craftsperson's availability prior to helping can be suggestive of their obligations or reasons for helping, which can influence their responsibility (Malle et al., 2014). However, we did not find a strong effect here. When analyzing the models fit to individual participants for whom the CRM-based model was the best fit, we found a large range of posterior means for the contributor predictor. Some participants placed little weight on this term, while others had strongly positive or

Table 6
Results of model comparison in Experiment 3. The baseline model is specified by `judgment ~ 1`*, and the CP model by* `judgment ~ 1 + potency`*, where potency was calculated using Equation 4. The CRM model is specified by* `judgment ~ 1 + contributor + probability of replacement`*, where contributor was either p_{low} or p_{high} and probability of replacement was calculated using 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 (Vehtari et al., 2017). "n best" is the number of participants whose judgments were best predicted by each model.*

condition	model	r	RMSE	Δelpd (se)	n best
agent	CRM-based (additive)	0.92	2.27	0 (0)	28
	CP (multiplicative)	0.63	4.61	-579.6 (55.9)	9
	baseline	—	8.24	-793.4 (61.9)	13
object	CRM-based (additive)	0.98	2.54	0 (0)	37
	CP (multiplicative)	0.12	11.71	-611.3 (40.5)	2
	baseline	—	8.96	-710.6 (42.5)	11

strongly negative weights. The wide variation in how much participants cared about the prior availability of the craftsperson who actually helped likely led to the inconclusive effect on overall judgments.

The key difference between the CRM-based model and the CP model is that the former assumes *additive* effects of these two factors, 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 sensitive when a replacement would have been especially needed. The CRM-based model doesn't imply such an interaction, but has more parameters. Despite the additional complexity, it outperforms the CP model, especially in the object condition, largely because the effects of the contributor and replacements emerged in opposite directions (see Table 5).

Why did we find an opposite effect of the contributor here? One possibility is that participants used the information about the prior availability of the contributor to make inferences we did not consider. This was hinted at in their comments about what factors influenced their judgments. A few participants tried to hypothesize why the contribution was more or less likely, or made inferences about the quality of the action because of availability (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."). In addition, the terms "busy" and "brittle" may have been ambiguous and confused with other concepts like effort (e.g. "When both [gears] were green, I feel like they were both doing equal effort. When one was red and the other green, I believed the green was doing more effort. When both were red, I thought it was equal effort."). If busyness or brittleness licensed an inference about effort, then this would explain the inverse relationship between the availability of the contributor and responsibility observed here. Future work is needed to de-confound these possible unexpected interpretations.

General discussion

What underlies people's intuitions of responsibility in groups is a complex question with important applications to our everyday lives. From determining who is at fault after a regretful company decision, to awarding an MVP after a sports team's win, the responsibility allocations that people make have numerous moral, political, and legal consequences. In this paper, we developed the Counterfactual Replacement Model (CRM), a computational model that predicts responsibility in terms of how easily one's contribution could have been replaced (see Figure 2).

The CRM explains responsibility by considering how a group situation would have turned out had a particular contribution not been made. It computes how likely the contribution could have been counterfactually replaced, and predicts that the more likely a successful replacement would have been made, the less responsible the contribution is for the actual outcome. To test the model, we designed an experimental setting where we instantiated and manipulated two parameters: the number of possible replacements (n) and the individual probability that each replacement i would have been able to contribute instead (p_i). These parameters affect the replaceability of the contribution according to Equation 1. Our paradigm allows us to explicitly compute the probability of counterfactual replacement and generate quantitative predictions from the model.

We tested the CRM across three experiments. In Experiment 1, we varied n only and assumed a uniform probability p_i that any replacement would have succeeded. In Experiment 2, we varied both n and p_i by introducing two types of possible replacements with differing success probabilities, p_{low} and p_{high}. This work concretely tested two different values for p_i, but in principle p_i could be distinct for every replacement. In Experiment 3, we tested an extension of the CRM that independently considers both the probability of counterfactual replacement and the contributor's prior probability of success. Across all three experiments the CRM captured participants' judgments well in both the social domain, where the contributions were made by agents, and the physical domain, where the contributions were made by objects.

The parameters p_{low} and p_{high} were fit to participants' judgments in each experimental condition. The CRM then computes the probability of replacement explicitly for each scenario given the fitted values of p_i and n in each trial, using Equation 1. 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 quantity 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 may have been sensitive to the relative difference between these two scenarios, they probably did not have an accurate sense of actual values. Thus, one improvement to the CRM is to use people's subjective estimates of the probability of counterfactual replacement in each trial instead of Equation 1. Our framework still holds because it predicts responsibility as a function of people's beliefs about these probabilities, regardless of how mathematically accurate those beliefs are.

In Experiment 3, the CRM-based model was contrasted with a model based on counterfactual potency (Petrocelli et al., 2011), which fixes an interaction between the effects of the contributor and the replacements. 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. The results suggest that it is not quite borne out of potency. More specifically, we found that the contributor's prior availability and the probability of replacement had opposite effects, contrary to the predictions of the potency model. Participants judged contributors to be more responsible when they were *more* likely to have contributed in the first place, or *less* likely to be successfully replaced. These effects could have arisen from unexpected inferences about factors such as the effort or quality of contribution, which then influenced judgments. In the following sections, we discuss how the CRM relates to prior work on these influencing factors, as well as counterfactual models of causal judgments. We also propose future directions of this work that address some of its limitations, including the problem of counterfactual selection and mental simulation of counterfactual replacements.

Other influences on responsibility

There is a large body of empirical work on responsibility and the many factors that affect responsibility attribution towards agents, including consideration of the agent's mental states like their intentions (Cushman, 2008; Lagnado & Channon, 2008), reasons for acting (Cushman, 2008), skill and capacity for acting (Gerstenberg, Ejova, & Lagnado, 2011; Malle et al., 2014; Weiner & Kukla, 1970), and character as a whole (Gerstenberg et al., 2018; Langenhoff et al., 2021; Uhlmann, Pizarro, & Diermeier, 2015), among other factors that we did not consider here. However, many of these theories are qualitative and do not generate testable predictions or touch on the cognitive process by which people actually reach their judgments. Future work should seek to combine these different findings and frameworks towards a comprehensive account of responsibility. One possible approach is to build on computational models like the CRM and others (e.g. Langenhoff et al., 2021) by developing and integrating formal implementations of the conceptual factors identified above. For instance, Kleiman-Weiner, Gerstenberg, Levine, and Tenenbaum (2015) offer a computational definition of intentions, which could be incorporated.

Counterfactual models of causal judgments

The CRM, which predicts responsibility judgments, shares important elements with counterfactual models of causal judgments. Responsibility and causality are closely related concepts; attributing responsibility requires assessing the causal connection between different contributions and the outcome (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015). Causal judgments, in turn, are intimately linked to counterfactual simulations (Gerstenberg & Tenenbaum, 2017; Pearl, 2000). For example, to explain whether the white cue ball is responsible for the black 8 ball landing in the pocket, people use their intuitive understanding of the physical dynamics at hand to mentally simulate what would have happened to the 8 ball without the cue ball. Gerstenberg, Goodman, Lagnado, and Tenenbaum (2021) developed the counterfactual simulation model (CSM) to capture people's causal judgments in situations like these. The CSM predicts that people compare what actually happened with what they believe would have happened in relevant counterfactual scenarios. The more clear it is that the 8 ball would have been safe if the cue ball had not been there, the more people are predicted to agree that the cue ball caused the 8 ball to be pocketed. The CSM yields quantitative predictions that align closely with participants' causal judgments across a variety of dynamic physical scenarios (Gerstenberg et al., 2021; Gerstenberg & Stephan, 2021; Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021; Zhou, Smith, Tenenbaum, & Gerstenberg, 2022).

The CRM is structurally similar to the CSM in interesting ways. First, both operate over a generative model of the situation at hand. Although the models of the situations look very different—physical events implemented in a computer game engine, compared to group tasks like the one in Figure 1—both are assumed to be represented mentally and used as a basis to form counterfactual scenarios. Secondly, both the CRM and CSM derive relevant counterfactuals by implementing "operators" in the generative model. In the CSM, the primary operator was removing a particular object from the scene. By removing the cue ball, for example, the model creates a situation where one can then evaluate what would have happened to the 8 ball, and compare that to what actually happened. In our paradigm, the operator was removing a particular contribution. This then led to consideration of whether (and how likely) a replacement could have stepped in, which determined whether the counterfactual group outcome would have been the same as the actual outcome. More generally, however, there are many possible counterfactuals that can come to mind. What if the carpenter who helped had used a different type of wood or followed a different blueprint? How would the ship have been different in each of those counterfactual scenarios? We discuss the problem of counterfactual selection next.

Counterfactual selection

A major assumption of our model is that participants make responsibility judgments by reasoning about one specific counterfactual scenario, which is how the outcome would have turned out had a particular contribution been replaced. In many contexts, such as when a particular role must be filled in a heist or a sports game, this is the most relevant counterfactual. But often there are a variety of counterfactual scenarios surrounding an event or an agent's action, and how easily each one comes to mind can be different for different people. The easier is it to imagine a counterfactual that would have changed the outcome, the more responsibility people tend to attribute to the cause in question (Alicke, 2000; Alicke, Davis, Buckingham, & Zell, 2008; Gilbert, Tenney, Holland, & Spellman, 2015; Macrae, Milne, & Griffiths, 1993; Malle et al., 2014; Wells & Gavanski, 1989). In the law, decisions are made by evaluating different counterfactuals as well. In some cases, a defendant's negligent action is considered to be an actual cause of the injury (and thus the defendant should be legally punished, see Summers, 2018) if the injury would not have happened *but for* that action. In other cases, jurors decide whether a defendant should be held liable for an injury by considering 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). Future work should explore what kinds of counterfactuals people spontaneously consider in different contexts, and how that interacts with the idea of replaceability to influences responsibility.

Counterfactual simulation

Another assumption of the CRM is a deterministic relationship between a successful replacement and a successful outcome. That is, we assume that as long as at least one other carpenter had said "yes" to helping, presumably the ship would have still been built. In a more complex situation where, for instance, the quality of the ship also matters, then it becomes relevant not only how likely a replacement carpenter could have been found, but also the exact manner in which the replacement contribution would have been made. The key evaluation is not whether a ship would have still been successfully built, but whether the replacement have done a relatively better or worse job.

In real world situations, people often have additional information and the capacity for mental simulations that enable them to answer such questions. Would a basketball team 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. To predict responsibility judgments in situations like this, a model would need to be able to generate sequences of counterfactual states over time, rather than a binary counterfactual outcome. Future work could investigate machine learning and reinforcement learning algorithms to expand the applications of the CRM from calculations over simple variables, to simulation of continuous action and state sequences.

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 parameters determine the replaceability of a particular contribution. We hope that this work brings us closer towards a comprehensive computational account of responsibility attribution.

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.

Alicke, M. D., Davis, T., Buckingham, J., & Zell, E. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin*, *34*(10), 1371–1381.

Alicke, M. D., Mandel, D. R., Hilton, D., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science*, *10*(6), 790–812.

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

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.

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.

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

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

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

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2263–2268). Austin, TX: Cognitive Science Society.

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.

Gilbert, E. A., Tenney, E. R., Holland, C. R., & Spellman, B. A. (2015). Counterfactuals, control, and causation: Why knowledgeable people get blamed more. *Personality and Social Psychology Bulletin*, *41*(5), 643–658.

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.

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.

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.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128). Austin, TX: Cognitive Science Society.

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., & 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.

Macrae, C. N., Milne, A. B., & Griffiths, R. J. (1993). Counterfactual thinking and the perception of criminal behaviour. *British Journal of Psychology*, *84*(2), 221–226.

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.

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.

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.

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.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2022). Mental Jenga: A counterfactual simulation model of physical support. *PsyArXiv*.

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

Appendix

Appendix A. Experiment 1 additional information

Table A1

Experiment 1 trial info. Trials E1 and E2 are examples, 1 and 2 are practice, 3-21 are test trials, and X is the attention check. We excluded participants if their highest and lowest ratings differed by more than 30 in trial X.

trial	carpenters/ yellow gears	blacksmiths/ green gears	tailors/ blue gears
E1	3	2	3
E2	4	2	2
1	1	3	2
2	4	2	3
3	1	2	1
4	1	1	3
5	1	4	1
6	2	1	2
7	1	2	3
8	1	2	4
9	1	3	3
10	3	4	1
11	4	4	1
12	2	2	2
13	2	3	2
14	4	2	2
15	3	3	2
16	3	4	2
17	4	4	2
18	3	3	3
19	4	3	3
20	4	3	4
21	4	4	4
X	1	1	1

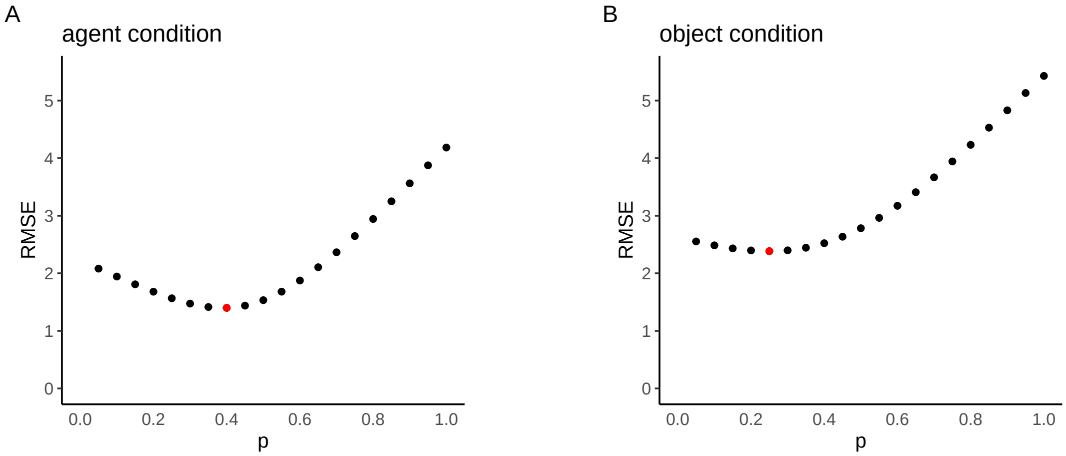

Figure A1. Results of fitting a uniform probability of success parameter in Experiment 1, in the (A) agent and (B) object conditions. The best-fitting values, indicated in red, minimize the squared error between model predictions and participants' judgments. They were $p = 0.4$ in the agent condition, 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. Experiment 2 additional information

Table B1

Experiment 2 trial info. Trials E1 and E2 are examples, 1 and 2 are practice, 3-22 are test trials, and X is the attention check. We excluded participants if their highest and lowest ratings differed by more than 30 in trial X.

trial	carpenters/ yellow gears		blacksmiths/ green gears		tailors/ blue gears	
	n_{high}	n_{low}	n_{high}	n_{low}	n_{high}	n_{low}
E1	3 total		2 total		3 total	
E2	2 total		2 total		4 total	
1	0	0	1	1	1	0
2	0	0	2	0	1	1
3	0	0	0	4	3	1
4	0	0	4	0	1	3
5	0	0	1	2	0	4
6	2	1	1	3	0	0
7	0	4	1	0	2	2
8	1	0	1	3	3	1
9	0	1	4	0	2	2
10	3	0	0	4	1	0
11	0	1	1	2	2	2
12	3	1	0	3	0	1
13	0	3	1	0	2	1
14	3	0	1	2	0	1
15	1	3	3	1	2	0
16	4	0	1	2	2	0
17	2	0	2	2	0	2
18	4	0	1	1	1	1
19	1	1	0	3	1	2
20	2	1	1	2	1	1
21	2	0	0	3	1	1
22	3	0	1	1	0	2
X	0	0	0	0	0	0

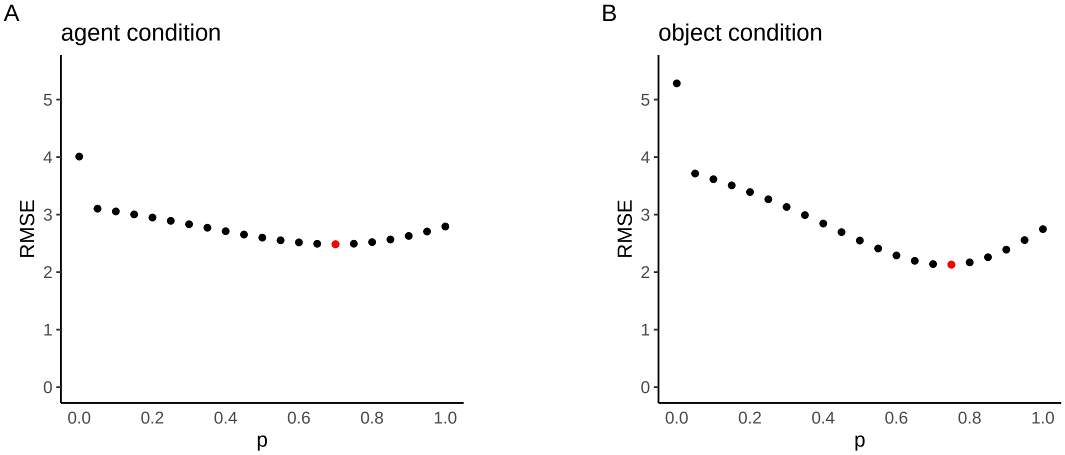

Figure B1. Results of fitting a uniform probability of success parameter in Experiment 2, in the (A) agent and (B) object conditions. The best-fitting values, indicated in red, minimize the squared error between model predictions and participants' judgments. They were $p = 0.7$ in the agent condition, 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.

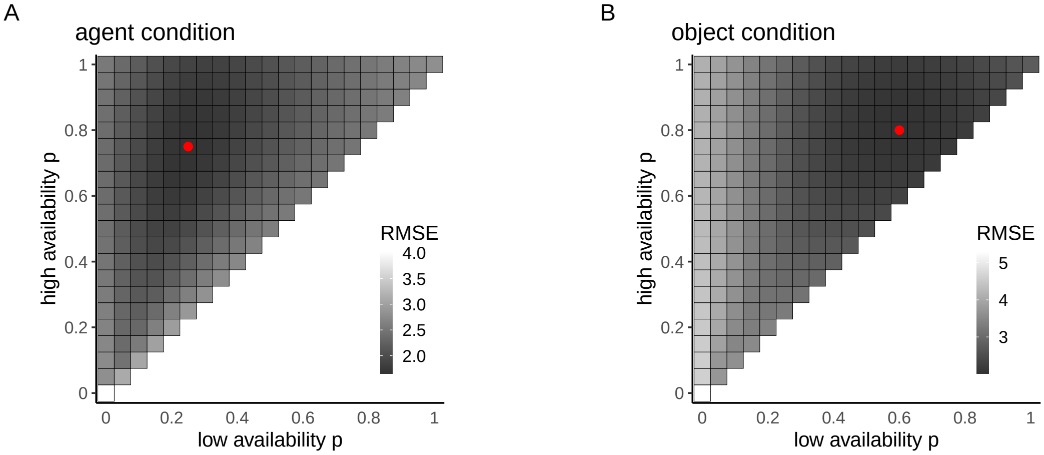

Figure B2. Results of a grid search over the two probability of success parameters for replacements with low and high availability in Experiment 2, in the (A) agent and (B) object conditions. The best-fitting parameters, indicated with the red dots, minimize the squared error between model predictions 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

Experiment 3 trial info. Trials E1 and E2 are examples, 1 and 2 are practice, and 3-21 are test trials.

trial	carpenters / yellow gears			tailors / blue gears		
	contribution	n_{high}	n_{low}	contribution	n_{high}	n_{low}
E1	-	1	0	-	1	1
E2	low	1	1	low	0	2
1	high	2	0	low	0	1
2	low	0	3	low	1	0
3	low	1	0	low	0	1
4	high	0	1	high	1	0
5	low	1	0	high	1	0
6	low	0	1	high	0	1
7	low	2	0	low	1	1
8	low	0	2	low	1	1
9	high	1	1	high	2	0
10	high	1	1	high	0	2
11	low	2	0	high	2	0
12	low	0	2	high	0	2
13	low	3	0	low	1	2
14	low	0	3	low	2	1
15	high	1	2	high	3	0
16	high	2	1	high	0	3
17	low	3	0	high	3	0
18	low	2	1	high	2	1
19	low	1	2	high	1	2
20	low	0	3	high	0	3
21	low	0	0	high	0	0

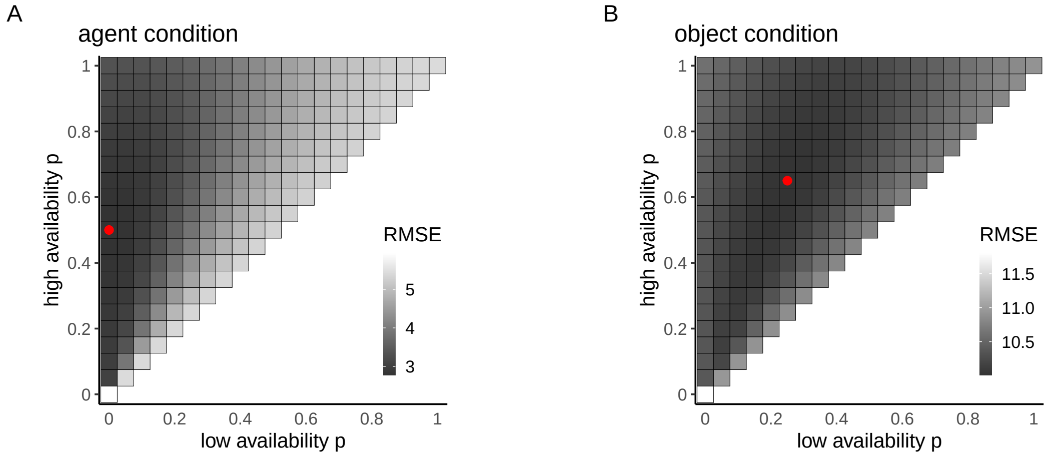

Figure C1. Results of a grid search over the two probability of success parameters for replacements with low and high availability in Experiment 3, in the (A) agent and (B) object conditions. The best-fitting parameters, indicated with the red dots, minimize the squared error between model predictions 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.