Cognition 242 (2024) 105646

Contents lists available at ScienceDirect

Cognition

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

Original articles

If not me, then who? Responsibility and replacement[☆]

Sarah A. Wu, Tobias Gerstenberg [*]

Stanford University, United States of America

ARTICLE INFO

Dataset link: https://github.com/cicl-stanford/responsibility_replacement

Keywords:
Responsibility
Counterfactuals
Social cognition
Mental simulation
Causality

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 Here, we look at how replaceability – the ease with which a person could have been replaced by someone else – affects responsibility.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 differed if the person of interest had been replaced. The model predicts that a person is held more responsible, the more difficult it would have been to replace them. To test the model's predictions, we design a paradigm that quantitatively varies replaceability by manipulating the number of replacements and 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.

1. 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 because she accomplished something fewer people would have been able to do? If Constance had not been there, Debbie and Lou could have easily found another pickpocket to replace her, given the high turnover. However, 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 where 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.

1.1. 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; Cushman, 2008; Lagnado & Channon, 2008; Lombrozo, 2010; Malle, Guglielmo, & Monroe, 2014; Shaver, 1985. Other factors pertain to the agent's causal role in bringing about the outcome (Gerstenberg et al., 2018; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021). For example, agents are generally held more responsible when their actions were pivotal (Gerstenberg & Lagnado, 2010; Lagnado,

[☆] All the data, study materials, and analysis code are available here: https://github.com/cicl-stanford/responsibility_replacement.

[*] Correspondence to: Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, United States of America.
E-mail address: gerstenberg@stanford.edu (T. Gerstenberg).

https://doi.org/10.1016/j.cognition.2023.105646
Received 7 March 2023; Received in revised form 6 October 2023; Accepted 10 October 2023

Gerstenberg, & Zultan, 2013; Zultan, Gerstenberg, & Lagnado, 2012). Here, we focus on the latter category of factors. We study how individuals' causal contributions affect their responsibility for group outcomes. While considering a person's character and mental states is critical for *moral* responsibility (Vincent, 2011), our model captures the *causal* responsibility that an individual carries for a joint outcome. This means that our model applies not only to agents but also to physical objects in situations with the same causal structure.

When multiple causes affect 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 be viewed as more responsible for the team's win (all else being equal). Second, contributions can differ in how much of a *difference* each made to the outcome e.g. Chockler & Halpern, 2004; Lagnado et al., 2013. For example, a citizen's vote is more responsible for a politician's election success when the outcome was close than when it was a landslide win. This intuition cannot be explained in terms of value since each vote counts the same. Finally, contributions can differ in how easily they could have been *replaced*. In *Ocean's 8*, the value of Nine Ball and Constance's contributions cannot easily be compared because they had unique jobs. Furthermore, both agents were pivotal, as the heist would not have succeeded without them. However, the pickpocket Constance's contribution was arguably 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.

1.1.1. 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, Landy, Cushman, Vélez, & Gershman, 2023). 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.

1.1.2. Responsibility and difference-making

The notion of value falls short when contributions are incommensurable or do not combine additively towards 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, agents may receive unequal responsibility even though they contributed the same value. For example, while different countries in the United Nations may have the same number of votes, their voting power differs based on the voting coalitions they tend to form (Felsenthal & Machover, 2004).

Chockler and Halpern (2004) define responsibility 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, 2013. 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 where 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).

Responsibility also depends on how critical one's contributions were perceived 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, responsibility in such situations is known to diffuse: the more (equally pivotal) bystanders are present, the less responsible each 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., casting a vote, or lending a helping hand), the extent to which their vote was critical and pivotal affects their perceived responsibility.

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 action brings about the outcome. For example, people may regard a deciding goal scored in the last minute of a game as more responsible for the team's success than a goal scored early on Henne, Kulesza, Perez, and 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 is conceivable that they could have acted differently. 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 would not 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 would not 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 the counterfactual potency as well.

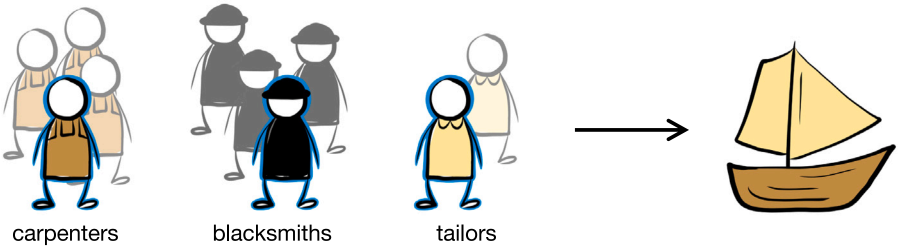

Fig. 1. Illustration of an example trial in the *agent condition*. One craftsperson of each type (highlighted in blue) helped build the ship. Here, there were three other carpenters, three other blacksmiths, and one other tailor who could have been potential replacements. Between trials, we varied the number of possible replacements of each type. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1.1.3. 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 thus deserve more credit. 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 judgments are affected by how a person's contribution compares to expectations about how they 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 norms in different domain, or from simulating what oneself would do (Simpson, Alicke, Gordon, & Rose, 2020). 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 compared to 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 can elicit different responsibility judgments for 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, 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; Johnson, 2003), or to absolve themselves in common goods dilemmas along the lines of "it makes no difference to the outcome what I do" (Glover & Scott-Taggart, 1975; Green, 1991; Hale, 2011; 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.

1.2. Overview of experimental paradigm

In this paper, we explore how replaceability affects responsibility attributions. Imagine a fictional village with three types of craftspeople who build ships together: carpenters, blacksmiths, and tailors (see Fig. 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 who can 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 were not for that particular tailor, the village would have had to rely on the only other tailor who might not have been available either. We focus on positive outcomes in which the ship is always successfully built here, and explore how the model extends to negative outcomes in the General Discussion.

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. Experiment 2 shows that responsibility judgments are also sensitive to how likely a possible replacement would have been available. In Experiment 3, in addition to manipulating the availability of the replacements, we also manipulate how likely the contributor would have needed to be replaced. In each experiment, we test people's responsibility judgments in social and physical contexts.

2. 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 when the probability is lower that a successful counterfactual replacement could have been made. At the extreme, the model predicts that when a successful replacement was very likely, like in crowded markets, very little responsibility is attributed (Falk & Szech, 2013). The notion of replaceability bears some resemblance to Kahneman and Tversky's (1982) simulation heuristic in that they both describe how judgments about present outcomes are affected by the availability or ease of imagining counterfactual alternatives see also Kahneman & Miller, 1986; Phillips & Knobe, 2018; Wells & Gavanski, 1989. The CRM suggests a concrete mechanism of how these counterfactual simulations may play out, and how they affect responsibility judgments. In fact, there are many ways in which a candidate replacement could play out. For example, in 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

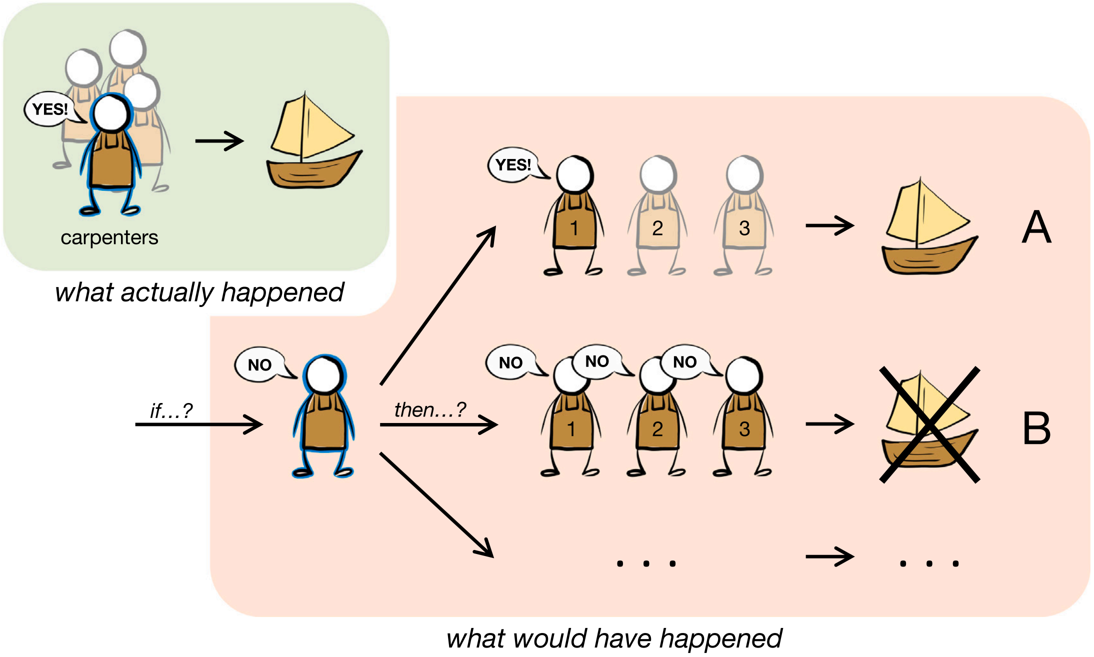

Fig. 2. Schematic diagram of the model. The model predicts the responsibility attributed to the carpenter highlighted in blue for the successful ship by considering what would have happened if that carpenter had said "no". Two possible counterfactual scenarios are shown. In scenario A, the first of the 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 Fig. 1. For example, to determine the extent to which the carpenter is responsible, the CRM simulates what would have happened if the carpenter had been unable to help build the ship. Fig. 2 illustrates a diagram 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 the ship would not have been built (scenario B). By relying on a causal 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 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 p_i represents the probability that replacement i would have stepped in, 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 who actually contributed to the outcome (and who could have been replaced). Conversely, the easier it would have been to replace someone, the less responsible that person is for the group outcome. Replaceability increases with increasing values of n and p_i. The more potential replacements there were (higher values of n) and the more likely those replacements would have said "yes" (higher values of p_i), the more likely a successful counterfactual replacement would have been made.

Note that replaceability is different from causal discounting e.g. Khemlani & Oppenheimer, 2011, in which the observation of one cause leads to a decrease in the belief of another cause. In our case, the actions of the contributor and replacements are mutually exclusive. The contributor always actually helped, while it is never observed whether the replacements would have been able to help. Responsibility judgments also depend only 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.

3. 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 Fig. 1 and varied the number of possible replacements for each person while keeping the outcome the same. In line with Eq. (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 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 Fig. 3). Similar to the ships, each machine requires exactly one gear of each type to work properly. However, the gears are sometimes broken, in which case other gears of the same type need to be used instead. Participants in this condition saw scenes in which the machine worked successfully and were asked to judge how responsible each gear was. Just like in the agent condition, we manipulated the number of possible replacement gears of each type.

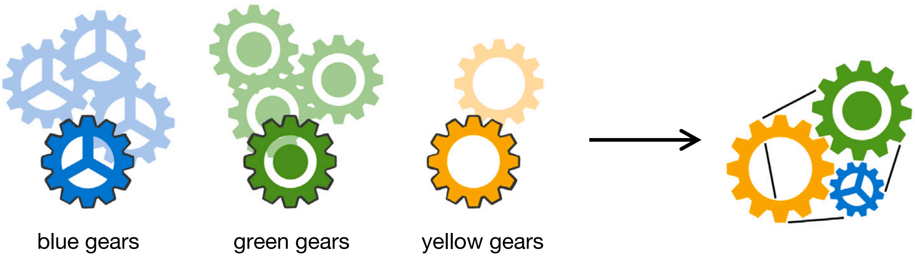

Fig. 3. Illustration of an example trial in the *object condition*. One gear of each type (highlighted in black) helped form the machine. Here, there were three other blue gears, three other green gears, and one other yellow gear that could have been potential replacements. Between trials, we varied the number of possible replacement gears of each type. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Methods

All materials including data, experiments, and analysis 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]

3.1.1. 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.

3.1.2. Procedure & design

Participants were first guided through instructions with two examples and then answered three comprehension questions to make sure they understood the setting (see Appendix A for details). 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 Fig. 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 whose endpoints were labeled “not at all” (0) and “very much” (100).

We emphasized that in every trial, the three contributors played an equal role in bringing about a 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 B.1 in the Appendix for details). For instance, Fig. 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 rating for each contributor 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 min (SD = 5.6).

3.2. Results

Fig. 4 shows participants’ mean responsibility judgments as a function of the number of possible replacements. The more replacements there were for a particular contributor, the less responsible people tended to hold that contributor for the outcome, regardless of whether it was an agent or an object. The effect was non-linear: additional replacements resulted in increasingly smaller differences in responsibility judgments. We discuss the results from each condition in turn.

3.2.1. Agent condition

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 equally to the outcome.[2] 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 Fig. B.1 in the Appendix for details). Then, given n possible replacements, Eq. (1) becomes

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

The black and white symbols in Fig. 4 show the predictions of the contribution model and the CRM, respectively. Participants’ judgments in the agent condition were well-captured by the CRM with a correlation of $r = 0.99$ and RMSE $= 1.40$. The replaceability predictor was credible (see Table 1).[3] 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[4] and used the same cross-validation procedure to evaluate which model explained each participant’s responses best. Table 1 shows that the CRM accounts best for the overall data and for the majority of individual participants (32 out of 50).

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

[2] Each craftsperson contributed equally in a descriptive sense, not necessarily in a commensurable way. The contribution model relates to the idea of value described in the Introduction, but does not explicitly compute or compare units of value; it simply reflects the fact that each person made a contribution to the outcome.

[3] We adopt the convention of calling an effect *credible* if the 95% HDI of the estimated parameter in the Bayesian model excludes 0.

[4] When fitting the CRM to individual participants, we constrained the replaceability predictor to be negative, because the CRM predicts that the more replaceable someone is, the less responsible they are held.

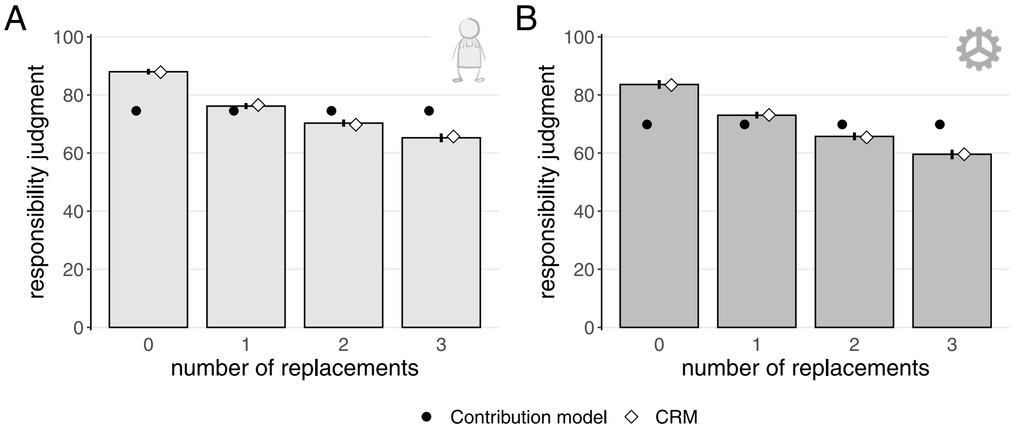

Fig. 4. Mean responsibility judgments for the (A) agent and (B) object conditions as a function of the number of replacements in Experiment 1. The black and white symbols show model predictions. Error bars are bootstrapped 95% confidence intervals.

3.2.2. Object condition

Mean responsibility judgments in the object condition were similar to those in the agent condition. They were again well-captured by the CRM with a correlation of $r = 0.96$ and RMSE = 2.64, and the replaceability predictor was credible. Table 1 shows that, in this condition too, the CRM fares better in the cross-validation on the overall data and best explains a majority of 37 out of 50 individual participants' judgments.

3.3. Discussion

The results of Experiment 1 show that even when each person's contribution towards 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 has shown that the number of contributors in a group 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 matters not only what one did, but also how easily one's contribution could have been replaced by someone else. Although the differences in responsibility are relatively small compared to the full scale measured, they are statistically credible, and the majority of individual participants assign responsibility in a way that is consistent with the CRM.

As predicted by the CRM, responsibility reduces non-linearly with each additional replacement. The largest difference in responsibility is between a contributor with no replacements and a contributor with just one. With an increasing number of replacements, the product of the probability terms approaches zero (see Eq. (1)). So, the model predicts that the more replacements there are, the less influence the absolute difference of the number of replacements has on responsibility.

The CRM accounted well both for the overall pattern of responses, as well as for the responses of individual participants. Only a small number of participants' responses were best explained by the contribution model. Participants' comments about what factors influenced their responses reflected these individual differences. For some participants, 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 participants, 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 is 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 predicts responsibility judgments as a function 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

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. The results show that replaceability is a credible 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)	32
Contribution	74.32 [68.55, 80.03]			8.24	−1591.1 (74.2)	18
Object condition						
CRM	85.72 [75.71, 91.71]	−41.86 [−64.34, −19.90]	0.96	2.39	0 (0)	37
Contribution	69.85 [64.25, 75.74]			8.96	−1777.8 (80.2)	13

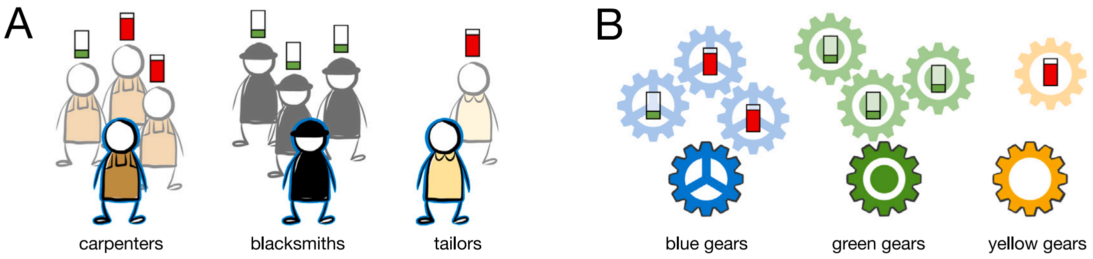

Fig. 5. Example trials in Experiment 2 for the (A) agent and (B) object conditions. A fuller red bar indicates that a craftsperson is more busy or that a gear is more brittle. An emptier green bar indicates that a craftsperson is less busy or that a gear is less brittle. More busy craftspeople and more brittle gears are less likely to be able to replace the actual contributor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

situation.[5] Chockler and Halpern's (2004) responsibility model also predicts the same non-linear decreasing effect. This is because the more potential replacements a contributor has, the farther their actions were from being pivotal (which would have required all the replacements to be unavailable), and thus the less responsible the contributor should be held. Since pivotality is inversely proportional to the number of counterfactual changes required, the difference in responsibility gets increasingly smaller as the number of replacements grows. To provide a more stringent test of the CRM, we manipulated both the number and availability of replacements in Experiment 2.

4. 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 merely 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 actually available (see Eq. (1)). For instance, a carpenter with a readily available replacement is more replaceable than one whose replacement has limited availability. In a counterfactual scenario, the replacement with limited availability would be less likely to actually step in to help build the ship. The CRM predicts that the carpenter whose replacement has limited availability is thus more responsible. In this experiment, we test whether the availability of replacements influences responsibility judgments.

4.1. 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).

4.1.1. 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* or *object* condition with $n = 50$ in each.

4.1.2. Procedure & design

The procedure and design were the same as that of Experiment 1, except that we additionally introduced the availability of each replacement. Fig. 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. Importantly, busyness and brittleness are probabilistic notions. We explained to participants that there was a small chance that replacements who were less busy or brittle (high availability) might still be unable to help if needed, or that replacements who were more busy or brittle (low availability) might actually be able to help. 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 C.1 in the Appendix). For example, one possible set consists of two replacements, one with high availability and one with low availability (set 5 in Fig. 6). Each trial featured three different sets (one for each contributor in the scene). We designed 20 trials and ensured that each set 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 in any one scene, in order to avoid visually overwhelming participants. 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 min (SD = 6.5) to complete the experiment.

4.2. Results

Fig. 6 shows participants' mean responsibility judgments across all possible sets of replacements. 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 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 Eq. (2). Another model is the CRM$_{\text{full}}$, which assumes two different probabilities p_{low} and p_{high} for 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_{low} is the number of replacements having low availability and n_{high} is the number with high availability. The parameters p, p_{low}, and

[5] Note, however, 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).

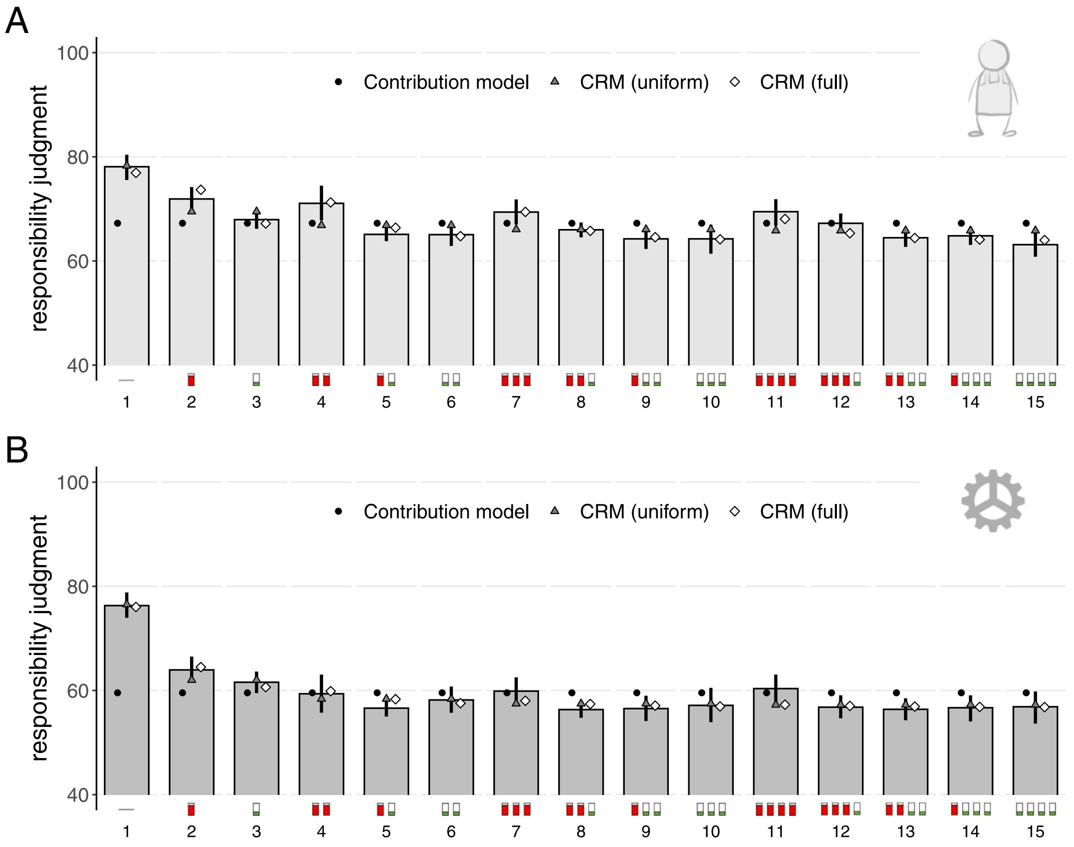

Fig. 6. Mean responsibility judgments for the (A) agent and (B) object conditions for each set of replacements in Experiment 2. Each contributor had up to four possible replacements, each of which had either low or high availability. The sets are ordered by increasing number and availability of replacements. 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 that was mapped from 0 to 100.

p_{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 Figs. C.1 and C.2 in the Appendix 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.

The results in Fig. 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 (set 15: mean responsibility 63.3, 95% CI [58.9, 67.3]) to be less responsible than a craftsperson whose four replacements all had low availability (set 11: 69.5 [65.1, 77.38]). We discuss the results from each condition in turn.

4.2.1. Agent condition

Mean responsibility judgments in the agent condition were well captured by the CRM$_{\text{full}}$ with a correlation of $r = 0.91$ and RMSE = 1.66. The best-fitting availability values were $p = 0.7$ for the CRM$_{\text{uniform}}$, and $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.75$ for the CRM$_{\text{full}}$. Fig. 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

Table 2
Results of model comparison for Experiment 2. '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 predictor, while the CRM models additionally computed replaceability by assuming either a uniform p (Eq. (3)) or varying p (Eq. (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.

Model	Intercept	Replaceability	r	RMSE	Δelpd (se)	n best
Agent condition						
CRM$_{\text{full}}$	76.91 [66.32, 87.33]	−12.96 [−24.23, −1.44]	0.91	1.66	0 (0)	26
CRM$_{\text{uniform}}$	78.27 [69.72, 86.76]	−12.56 [−19.99, −5.09]	0.78	2.49	−411.7 (35.9)	7
Contribution	67.54 [62.80, 72.35]			4.11	−586.4 (40.2)	17
Object condition						
CRM$_{\text{full}}$	76.01 [66.18, 85.59]	−19.17 [−30.55, −7.88]	0.92	2.01	0 (0)	21
CRM$_{\text{uniform}}$	76.52 [67.55, 85.54]	−19.35 [−29.84, −8.98]	0.92	2.13	−61.8 (6.8)	13
Contribution	59.49, [54.41, 64.56]			4.11	−307.3 (24.8)	16

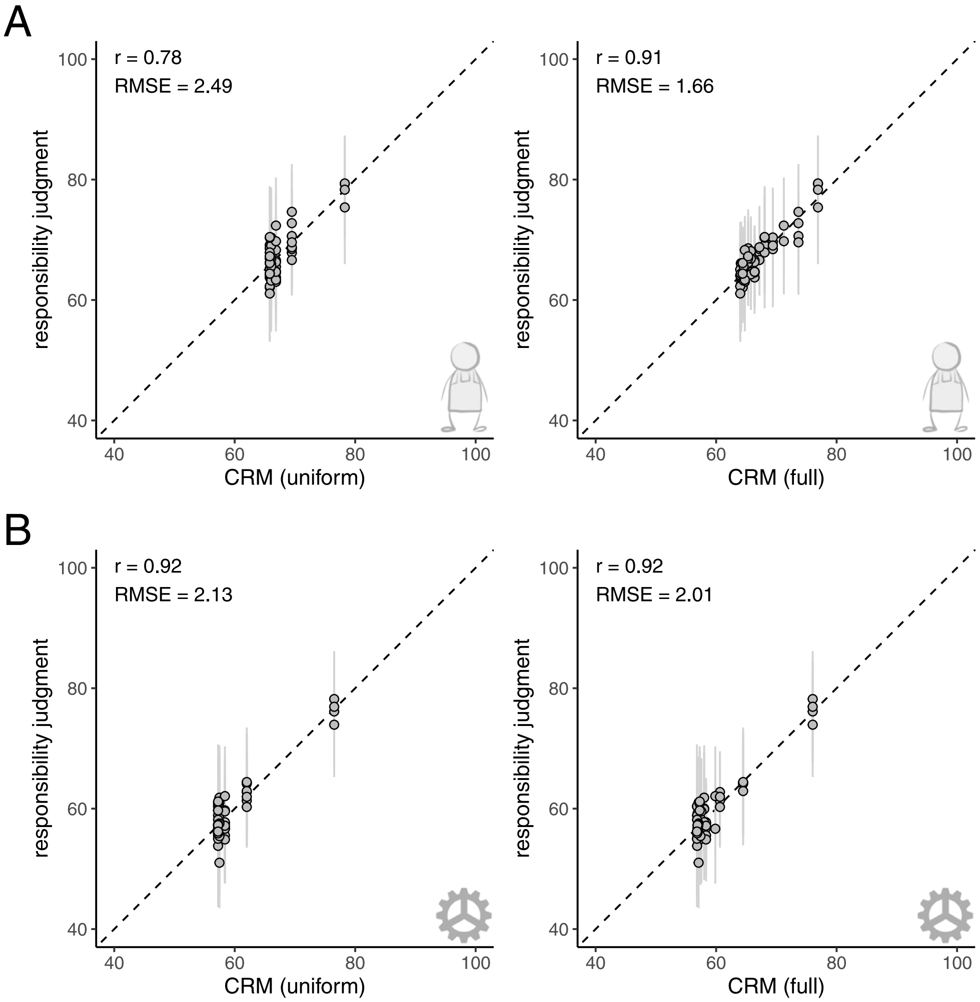

Fig. 7. Scatter plots showing the relationship between mean responsibility judgments and the uniform and full versions of the CRM in Experiment 2, in the (A) agent and (B) object conditions. Each point represents mean judgments for one contributor in one trial. Error bars are bootstrapped 95% confidence intervals. r = Pearson correlation coefficient, RMSE = root mean squared error. Note that the axes are truncated; participants judged responsibility on a scale from 0 to 100.

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 more individual participants than either of the other models. The replaceability predictor was credible in the full model.

4.2.2. 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 CRM$_{\text{uniform}}$, and $p_{\text{low}} = 0.6$ and $p_{\text{high}} = 0.8$ for the CRM$_{\text{full}}$. Here, the replaceability predictor in the full model was again credible. 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 in cross-validation and best explains the most individual participants.

4.3. 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 because it affects the probability of successful replacement – the addition of many replacements means little if they are all very busy, for example, but matters more if they have high availability. The results show that participants' responsibility judgments were sensitive to both the number of replacements and their individual availability and best explained by a full version of the CRM that considers both of these factors. Although the differences in responsibility between the situations were small relative to the full response scale, they are credible and cannot be captured by any existing models of responsibility. The better performance of the full CRM over one that assumes uniform replaceability demonstrates that responsibility judgments cannot be explained by a 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 were 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 contribution 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 brittleness of a machine part, perceived as generally reliable, may not suggest as much variation as the busyness 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 avoided specifying the prior availability of the actual contributor because we did not 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 must have been more available. On the other hand, perhaps the overall low availability 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 explore directly how the prior availability of the contributor affects responsibility judgments.

5. Experiment 3: Availability of contributor

Experiment 3 investigates how a contributor's own availability 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 does not consider features of the contributing cause itself, or how likely 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.

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. 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. Because if-likelihood and then-likelihood combine multiplicatively to determine counterfactual potency, this model predicts that Nine Ball would receive little responsibility for the successful outcome in either of these cases.

However, it is also possible that in contrast to the predictions of counterfactual potency, a lower if-likelihood may actually result in *more* responsibility. Consider the difference between a killer who is completely determined and one who wavers back and forth before committing the act. The counterfactual scenario in which the determined killer had not acted has low if-likelihood because it seems implausible for them not to 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. Somewhat counterintuitively, potency predicts that the determined killer would be less responsible than the hesitant one. Because the intuitions are mixed about how differences in if-likelihood may affect responsibility, we test this in our paradigm in Experiment 3 by manipulating the contributor's own availability.

5.1. 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).

5.1.1. 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 into the *agent* and *object* conditions with $n = 50$ in each.

5.1.2. Procedure & design

The setup and design of the experiment followed that of Experiments 1 and 2. On each trial, participants saw the availability of all craftspeople or gears, including the ones that actually helped and all of their potential replacements. Low and high availability were described in the same way as in Experiment 2, with low availability (more busy craftspeople or more brittle gears) reflecting lower likelihoods of successfully stepping in, and high availability (less busy craftspeople or less brittle gears) reflecting higher likelihoods of stepping in. We designed 20 different configurations in which 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 D.1 in the Appendix for details). For example, configuration 8 in Fig. 8 consists of a contributor with high prior availability and two low availability replacements.

We then designed 19 different trials featuring two configurations each. We used two contributors in each trial instead of three so that the scenes did not become visually overwhelming. Furthermore, to isolate the influence of the contributor versus the replacements, the contributors in each trial always had the same number of replacements and differed only in their own prior availability or in the availability of their replacements. Each configuration appeared in two different trials. For example, configuration 8 in Fig. 8 was contrasted with configuration 7 in one trial (different prior availability of contributor, same availability of replacements), and with configuration 10 in another trial (same prior availability of contributor, different availability of replacements). The exception is that configurations 1 and 2 were contrasted only with each other because they both have zero replacements. Participants took an average of 7 min (SD = 2.8).[6]

5.2. Results

Fig. 8 shows participants' mean responsibility judgments across the 20 different configurations 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 Eq. (3)), as well as an additional fixed effect of the prior availability of the contributor, $p_{\text{contributor}}$. This was equal to either p_{low} or p_{high}. We call this a CRM-based account because it computes replaceability as in the CRM and includes an additional predictor based on the availability of the contributor, which may have positive or negative effect on responsibility. The CRM-based account does not specify a cognitive mechanism by which the availability of

[6] For reporting this time, we excluded one outlying participant who took 10.5 h to complete the experiment (most likely by leaving their browser open before submitting), but included their data otherwise.

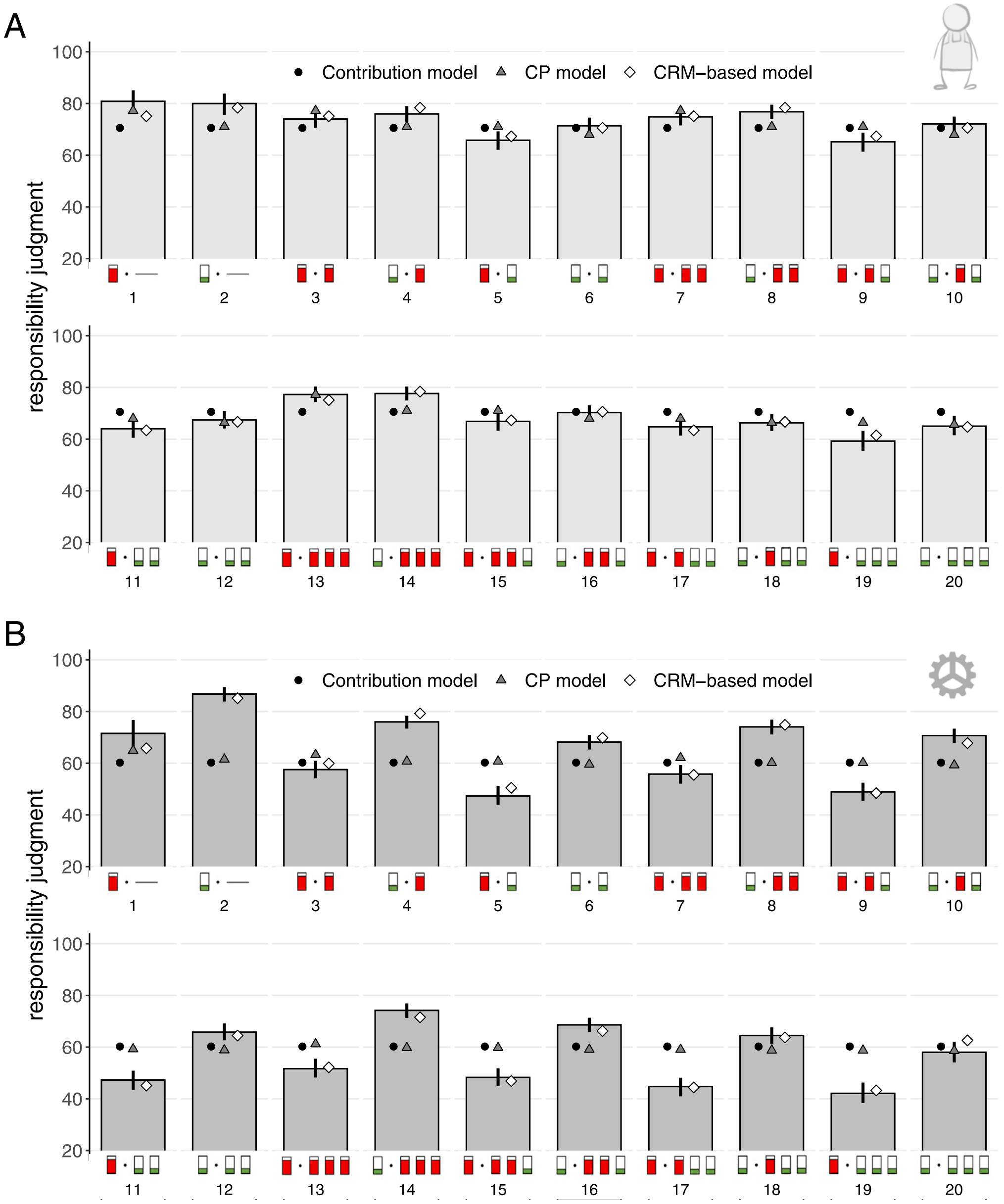

Fig. 8. Mean responsibility judgments in the (A) agent and (B) object conditions in Experiment 3. Each configuration on the x-axis is formatted as "contributor · replacements". The actual contributor had low or high prior availability and up to three possible replacements, each of which also had low or high availability. The configurations 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 that was mapped from 0 to 100.

the contributor affects responsibility judgments, but we discuss some possibilities in the General Discussion.

The second model we fit to responsibility judgments 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

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

The parameters p_{low} and p_{high} were fit to minimize the squared error between model predictions and participants' 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 Fig. D.1 in the Appendix 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-based account and CP model also included random slopes.

Fig. 8 reveals two main trends that hold across both conditions. 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. This is particularly noticeable in the object condition. We discuss the results in more detail from each condition in turn.

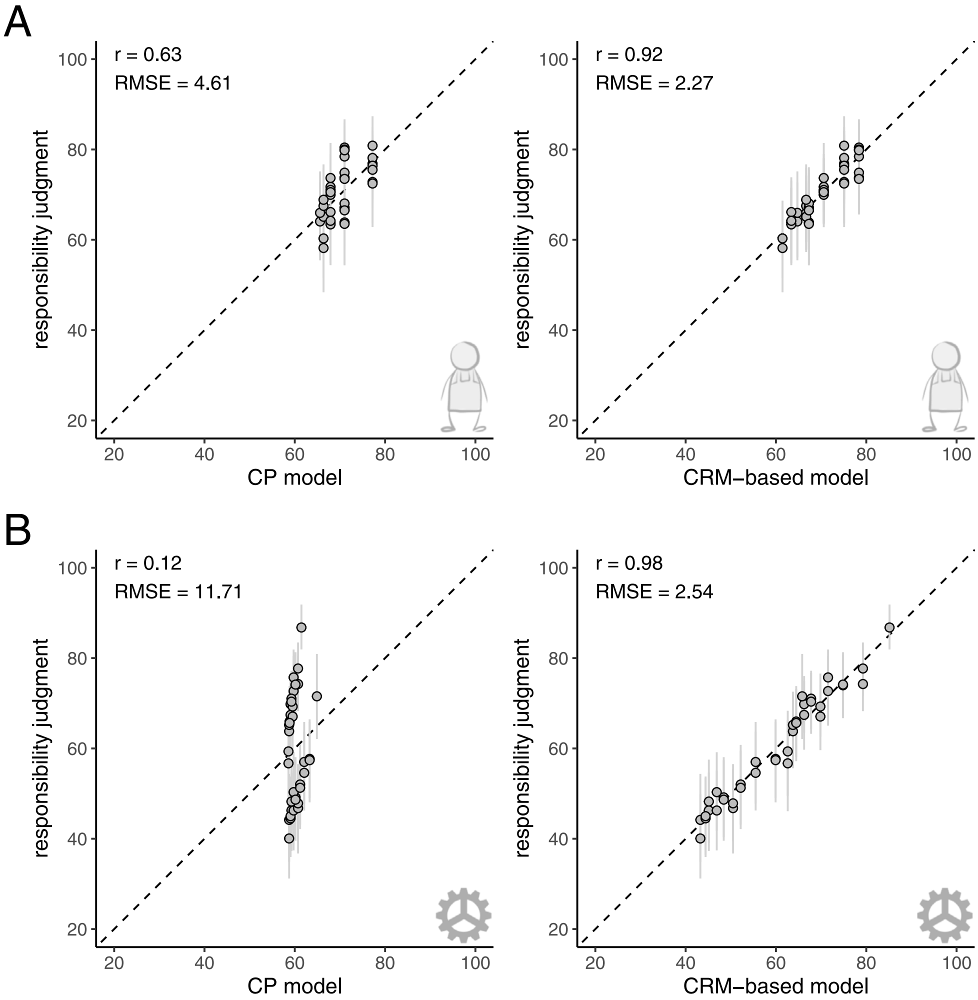

Fig. 9. Scatter plots showing the relationship between mean responsibility judgments and model predictions in Experiment 3, in the (A) agent and (B) object conditions. Each point represents mean judgments for one contributor in one trial. Error bars are bootstrapped 95% confidence intervals. r = Pearson correlation coefficient, RMSE = root mean squared error. Note that the y-axis is truncated; participants judged responsibility on a scale that was mapped from 0 to 100.

5.2.1. Agent condition

The filled symbols in Fig. 8 indicate model predictions. Participants' judgments in the agent condition were well-captured by the CRM-based account with a correlation of $r = 0.92$ and RMSE = 2.27 (see Fig. 9). The best-fitting availability values were $p_{\text{low}} = 0$ and $p_{\text{high}} = 0.5$.[7] In the CRM-based account, replaceability was a credible 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.[8] 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 10 out of 50 individuals. The CRM-based account best explains the overall data, and the judgments from 30 out of 50 individual participants.

5.2.2. 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 account with a correlation of $r = 0.98$ and RMSE = 2.54. 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-based account (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 credible. The results of the cross-validation show that the CRM-based account best explains judgments overall and also best captures a majority of 37 out of 50 individual participants.

5.3. 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

[7] The fact that the best-fitting value for p_{low} was 0 implies that the number of low availability replacements did not affect responsibility. For example, participants provided very similar responsibility judgments in configurations 3, 7, and 13 in Fig. 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 Fig. D.1 shows. So even if the parameter took on a slightly higher value, the model would still capture judgments well.

[8] We left both the contributor and replaceability predictors unconstrained for the CRM-based model here because Stan (Carpenter et al., 2017) does not support setting different bounds on different predictors. The posterior estimates on the replaceability predictor were in the predicted direction for most of the participants best fit by the CRM-based account (25 out of 30 in the agent condition, and 29 out of 37 in the object condition).

Table 3
Experiment 3 model comparison. ‘Intercept‘, ‘Contributor‘, ‘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-based account also included the availability of the contributor and replaceability (Eq. (3)) and the CP model also included potency (Eq. (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-based	74.91 [67.96, 81.84]	6.58 [−8.43, 22.72]	−15.59 [−22.04, −9.09]	0.92	2.27	0 (0)	30
CP	64.82 [57.02, 72.70]	Potency: 12.40 [3.65, 21.32]		0.63	4.61	−579.9 (55.9)	10
Contribution	70.55 [64.61, 76.60]				8.24	−793.0 (61.8)	10
Object condition							
CRM-based	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

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-based account but not the CP model. 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 CP model predicts responsibility judgments to be a multiplicative combination of replaceability and the prior availability of the contributor. 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 of the contributor here, which resulted in the CRM-based account outperforming the CP model in both conditions.

The effect of the contributor's availability on responsibility judgments was stronger in the object condition. Looking at the subset of participants who were best fit by the CRM-based account, we found that more participants assigned a positive weight to the contributor in the object condition (86%) compared to the agent condition (42%). 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 credible, overall effect in the agent condition.

Why did we find that a contributor was held somewhat *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.

6. 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 likely a group outcome would have turned out differently, had a particular contribution not been made. 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 their 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 studied 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 individual availability. In Experiment 3, we additionally manipulated the prior availability of the contributor. Across all three experiments, participants' judgments were sensitive to replaceability. This was true both in a social domain, where the contributions were made by agents, and in a physical domain, where the contributions were made by components of a mechanistic device. The CRM outperformed alternative models that consider only the actual contributions, and the CRM-based account in Experiment 3 outperformed a model based on counterfactual potency (CP, Petrocelli et al., 2011). In contrast to what the CP model predicts, contributions that were unlikely to have needed replacement were held *more* responsible for the outcome.

In the following sections, we discuss several aspects of the CRM in more detail and propose directions for future work. First, we discuss the process of computing replaceability and simulating counterfactuals. Then, we discuss the relationship between prior availability and normality. In the last few sections, we expand the discussion to responsibility for negative outcomes and omissive causation, responsibility for agents vs. objects, and the problem of counterfactual selection.

6.1. Computing replaceability and simulating counterfactuals

The CRM computes replaceability by taking into account the individual replacements' availabilities. For instance, consider $n = 3$ replacements in a situation with $p_{\text{low}} = 0.25$ and $p_{\text{high}} = 0.75$. If all three replacements have high availability, then the probability of successful replacement is $1-(1-0.75)^3 = 0.98$. If all three replacements have low availability, then the probability falls to $1 - (1 - 0.25)^3 = 0.58$. But people tend to believe that the first outcome is only slightly more likely than the second, failing to recognize how rapidly conjunctive probability drops off. Prior research has shown that people have difficulty estimating the probability of outcomes to which multiple factors contribute (Bar-Hillel, 1973; Gerstenberg, Lagnado, & Zultan, 2023; Nilsson, Rieskamp, & Jenny, 2013). So, while it is unlikely that participants in our experiments computed probabilities in the exact same manner as the CRM, their responsibility judgments suggest that they were nonetheless sensitive to the differences in the probabilities. For example, in Experiment 2, participants assigned more responsibility to a contributor with a single replacement when that replacement had low availability, compared to when they had high availability. Follow-up work might empirically test for and use people's subjective estimates of the probability of replacement in each trial directly, instead of computing replaceability from the model.

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, Goodman, Lagnado, & Tenenbaum, 2021; Pearl, 2000). Here, 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. However, in many situations, there is uncertainty not only about whether a replacement could have been found, but also about whether that replacement would have been as successful at the task. In a more complex case where, for instance, the quality of the ship also matters, responsibility would depend on not only whether a replacement carpenter could have been found, but also how good of a job they would have done. Complex counterfactual simulations may require people to abstract certain elements of their causal models or to rely on heuristics to compute.

6.2. Prior availability and normality

In Experiment 3, we manipulated the normality of the contributor's action 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. 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 contributor availability relate to normality in our paradigm? On the one hand, for a contributor with high prior availability, an act of helping is more normal for them as they are more likely to say "yes" compared to a contributor with low prior availability. On the other hand, having high availability may imply being previously engaged in fewer tasks, which would make a particular instance of helping more abnormal for that contributor.

If we assume that a contributor's helping actions are more abnormal when they have high availability (because they helped less often or were used less often in the past, which produced the high availability in the first place), then our findings are consistent with prior research showing that people tend to attribute more causality to abnormal factors in conjunctive causal structures in which each person's contribution was 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 of 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, instead, to view a contributor's helping as more normal when they had high prior availability. A highly available contributor was more likely to say "yes" in the first place, so the fact that they did so, rather than not, is the more normal event (Kahneman & Miller, 1986). Norms extend to physical objects as well, specifically norms of proper functioning (Hitchcock & Knobe, 2009; Kominsky & Phillips, 2019). We generally expect mechanical parts such as gears that are less brittle to work more often, so it is similarly the more normal event when they do so. In that case, our results in Experiment 3 run counter to what some prior research has found.

One possible explanation is that participants' judgments were affected by additional inferences they made from a contributor's availability. Prior work has shown that consideration of an agent's skills and capacities affect responsibility attributions (Gerstenberg, Ejova, & Lagnado, 2011; Malle et al., 2014; Weiner & Kukla, 1970). For example, some participants inferred effort from 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".), and others inferred quality (e.g., "I rated green gears higher since durability means it most likely will support the machine longer".). In the case of agents specifically, busyness is suggestive of a craftsperson's skills (i.e. good craftspeople tend to be in greater demand). Some participants might have inferred that busier craftspeople were more skilled and thus should be *more* responsible (e.g. "Business [sic] likely increased output and productivity".). The various inferences supported by 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. Future work needs to study more closely what inferences participants make based on a contributor's prior availability. The CRM only considers the probability with which a successful replacement could have been found, but the additional inferences that people draw about the nature of the contribution clearly matter, too.

6.3. Negative outcomes and omissive causation

In our experiments, we only looked at positive outcomes. How would replaceability affect responsibility for negative outcomes? Prior research has shown asymmetries between praise for good outcomes and blame for bad ones. For example, blame judgments for individual actions tend to be more extreme than praise judgments (Guglielmo & Malle, 2019). The predictions of the CRM are similar for positive and negative outcomes: a person would be blamed less for a negative outcome, the more likely someone else would have replaced them and done the same. Prior work suggests that this prediction aligns with people's intuitions (Falk & Szech, 2013; Gantman et al., 2020). For example, in Falk and Szech's (2013) experiment, participants felt less responsible for killing a mouse (a bad outcome) when there were more others who could have taken their place.

Our experiments also focused on situations in which agents or objects actively contributed to the outcome. What about situations in which someone failed to contribute? Prior work has shown that negligence readily elicits blame e.g. Sarin & Cushman, 2022 and that causal judgments about omissions are influenced by various factors such as the normality of the omission e.g. Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Henne, Pinillos, & De Brigard, 2017; Khemlani, Bello, Briggs, Harner, & Wasylyshyn, 2021; Livengood & Machery, 2007. Gerstenberg and Stephan (2021) showed that people's causal judgments about omissions in physical scenarios are well predicted by a model that simulates how likely the outcome would have been different, if the event that was omitted had actually happened instead.

The CRM naturally yields predictions about cases of omissive causation. Suppose Alice leaves for vacation and asks one of her three neighbors to water her plants while she is away. He forgets, so the plants die. How responsible is the neighbor for the plants dying? Alice could have asked either one of her other neighbors as well. Intuitively, if either of them would also have forgotten to water her plants, then the neighbor who actually forgot seems somewhat less responsible (compared to a situation in which the other neighbors would have certainly remembered). The higher the probability that a replacement would have made the same omission, the less responsible the person in question seems to be. Thus, there may be a similar effect of the potential replacements' individual probabilities of success (where "success" in this case means making the same omission) on responsibility for omissions and actions.

6.4. Agents vs. objects

Across all three experiments, responsibility judgments were very similar in the agent and object conditions. Since the causal structure was the same across both conditions, perhaps it is not that surprising that responsibility judgments were similar. Prior work by Lagnado and Gerstenberg (2015) which focused on the effects of causal structure on responsibility judgments to individuals in groups, also found very similar responses in social and non-social domains.

However, there are conceptions of responsibility that treat agents and objects differently. In particular, the factors that affect judgments of an agent's *moral* responsibility having to do with their mental states or moral character e.g. Cushman, 2008; Vincent, 2011; Zhao & Kushnir, 2022 do not apply to objects. It would not make sense to say that a broken gear "acted irresponsibly" in the same way that a carpenter who purposely fumbled his job did. These agentive factors can be intertwined with replaceability, as people who are exceptionally virtuous, talented, etc. are also often rare. For example, a firefighter who ran into a burning building to rescue a baby may be considered more responsible for the rescue than the passerby who called 911, because fewer people could have done what the firefighter did. That said, the firefighter's responsibility may also be on account of the very virtue of her self-sacrificing actions which make her more irreplaceable to begin with. Such possible agentive confounds of replaceability are not a problem for the physical domain, but need to be teased apart in future work on moral responsibility in the social domain.

Many situations in our everyday lives do not involve agents or objects exclusively. The contributions of functional artifacts or machines can sometimes be replaced by those of people, or vice versa. For example, a tailor could be replaced not only by other tailors but also by an automatic sewing machine in a factory. In our experiments, the set of possible replacements in each trial was given explicitly, but the CRM itself does not restrict what the replacements are and only considers the probability that each one would have succeeded. When the replacements are not specified, people must rely on their understanding of the situation to consider what relevant counterfactuals to select. We discuss this in the next section.

6.5. 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, Pizarro, & Diermeier, 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 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, more work is needed to better understand what kinds of counterfactuals most naturally come to people's minds (Byrne, 2005; Gerstenberg & Stephan, 2021; Kominsky & Phillips, 2019). At minimum, this seems to depend on someone's knowledge of the situation, and on what is 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.

The flexibility of counterfactual selection suggests that replaceability may play a role in explaining various phenomena in causal judgments beyond the factors manipulated here. For instance, the effect of norms on judgments of responsibility and causality has been explained in terms of the ease of imagining relevant counterfactual situations in which the outcome could have been different (Icard et al., 2017; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Phillips, Luguri, & Knobe, 2015). Norm-violating agents or objects tend to be held more responsible because it is easier to imagine a counterfactual scenario in which they would have acted or functioned in a norm-conforming way – an action-centered counterfactual. Replaceability offers a complementary explanation in the form of a person-centered counterfactual. Norm-violating agents are more responsible because they are less replaceable; it is harder to imagine others who would have also violated the norm and thus could have replaced them, compared to others who would have conformed to a norm.

7. 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.

CRediT authorship contribution statement

Sarah A. Wu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Tobias Gerstenberg:** Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Data availability

The data are available here: https://github.com/cicl-stanford/responsibility_replacement.

Acknowledgments

We thank the members of the Causality in Cognition Lab at Stanford University, Ruth Byrne, and an anonymous reviewer for their valuable comments and suggestions on this paper. Early versions of this work were presented at the 43rd Annual Meeting of the Cognitive Science Society, the 47th Annual Meeting of the Society for Philosophy and Psychology, and an Experimental Jurisprudence workshop at the University of Michigan in 2022. We thank the discussants there for their thoughtful feedback. This work was supported by the NSF GRFP, United States of America and a grant from the Stanford Institute for Research in the Social Sciences, United States of America to S.A.W., and a grant from the Stanford Human-Centered Artificial Intelligence Institute, United States of America to T.G. All experiments were approved by Stanford's Institutional Review Board.

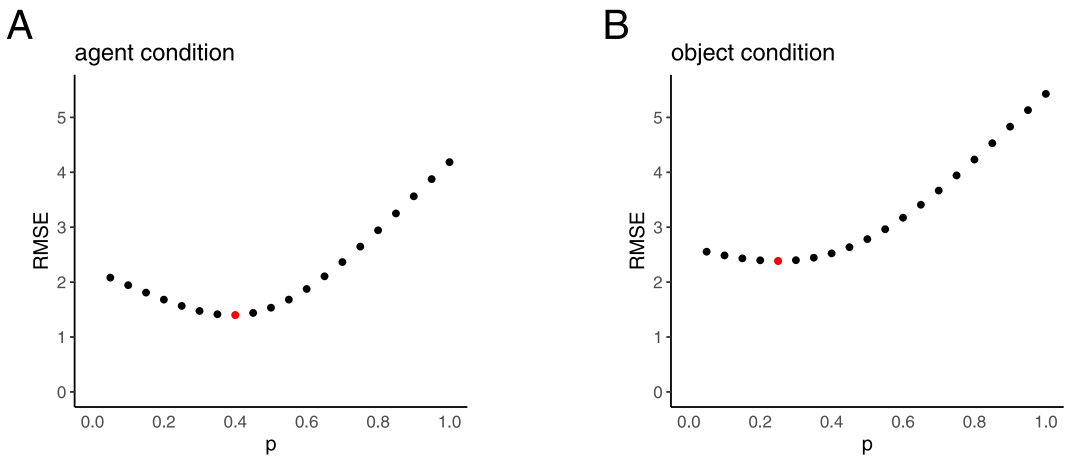

Fig. B.1. Results of fitting a uniform probability of success parameter in Experiment 1, in the (A) agent and (B) object conditions. The best-fitting values, indicated in red, minimize the squared error between model predictions and participants' judgments. The best-fitting parameter value was $p = 0.4$ 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 A. Sample experiment text

As an example, participants in the *agent* condition of Experiment 1 read the following text as part of the instructions. Each page also contained an illustration. Thus, the text alone does not capture all the information that was conveyed to participants. We invite readers to try out the full experiments here: https://github.com/cicl-stanford/responsibility_replacement.

Page 1 Imagine a coastal land with many villages that produce ships. Each ship is made out of wood, metal, and fabric.

Page 2 The villages are inhabited by various craftspeople including carpenters (who work with wood), blacksmiths (who work with metal), and tailors (who work with fabric). Each ship requires one craftsperson of each type to build.

Page 3 Some villages have the same number of each type of craftsperson, but other villages have more of one type and less of another. For instance, the village Aramoor has 3 carpenters, 2 blacksmiths, and 3 tailors.

Page 4 Whenever a village gets an order for a new ship, it must find craftspeople to fulfill the order. Craftspeople tend to be very busy, so many of them have to say "no" when asked to help build the ship.

Page 5 Aramoor was unable to find a blacksmith available to help, so unfortunately it failed to build a ship. The craftspeople who were available to help are outlined in red.

Page 6 As another example, the village Skystead has 4 carpenters, 2 blacksmiths, and 2 tailors.

Page 7 Skystead was able to find one craftsperson of each type available to help (outlined in red), so here building the ship was a success! In this experiment, we will show you scenes like this depicting different villages that successfully built ships. We are interested in seeing how responsible you think each craftsperson who ended up helping was for the success.

Participants were required to answer the following comprehension questions correctly before moving on to the test trials.

1. True or False: Each ship requires exactly one carpenter, one blacksmith, and one tailor to build. (Correct answer: True)
2. True or False: All villages have the same number of carpenters, blacksmiths, and tailors. (Correct answer: False)
3. True or False: Craftspeople are always available to work on a ship if asked. (Correct answer: False)

Appendix B. Experiment 1 details

See Fig. B.1 and Table B.1.

Table B.1
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. The contributor with only one replacement happened to be the tailor in the agent condition (see Fig. 1) and the yellow gear in the object condition (see Fig. 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

Appendix C. Experiment 2 details

See Figs. C.1 and C.2 and Table C.1.

Table C.1
Number of replacements with low availability and number of replacements with high availability in each trial in Experiment 2. Each set of replacements was randomly permuted among the three contributors. 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_{low}	0	0	0	1	4	0	1	0	1	1	3	0	3	0	0	0	1	1	0	0
	n_{high}	0	0	0	2	0	1	0	3	0	3	0	3	1	4	2	4	1	2	2	3
Blacksmiths	n_{low}	4	0	2	3	0	3	0	4	2	3	0	2	1	2	2	1	3	2	3	1
	n_{high}	0	4	1	1	1	1	4	0	1	0	1	1	3	1	2	1	0	1	0	1
Tailors	n_{low}	1	3	4	0	2	1	2	0	2	1	1	1	0	0	2	1	2	1	1	2
	n_{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_{low}	0	0	2	3	4	1	0	0	2	1	0	1	0	0	0	0	1	1	3	0
	n_{high}	0	0	1	1	0	3	4	1	1	0	1	0	2	2	2	4	1	1	0	3
Green gears	n_{low}	1	0	4	1	0	0	2	0	2	1	3	2	3	2	2	1	2	1	1	1
	n_{high}	3	4	0	2	1	1	2	3	2	3	0	1	1	1	0	1	1	2	1	1
Blue gears	n_{low}	4	3	0	0	2	3	1	4	1	3	1	0	1	0	2	1	3	2	0	2
	n_{high}	0	1	0	0	2	1	0	0	0	0	2	3	3	4	2	1	0	1	2	0

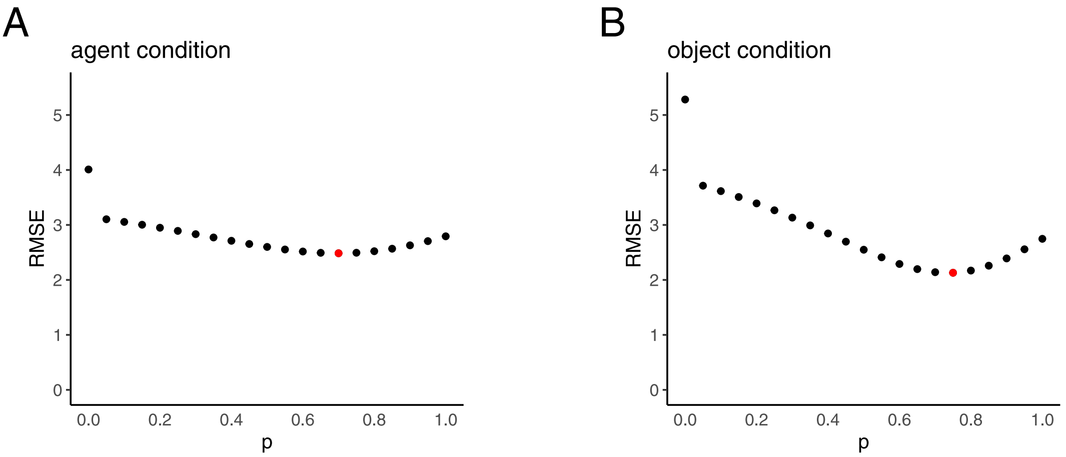

Fig. C.1. Results of fitting a uniform probability of success parameter in Experiment 2, in the (A) agent and (B) object conditions. The best-fitting values, indicated in red, minimize the squared error between model predictions and participants' judgments. This value was $p = 0.7$ in the agent condition 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

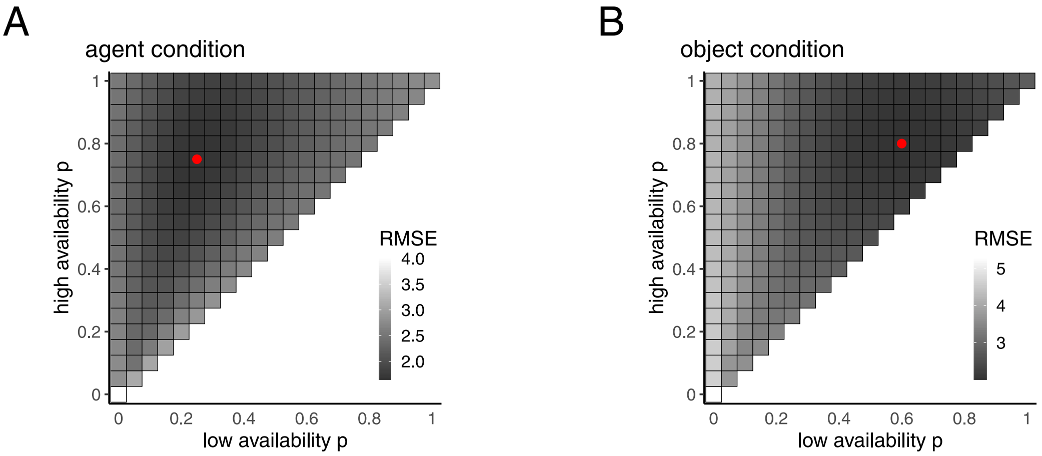

Fig. C.2. Results of a grid search over the two probability of success parameters for replacements with low and high availability in Experiment 2, in the (A) agent and (B) object conditions. The best-fitting parameters, indicated with the red dots, minimize the squared error between model predictions 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 D. Experiment 3 details

See Fig. D.1 and Table D.1.

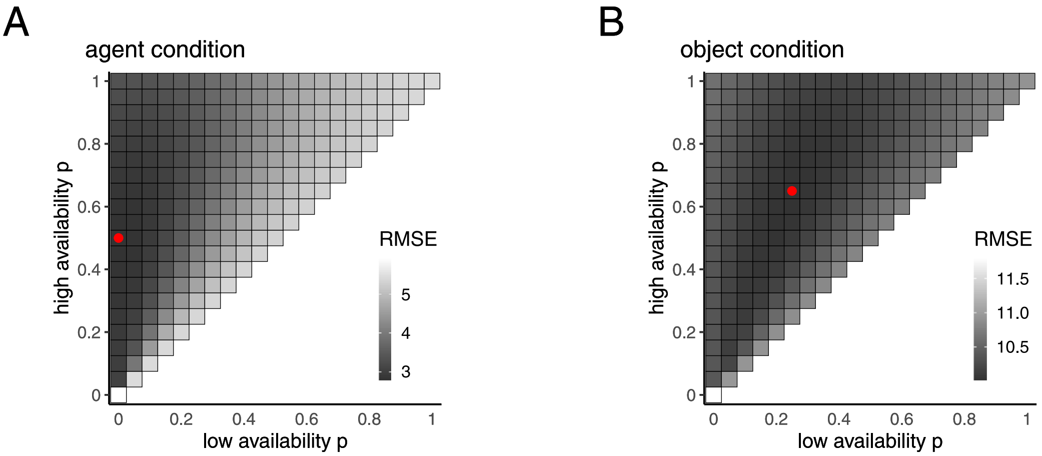

Fig. D.1. Results of a grid search over the two probability of success parameters for replacements with low and high availability in Experiment 3, in the (A) agent and (B) object conditions. The best-fitting parameters, indicated with the red dots, minimize the squared error between model predictions 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 highly available craftsperson had a 0.5 chance of saying "yes" if asked to help with building the ship.

Table D.1
Information about the replacements and the prior availability of the contributor (H = high, L = low) for each trial in Experiment 3. The configurations were randomly assigned between the two contributors in each trial. That is, the carpenters share the same configuration as the yellow gears in some trials and the blue gears in other trials, and the tailors have the opposite pattern.

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

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). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*(2), 134–143. http://dx.doi.org/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. http://dx.doi.org/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. *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*, Article 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 positive difference: Criticality in groups. *Cognition*, *238*, Article 105499. http://dx.doi.org/10.1016/j.cognition.2023.105499.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, Article 104842.

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). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLoS One*, *14*(3), Article e0213544. http://dx.doi.org/10.1371/journal.pone.0213544, Publisher: Public Library of Science.

Hale, B. (2011). Nonrenewable resources and the inevitability of outcomes. *The Monist*, *94*(3), 369–390. http://dx.doi.org/10.5840/monist201194319.

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*, Article 104708.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164. http://dx.doi.org/10.1016/j.cognition.2019.05.006.

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, *95*(2), 270–283.

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. http://dx.doi.org/10.1016/j.cognition.2017.01.010.

Johnson, B. L. (2003). Ethical obligations in a tragedy of the commons. *Environmental Values*, *12*(3), 271–287. http://dx.doi.org/10.3197/096327103129341324.

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

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

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.

Khemlani, S., Bello, P., Briggs, G., Harner, H., & Wasylyshyn, C. (2021). Much ado about nothing: The mental representation of omissive relations. *Frontiers in Psychology*, *11*, Article 609658.

Khemlani, S., & Oppenheimer, D. M. (2011). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological Bulletin*, *137*(2), 195–210.

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

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

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: Intuition and diversity, vol. 2*. The MIT Press.

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), Article e12792. http://dx.doi.org/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*, Article 101412.

Livengood, J. (2013). Actual causation and simple voting scenarios. *Noûs*, *47*(2), 316–345.

Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, *31*(1), 107–127.

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). 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.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65–94.

Phillips, J., Luguri, J., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

R. Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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. http://dx.doi.org/10.1111/j.1747-4469.1996.tb00098.x.

Sarin, A., & Cushman, F. A. (2022). One thought too few: Why we punish negligence. http://dx.doi.org/10.31234/osf.io/mj769, preprint, PsyArXiv.

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.

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*. New York: Springer-Verlag.

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.

Vincent, N. A. (2011). A structured taxonomy of responsibility concepts. In N. a Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral responsibility: Beyond free will and determinism*. Dordrecht: Springer.

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., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, Article 105609. http://dx.doi.org/10.1016/j.cognition.2023.105609.

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*, Article e13257.

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