Cognition 238 (2023) 105499

Contents lists available at ScienceDirect

Cognition

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

Original articles

Making a positive difference: Criticality in groups

Tobias Gerstenberg [a,*], David A. Lagnado [b], Ro'i Zultan [c]

[a] *Stanford University, United States of America*
[b] *University College London, United Kingdom*
[c] *Ben-Gurion University of the Negev, Israel*

ARTICLE INFO

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

Keywords:
Criticality
Causality
Counterfactuals
Responsibility
Pivotality

ABSTRACT

How critical are individual members perceived to be for their group's performance? In this paper, we show that judgments of criticality are intimately linked to considering responsibility. Prospective responsibility attributions in groups are relevant across many domains and situations, and have the potential to influence motivation, performance, and allocation of resources. We develop various models that differ in how the relationship between criticality and responsibility is conceptualized. To test our models, we experimentally vary the task structure (disjunctive, conjunctive, and mixed) and the abilities of the group members (which affects their probability of success). We show that both factors influence criticality judgments, and that a model which construes criticality as anticipated credit best explains participants' judgments. Unlike prior work that has defined criticality as anticipated responsibility for both success and failures, our results suggest that people only consider the possible outcomes in which an individual contributed to a group success, but disregard group failure.

1. Introduction

Three psychologists are collaborating on an exciting new project studying responsibility attributions in groups. One is developing the theoretical model, while the others are running two separate experiments. The three colleagues learn about a special issue in a journal dedicated to responsibility, and wish to submit their work. Unfortunately, the submission deadline is near, and it is not certain that any of the three tasks will be completed in time. The three colleagues nonetheless decide to make an effort to prepare the submission. For a successful submission, they need a theoretical model and at least one experimental study. How critical is each of them for making the deadline?

Most research on responsibility attribution has focused on the problem of how people assign responsibility to individuals (e.g. Alicke, 2000; Shaver, 1985; Weiner, 1995). When people credit or blame others, they care about whether the outcome was intended (Lagnado & Channon, 2008), foreseeable (Brickman, Ryan, & Wortman, 1975) and under the control of the agent (Gerstenberg, Ejova, & Lagnado, 2011; Gerstenberg et al., 2018; McClure, Hilton, & Sutton, 2007). Researchers have also looked at how responsibility is attributed in groups (Douer & Meyer, 2022; El Zein, Bahrami, & Hertwig, 2019; Forsyth, Zyzniewski, & Giammanco, 2002; Gantman, Sternisko, Gollwitzer, Oettingen, & Van Bavel, 2020; Gerstenberg & Lagnado, 2010, 2012; Gerstenberg, Lagnado, & Kareev, 2010; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018; Lagnado & Gerstenberg, 2015; Lagnado, Gerstenberg, & Zultan, 2013; Teigen & Brun, 2011; Wu & Gerstenberg, 2023; Zultan, Gerstenberg, & Lagnado, 2012).

How much responsibility an individual group member receives depends not only on their performance. It also matters what the other group members did, and how the individual contributions combined to determine the group outcome (Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Zultan et al., 2012). While in some situations all group members need to perform well for the group to succeed, other situations require fewer members to do their best (Steiner, 1972). Responsibility attributions are sensitive to these structural differences (e.g. Lagnado et al., 2013). In these studies, participants assign responsibility *ex-post*. They first learn what happened and then attribute credit or blame. However, one can also assign responsibility *ex-ante*, that is, before the outcome has occurred. The word "responsibility" is polysemous and refers to several related but distinct concepts (see Hart, 2008; Sousa, 2009). It can be your responsibility to make sure everything goes well. And you can be the one who is held responsible when things went south.

In Lagnado et al. (2013), we linked the two notions of responsibility in the *criticality–pivotality framework*. Within this framework, ex-post responsibility is constructed as a function of both ex-ante criticality and ex-post pivotality. Criticality is a forward-looking concept and captures the extent to which a person's contribution is important for the

[*] 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.2023.105499
Received 8 February 2023; Received in revised form 9 May 2023; Accepted 20 May 2023

outcome. Pivotality is a backward-looking concept and it expresses how close a person's contribution was to actually having made a difference to the outcome. In Lagnado et al.'s (2013) experiment, participants evaluated players in team tasks that differed in how the individual contributions combined to bring about the outcome. For example, in one of the team tasks, both players A and B had to succeed in addition to at least one out of players C and D. When asked to evaluate how critical each player was (before the players' had performed their tasks), participants judged players A and B to be more critical than players C and D in this task. And when evaluating how responsible player C was for the team's success, player C received more responsibility when player D failed (and the rest succeeded) compared to when all players succeeded. When player D failed, player C's contribution was pivotal for the positive outcome — the team would have lost had player C also failed in that situation. However, when both player C and D succeeded, then the team would have won even if player C (or D) had failed. Lagnado et al. (2013) found that both criticality and pivotality were important for capturing participants' responsibility judgments (see also Quillien & Lucas, 2022).

In this paper, we focus on the notion of criticality. We first review prior work on perceptions of criticality in group outcomes. We then introduce different models of criticality judgments that we test in subsequent experiments. We reanalyze the data from Lagnado et al. (2013) and find that three models generate highly inter-correlated predictions that closely match participants' criticality judgments in that study. In Experiment 1, we present new situations designed to separate the predictions of these models. Experiment 2 then introduces a stochastic environment that helps to further tease the models apart. We find that across all of the experiments, a model captures participants' judgments best which assumes that people consider each player's chances of being pivotal for a positive outcome.

1.1. Criticality in public goods games

Early work on criticality looked at individuals' decisions of whether or not to contribute to a public good (Hardin, 1968). In the step-level public goods game, a public good is provided if at least k out of n players contribute. In this setting, Rapoport (1987) defined criticality as the probability of being both necessary and sufficient for the provision of the public good. The probability of being critical π is

$$\pi = \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \tag{1}$$

whereby n equals the number of players in the group and k denotes the provision point that has to be reached in order for the public good to be provided (Au, Chen, & Komorita, 1998). The parameter p is given by the probability that the other players in the group will contribute.[1] For example, in a public goods game with $n = 6$ players, a provision point of $k = 4$ and a probability of $p = 0.6$ that each of the other players will contribute, the probability that the player under consideration would be critical is $\pi = \binom{5}{3} \cdot 0.6^3 \cdot 0.4^2 \approx 0.35$.

As predicted by this model, an increased group size generally leads to a decrease in participants' ratings of perceived criticality and fewer contributions to the public good (Kerr, 1989; but see Isaac, Walker, & Williams, 1994). Keeping the ratio between provision point and group size identical, a person's criticality decreases with an increased group size. For example, holding the probability that the others will contribute at $p = 0.5$, a player's criticality is $\pi = 0.25$ when $n = 3$ and $k = 1$, compared to $\pi = 0.0187$ when $n = 30$ and $k = 10$. Kerr (1989) showed that people's assumption that individual contributions are less critical in larger groups overgeneralizes to situations in which this is not the case (i.e. when the provision point k and the probability of contribution p are manipulated to counteract the effect of group size).

In addition to group size (Au, 2004; Kerr, 1989; Kollock, 1998), a number of other factors influence perceptions of criticality such as the players' initial endowment (De Cremer, 2007; Rapoport, 1988; Rapoport, Bornstein, & Erev, 1989), the impact that a player's decision to contribute has on the likelihood of the public good provision (Kerr, 1992), the expectation that other players are going to contribute (Kerr & Kaufman-Gilliland, 1997), uncertainty about the provision point (Dannenberg, Löschel, Paolacci, Reif, & Tavoni, 2015), and, in a sequential version of the step-level public goods game, the order in which the players contribute (Anselm et al., 2022; Au, 2004; Au et al., 1998; Au & Chung, 2007; Bartling, Fischbacher, & Schudy, 2015).

The degree to which an individual perceives her contribution to be critical for the provision of the public good affects how likely they will contribute (Duch, Przepiorka, & Stevenson, 2015; Kerr, 1989, 1992, 1996; see also Falk, Neuber, & Szech, 2020). People are also more likely to contribute the more responsible they feel to the others in the group (De Cremer & van Dijk, 2002; Spadaro et al., 2022). In fact, there is a close connection between work on criticality in public goods games, and work on a-priori voting power which looks at the influence that an individual has over an election outcome (see e.g. Felsenthal & Machover, 2004; Gelman, Katz, & Tuerlinckx, 2002).

1.2. Responsibility for voting outcomes

The chance of casting the pivotal vote in an election is usually very small (Riker & Ordeshook, 1968). So, from a cost–benefit standpoint, voting may seem irrational (Meehl, 1977). It takes effort to do so, and the chances of making a difference are minuscule. One way to address the so-called "paradox of voting" is to move from an all-or-none concept of casting the pivotal vote toward a more graded notion of pivotality (Bartling et al., 2015; Braham & Hees, 2009; Braham & van Hees, 2013; Chockler & Halpern, 2004; Felsenthal & Machover, 2009; Goldman, 1999; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021). People may be motivated by the partial responsibility they perceive for having contributed to the outcome, and by the anticipated blame they would feel for a negative outcome if they had not voted.

Chockler and Halpern (2004) developed a model which assigns graded responsibility based on how close a person's action was to having been pivotal. The more things would have needed to change about the situation to make the person's action pivotal the less responsible the person is (see Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Langenhoff et al., 2021; Zultan et al., 2012, for empirical tests of the model). In a voting context, a person would be held less responsible if many of the other votes would have needed to change to make their's pivotal. Chockler and Halpern further propose that the extent to which a person is deserving of blame for a negative outcome depends on their anticipated responsibility. Blame is higher, the more likely it was that a person's action would be pivotal. Recently, Engl (2022) applied Chockler and Halpern's (2004) model to explain how people assign responsibility to players in various economic games.

In the situations discussed so far, each individual's contribution counts the same toward the group outcome. However, this need not be the case (see Gelman et al., 2002). For example, in the United Nations, some countries have more votes than others. Similarly, in presidential elections in the United States, some states cast more votes than others. Our example of the three psychologists in the introduction could be represented by the theorist having two votes while the two experimentalists each have one vote, with three votes being required for a successful outcome. This formulation is structurally equivalent to stating that success requires the theorist and at least one of the two experimentalist. In our experiments, we manipulate the extent to which

[1] Note that this definition of criticality rests on the *homogeneity assumption*, which states that all other players will contribute to the public good with the same probability. See Rapoport (1987) for a definition of criticality that relaxes this assumption and allows different players to have unequal probabilities of contribution.

individuals' contributions affect the group outcome, and investigate how this affects people's judgments of criticality.

Rapoport's (1987) model of criticality, models of a–priori voting power (see, e.g. Felsenthal & Machover, 2004), and Chockler and Halpern's (2004) model of anticipated responsibility all have in common that they consider the probability that a person's action would have been pivotal for the outcome. Accordingly, a person is more critical the more likely they make a difference to the outcome. However, as our experiments below will show, people's intuitive concept of what it means to be critical for an outcome is asymmetric: people care more about a person's contribution being pivotal for a *positive* outcome than a *negative* one. For example, consider a situation in which player A and player B each cast a vote. Player A believes that player B's chance of voting 'yes' is 50%. No matter whether the situation is conjunctive (such that the outcome only happens if both vote 'yes') or disjunctive (such that the outcome happens if at least one of them votes 'yes'), player A's expected pivotality is the same (as is their criticality according to Eq. (1)). In the disjunctive situation, player A is only pivotal if player B votes 'no'. Similarly, in the conjunctive situation, player A is only pivotal if player B votes 'yes'. Because player B is equally likely to vote 'yes' or 'no', player A's expected pivotality is the same. However, as we will see below, people judge player A as being more critical in the conjunctive structure compared to the disjunctive structure.

2. Models of criticality

We consider situations in which team members can either succeed or fail in their individual task, and the team as a whole can either win or lose the team challenge. Team challenges have different causal structures that dictate how individual contributions translate into team outcomes. The top of Fig. 1 shows four different team challenges. Each of the challenges features three players: player A, player B, and player C. In the *disjunctive challenge*, at least one of the players needs to succeed in order for the team to win. In the *conjunctive challenge*, all three players need to succeed. In the *mixed 1 challenge*, player A needs to succeed in addition to at least one out of player B or player C. In the *mixed 2 challenge*, player A needs to succeed, or both players B and player C.

We will call players whose contributions combine in a conjunctive way *complements*, and players whose contributions combine in a disjunctive way *substitutes*. For example, in the conjunctive challenge, player B (or player C) is player A's complement. In the mixed 1 challenge, player C is player B's substitute. We compare participants' criticality judgments against the predictions of five different models (see Fig. 1).

2.1. The heuristic model

We introduced the heuristic model in Lagnado et al. (2013). The heuristic captures the basic intuition that in a conjunctive task, each team member is fully critical, whereas in a disjunctive task criticality diffuses among the members. The model first assigns full criticality to the team as a whole. Conjunctive subunits have the same criticality score as the composite unit. Disjunctive units, on the other hand, receive equally divided shares. This allocation of criticality to subunits repeats recursively until each individual player receives an individual criticality score.

For example, the heuristic predicts that in the disjunctive challenge, each player receives a criticality of $\frac{1}{3}$, as the full criticality is equally shared among the three disjunctive units (see Fig. 1). In the conjunctive challenge, each player is fully critical because the three conjunctive subunits share the same criticality as the whole team does. In the mixed 1 challenge, player A is fully critical, whereas player B and player C, who form a disjunctive unit, each receive a criticality of $\frac{1}{2}$. In the mixed 2 challenge, the heuristic predicts that each player receives a criticality of $\frac{1}{2}$. This is because player A, or players B and C together, form one disjunctive unit so the criticality is divided between these two units. Since player B and player C are complements whose contributions combine conjunctively, each of them receives the same criticality as their composite unit.

The heuristic only considers the causal structure of the challenge but does not take into account how likely each player is to succeed. The following models are sensitive to both the causal structure and the players' likelihood of success.

2.2. The necessity model

The necessity model was also introduced in Lagnado et al. (2013) and adapted from Pearl (1999). The model computes the probability with which each player's contribution is necessary for the team success. The necessity model assigns criticality C_i to player i based on that player's individual outcome $o_i \in \{0 = \text{fail}, 1 = \text{succeed}\}$ and on the team outcome $O \in \{0 = \text{loss}, 1 = \text{win}\}$, as follows

$$C_{i,\text{necessity}} = 1 - \frac{Pr(O = 1|do(o_i = 0))}{Pr(O = 1|do(o_i = 1))}, \tag{2}$$

where $do(o_i = 0)$ considers a situation in which the player failed, and $do(o_i = 1)$ one in which the player succeeded.[2] Intuitively, if a person's contribution makes no difference to the outcome, then the chances of the team succeeding are the same no matter whether the player failed (the numerator) or succeeded (the denominator). In this case, their criticality is 0. If there is no way for the team to win without player i's success (i.e. when the numerator $Pr(O = 1|do(o_i = 0))$ is zero), then the player's criticality is 1.

So, criticality varies from 0 for a player whose individual outcome has no effect on the team outcome to 1 for a player who must succeed for the team to win. If the players' probabilities of success are unspecified, the model assumes equal probabilities for success and failure. As Fig. 1 shows, the predictions of the necessity model align closely with that of the heuristic for the disjunctive, conjunctive, and mixed 1 challenge. The predictions of the two models come apart in the mixed 2 challenge. Unlike the heuristic model, which predicts equal criticality for players A and B, the necessity model predicts that player A is more critical than player B. If player A fails, both players B and C need to succeed in order for the team to win. Whereas when player B fails, only player A needs to succeed (which is more likely than both of the other players succeeding).

The necessity model makes graded predictions that are sensitive to the success likelihood of a player's complement or substitute. The model predicts that as player i's substitute's likelihood of success increases, player i's criticality goes down. Intuitively, if the substitute's probability of success is higher, there is a better chance that the team wins even if player i fails, $Pr(O = 1|do(o_i = 0))$, and hence player i's criticality decreases. The opposite holds for complements. Increasing the success likelihood of player i's complement makes it more likely that i's contribution will make a difference. Our next three criticality models consider the probability with which the target player's action is pivotal for the team outcome. The *credit model* computes the probability of a player being pivotal for a group win, the *blame model* for a group loss, and the *responsibility model* weighs a player's pivotality for a win or loss based on the probability of these outcomes occurring.

[2] Notice that for the structures that we consider, the interventional probability $Pr(O = 1|do(o_i = 0))$ is the same as the conditional probability $Pr(O = 1|o_i = 0)$. However, we chose to use the more general formulation here because it can also be applied to compute the criticality when the probabilities of individual players succeeding are confounded with one another (e.g. due to a common causal factor).

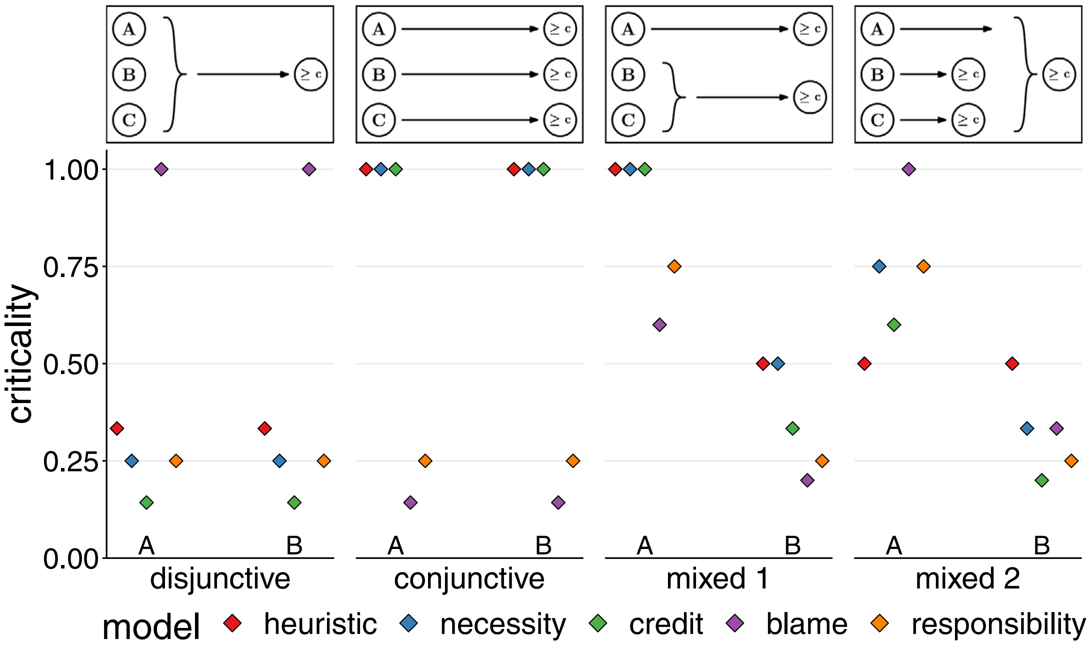


Fig. 1. Criticality model predictions for a selection of team challenges with model predictions. Each panel shows the model predictions for player A on the left side, and player B on the right side. In each team challenge, individual players A, B, and C perform a task in which they can either succeed or fail. We assume here that each player has a chance of $p = 0.5$ to succeed. The team challenge determines how individual performances translate into the team outcome. In the *disjunctive challenge*, at least one of the players needs to succeed in order for the team to win. In the *conjunctive challenge*, all three players need to succeed. In the *mixed 1 challenge*, player A needs to succeed and at least one out of player B or player C. In the *mixed 2 challenge*, player A needs to succeed or both players B and player C.

2.3. The credit model

The credit model predicts that a player is judged to be more critical the more probable it is that their action is pivotal for a team win. The criticality of player i is

$$C_{i,\text{credit}} = Pr(O = 0_{o_i=0} | O = 1), \tag{3}$$

where $O = 0_{o_i=0}$ expresses the counterfactual event of a team loss if player i failed, assuming that in fact the team won ($O = 1$). We adopt this notation for expressing counterfactuals from Pearl (2000) (see also Gerstenberg, 2022). This model computes a player's expected credit, assuming that credit = 1 whenever their action was pivotal for the team win.

As Fig. 1 shows, the credit model and the necessity model make the same qualitative predictions, both between and across the four situations. Both models generally predict that a player's criticality decreases the more likely substitutes are to succeed, and that a player will be judged highly critical if the team cannot win without them. However, as we will see later, the two models make different predictions when a player's probability of success is manipulated. The credit model assigns higher criticality as the probability of a player's success increases. The necessity model, in contrast, only predicts that the success probabilities of the other players affect the criticality of the judged player. How likely the judged player is to succeed does not affect their criticality. Intuitively, this is the case because the necessity model conditions both on what would happen if the judged player succeeded *and* on what would happen if they failed (see Eq. (2)).

Note that quantitatively, there are some differences between the necessity and the credit model even for the situations depicted in Fig. 1. When we fit the models to participants' judgments in the experiments below, we map the model predictions separately onto participants' judgments via a linear regression.

2.4. The blame model

Mirroring the credit model, the blame model predicts that a player's criticality depends on the probability that their action is pivotal for a team *loss*. The criticality of player i is

$$C_{i,\text{blame}} = Pr(O = 1_{o_i=1} | O = 0) \tag{4}$$

where $O = 1_{o_i=1}$ expresses the counterfactual of whether the outcome would have been a win had player i succeeded, assuming that in fact the team failed ($O = 0$).

The blame model predicts that players are more critical in disjunctive structures, and less critical in conjunctive ones. In disjunctive structures, there is only one scenario with a negative outcome, namely when all players failed. In this scenario, each player is pivotal for the loss. So, the model predicts that in disjunctive structures, each player is fully critical for the outcome. In contrast, in conjunctive structures, there is only one scenario with a positive outcome, namely when each player in the team succeeded. But there are many possible ways for the team to fail. Consider a team with three members as in Fig. 1. Here, out of the seven scenarios that lead to a negative outcome in the conjunctive task (all players fail, A and B fail, etc.), there is only one situation in which a player is pivotal for the team's loss. So, according to the blame model, their criticality would be $\frac{1}{7}$. In line with the other models, the blame model predicts that player A is more critical than player B in the mixed challenges. Player A is more likely to be pivotal for team losses than the other two players.

2.5. The responsibility model

The responsibility model computes the expected probability that a player is either pivotal for the team win or pivotal for the team loss. The criticality of player i is

$$\begin{aligned} C_{i,\text{responsibility}} = {} & Pr(O = 0_{o_i=0} | O = 1) \cdot Pr(O = 1) \\ & + Pr(O = 1_{o_i=1} | O = 0) \cdot Pr(O = 0). \end{aligned} \tag{5}$$

The responsibility model is simply a weighted average of the credit model and the blame model, whereby the weighting is determined by the likelihood of a positive or negative team outcome.

The responsibility model can be viewed as an extension of the *probabilistic criticality* model (see Eq. (1)). When the probability of each player in the team succeeding is the same, and when the structure is disjunctive (i.e., $k = 1$), or conjunctive (i.e., $k = n$), these two models make the same predictions. However, the responsibility model also handles situations where the player's probabilities of success within a team vary, and where players have different causal roles (such as in the two mixed challenges in Fig. 1).

The responsibility model predicts that when each player's probability of success is 0.5, then a player's criticality in disjunctive or conjunctive structures is the same (assuming the same team size). For example, in a disjunctive structure with three players, player A is pivotal for a win when both players B and C failed, and pivotal for a loss when all players failed. In the conjunctive structure, player A is pivotal for a win when all players succeeded, and pivotal for a loss when both players B and C succeeded. So for both structures, player A is pivotal for the outcome in two out of eight situations. So as long as each player has the same probability of succeeding or failing, player A's criticality is the same.[3]

Like the other models, the responsibility model also predicts that player A will be judged more critical than player B in the mixed challenges. There are more situations in which player A would be pivotal for the outcome than player B (or player C) would.

3. Re-analysis of Lagnado et al. (2013)

In prior work, we were interested in better understanding how people assign ex-post responsibility to individuals in groups (Lagnado et al., 2013). We argued that when people assign responsibility, they consider both how critical an individual's performance was for the team outcome, and how close their performance ended up to being pivotal. In the experiment reported in the paper, we showed participants eight different challenges and asked them to answer the following question: "How critical is Player A for the team's outcome in each challenge?" We presented the eight challenges in two sets of four whereby each set of challenges was displayed on the same screen with a separate slider underneath each challenge whose endpoints were labeled "not at all" (0) and "very much" (100). The study materials including the raw data, experiment screenshots, and analyses files for all experiments are available here: https://github.com/cicl-stanford/making_a_difference.

The top of Fig. 2 shows the eight challenges for which participants evaluated player A's criticality. The first set of challenges (1–4) manipulates the size of the group (2 or 4 players), and whether individual contributions combine disjunctively (one arrow) or conjunctively (several arrows). For example, challenge 1 features two team members, and the team wins if at least one out of player A and player B succeed in their individual task. Challenge 4 features four team members and the team only wins if all of the players succeed. The second set of challenges (5–8) introduced mixed causal structures. In challenge 5, for example, the team wins if both player C and player D and at least one out of player A or player B succeeds. In challenge 8, the team wins if player A succeeds as well as at least one out of players B, C, or D.

Fig. 2 shows participants' criticality ratings for the eight different causal structures tested in Lagnado et al. (2013) together with the models' predictions. Note that we calibrated each model by fitting it to participants' mean judgments on the eight trials using a linear regression. Table 1 shows each model's raw predictions before it was fitted. For these model predictions, we assumed that each player is just as likely to succeed or fail in their task.[4]

Table 1
Raw model predictions for player A in team challenges with different structures. Fig. 2 shows a visualization of the different structures together with the model predictions that were fitted to the data.

Challenge	Heuristic	Necessity	Credit	Blame	Responsibility
1	0.50	0.50	0.33	1.00	0.50
2	1.00	1.00	1.00	0.33	0.50
3	0.25	0.12	0.07	1.00	0.12
4	1.00	1.00	1.00	0.07	0.12
5	0.50	0.50	0.33	0.08	0.12
6	1.00	1.00	1.00	0.23	0.38
7	0.33	0.25	0.14	0.11	0.12
8	1.00	1.00	1.00	0.78	0.88

Table 2
Model fits for the different studies. *Note*: In the re-analysis of Lagnado et al. (2013) the response scale was from 0 to 100, in Experiment 1 it was from 0 to 20, and in Experiment 2 it was from 0 to 10. rmse = root mean squared error, r = Pearson correlation, n = number of individual participants best fitted by the model. The baseline model just fits an intercept to the data.

Model	Re-analysis			Experiment 1			Experiment 2		
	r	rmse	n	r	rmse	n	r	rmse	n
Heuristic	.97	5.48	9	0.89	1.46	7			0
Necessity	.97	5.85	5	0.98	0.67	9	0.77	0.97	3
Credit	.97	5.44	16	0.97	0.73	4	0.85	0.82	50
Blame	.10	23.22	1	0.32	2.98	0	0.40	1.41	3
Responsibility	.62	18.28	4	0.54	2.65	3	0.36	1.43	9
Baseline		23.32	5		3.14	17		1.54	5

Table 2 shows how well each model captures participants' responses using several evaluation criteria. Three models – the *heuristic model*, the *necessity model*, and the *credit model* – clearly outperform the *blame model* and the *responsibility model*. These three models achieve similar correlations with participants' mean criticality judgments and a similar prediction error. We also computed for each participant which model best fit their judgments based on the sum of squared errors. The *credit model* best fits more participants than the *heuristic* and *necessity models* combined.

The re-analysis of Lagnado et al.'s (2013) shows that only some of the models we introduced above are consistent with how participants evaluate the extent to which an individual is critical for the team's outcome. The *heuristic model*, the *necessity model*, and the *credit model* capture participants' judgments well but the *blame model* and *responsibility model* do not. The *blame model* incorrectly predicts that a player is more critical in a disjunctive than in a conjunctive challenge (see Table 1). Even when fitted to participants' criticality judgments in a way that allows for a negative coefficient on the models' predictions, the *blame model* fails to capture participants' judgments. The *responsibility model* incorrectly predicts that for a given team size, player A would be judged equally critical no matter whether the structure was disjunctive or conjunctive. For example, it predicts the same judgment for player A in challenges 1 and 2, and in challenges 3 and 4.

The *heuristic*, *necessity*, and *credit models* fail to capture one pattern in the data. Player A was judged more critical in challenge 1 compared to challenge 5, and in challenge 2 compared to challenge 6. One possible explanation for the lower criticality in challenges 5 and 6 is that participants' judgments are affected by the size of the team where players in larger teams are considered less critical. However, this does not really hold for the conjunctive challenges 2 and 4. An alternative explanation is that this difference came about as a consequence of the way in which we presented the challenges. Remember that challenges 1–4 were presented on the same screen, and challenges 5–8 on the same screen later in the experiment. The set of challenges to consider on a given screen may have created context effects. It is possible that if these four challenges were presented on the same screen that participants would evaluate player A's criticality to be the same in challenges 1 and 5, and in challenges 2 and 6.

[3] If instead each player's likelihood of success was greater than 0.5, then their criticality would be greater in the conjunctive challenged compared to the disjunctive one. This is the case because now the situation in which all players succeed (where each player is critical in the conjunctive challenge) would be more likely than the situation in which all players failed (where each player is critical in the disjunctive challenge).

[4] Note that while the (unscaled) *blame model* predicts a higher criticality rating in disjunctive structures than in conjunctive structures, the scaled model predicts the opposite because we allowed for the weighting parameter in the linear regression to be negative.

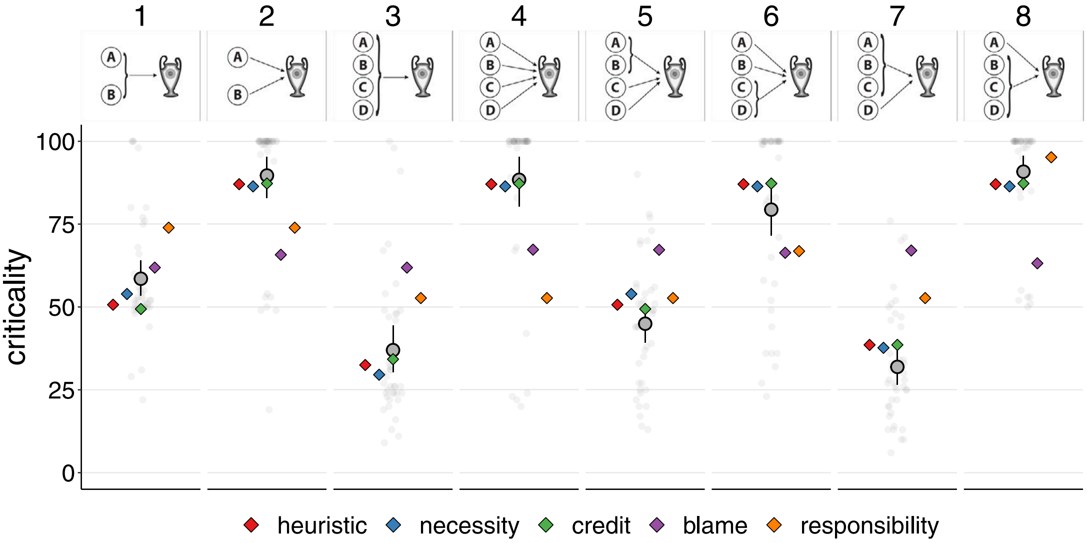


Fig. 2. Re-analysis of Lagnado et al. (2013). Criticality judgments for player A (gray points) and model predictions (colored diamonds). Large points show means with 95% confidence intervals. Small points show individual judgments. The *heuristic model*, *necessity model*, and the *credit model* capture participants' judgments equally well. The *blame model* and *responsibility model* do not capture participants' judgments well. *Note*: The model predictions were scaled via a linear regression from the values shown in Table 1 to the mean criticality judgments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

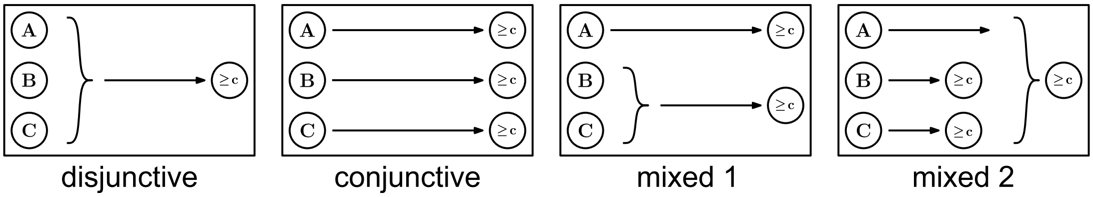


Fig. 3. Experiment 1. Diagrams of the different challenges. In the *disjunctive challenge*, the team wins if at least one of the players succeeds. In the *conjunctive challenge*, the team wins if all three players succeed. In the *mixed 1 challenge*, the team wins if player A succeeds and at least one of players B and C. In the *mixed 2 challenge*, the team wins if player A succeeds, or if both players B and C succeed.

Overall, while participants' criticality judgments are consistent with the *necessity model* and the *credit model*, it is also possible that they followed the simple *heuristic model* according to which a player is critical if their success is necessary for the team win, and where their criticality reduces with the number of players if they are part of a disjunctive (sub-)structure.

4. Experiment 1: Criticality in novel scenarios

Experiment 1 tests the different models' predictions against people's criticality judgments for a new set of causal structures (see Fig. 3). This time, each challenge featured three players, including a disjunctive challenge, a conjunctive challenge, a mixed challenge in which player A and at least one out of player B or player C needed to succeed in order for the team to win ("mixed 1"), and another mixed challenge where either player A, or both players B and C had to succeed for the team to win ("mixed 2").

For the mixed 2 challenge, the predictions of the *heuristic model* come apart from those of the *necessity model* and the *credit model*. Here, the *heuristic model* predicts that player A and player B are equally critical. It predicts that criticality is shared equally between players who form a disjunctive unit, and that players in a conjunctive unit do not share the criticality. In this case, player A and players B and C form a disjunctive unit, so the criticality is shared between them (each unit receiving a criticality of 0.5). Since players B and C form a conjunctive unit, the criticality of 0.5 for this unit is not shared, so that each individual player of that unit receives a criticality of 0.5

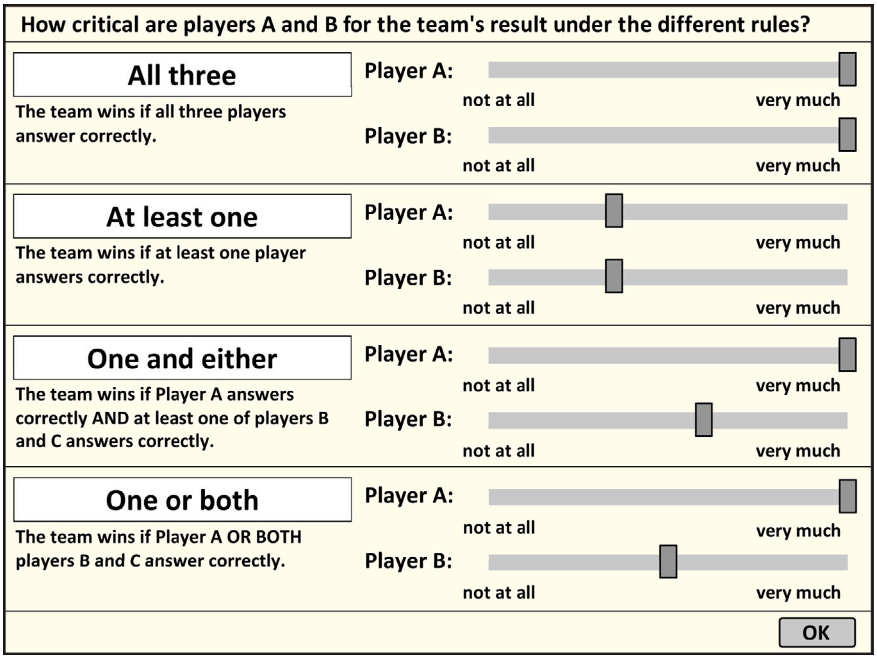
In contrast, both the *necessity model* and the *credit model* predict that player A is more critical than player B. To get the intuition for why the *credit model* predicts this, remember that this model considers how likely a player's success is pivotal for the team's win. Player A is pivotal when both player B and C fail, or when either of them succeeds (but not both). However, player B is only pivotal if player A fails and player C succeeds. So from the six possible situations that would result in a team win, player A is pivotal in three of them, whereas player B is only pivotal in one of them. Fig. 1 shows the model predictions for player A and player B in the four different challenges (assuming that each player is just as likely to succeed or fail in their task).[5] Notice that while there are some quantitative differences between the raw predictions of the *necessity model* and the *credit model*, their predictions are extremely highly correlated (r = .99). So while the results of this experiment can shed light on whether participants are using the *heuristic model* to judge criticality, they will not help tease apart the *necessity* and *credit models*.

4.1. Methods

4.1.1. Participants

Forty participants (*age*: M = 36, SD = 13; *gender*: 21 female, 19 male) were recruited online via Amazon Mechanical Turk and received $1 compensation.[6]

[5] The prediction that player A's criticality is greater than that of player B or C holds as long as player A's probability of success is not substantially lower than that of players B and C. In the experiments, we told participants that players were randomly assigned to the three different roles, so there is no



How critical are players A and B for the team's result under the different rules?

All three

The team wins if all three players answer correctly.

At least one

The team wins if at least one player answers correctly.

One and either

The team wins if Player A answers correctly AND at least one of players B and C answers correctly.

One or both

The team wins if Player A OR BOTH players B and C answer correctly.

Fig. 4. Experiment 1. Screenshot of the main task in which participants were asked to judge how critical player A and player B are for the team's result under different rules.

4.1.2. Design & procedure

The experiment was programmed with Adobe Flash. Participants were told that they would act as external observers asked to comment on a team challenge in a game show. Their task was to indicate how critical each player in the team will be for the team's outcome in the challenge. The game show has teams of three players competing for a prize. In the first round, each player answers one general-knowledge question. In order for the team to win and pass the first round, one or more players must give the correct answer.

Participants were then introduced to four versions of the game show that differed in which players, and how many of the players, must give correct answers in order for a team to win and pass the first round. Participants learned that players in the game show are randomly assigned to the roles A, B, and C. The four different challenges were labeled "At least one", "All three", "One and either", and "One or both" in the order as shown at the top of Fig. 5. Participants had to hover with the mouse over each of the labels to read a description of what it takes for the team to win that challenge.

Participants then completed a comprehension check phase in which they had to say for each of the four challenges whether the team would win or fail if

1. Player A answers correctly, and players B and C give a wrong answer;
2. Player B answers correctly, and players A and C give a wrong answer;
3. Player A and player B answer correctly, and player C gives a wrong answer;
4. Player A and player B and player C answer correctly.

Only participants who answered all of the comprehension check questions correctly were able to proceed to the main task. If a participant answered one ore more questions incorrectly, they were redirected to read the instructions again. Before the main task, participants received one reminder that they will be asked to indicate how critical each player will be for the team's result in the first round of the game show under the different rules.

Fig. 4 shows a screenshot of the main task. All of the four challenges were presented on the same screen. Each challenge was indicated by its label together with a short description of the rule. The vertical order in which the different challenges appeared on the screen was held constant across participants with the "All three" challenge at the top followed by the "At least one", "One and either", and "One or both" challenges. At the top of the screen was the question: "How critical are players A and B for the team's result under the different rules?". For each of the four rules, there was a separate slider for player A and for player B whose endpoints were labeled "not at all" and "very much". The sliders corresponded to 21-point Likert scales (0 = not at all, 20 = very much), and were initialized at 0. Each participant provided eight judgments, judging players A and B in the four different challenges. To proceed to the next screen of the experiment, participants' had to click on each of the sliders (but they did not have to move them).

After this main task, participants were asked to describe in a text box how they assessed the criticality of the players. They were also asked to share any other thoughts they had on the experiment. Finally, we asked participants for their age and gender.

4.2. Results

Fig. 5 shows participants' criticality judgments together with the predictions of the *heuristic model*, the *necessity model*, and the *credit model*. As mentioned earlier, the predictions of the *necessity model* and the *credit model* are virtually identical for the cases considered here.

reason to assume that a player's role reveals anything about their likelihood of succeeding.

[6] The key comparison in this experiment is the predicted difference between player A's criticality and player B's criticality in the mixed 2 challenge. We performed a power-analysis to determine the sample size for this experiment. We anticipated an effect size of $d = 0.5$ for this difference in a paired-samples t-test, which for a desired power of 0.8 at an alpha level of 0.05 (two-sided), yields a sample size of 34. In fact, we observed an effect of $d = 0.76$, which for our sample size of 40 yields a power of 0.997.

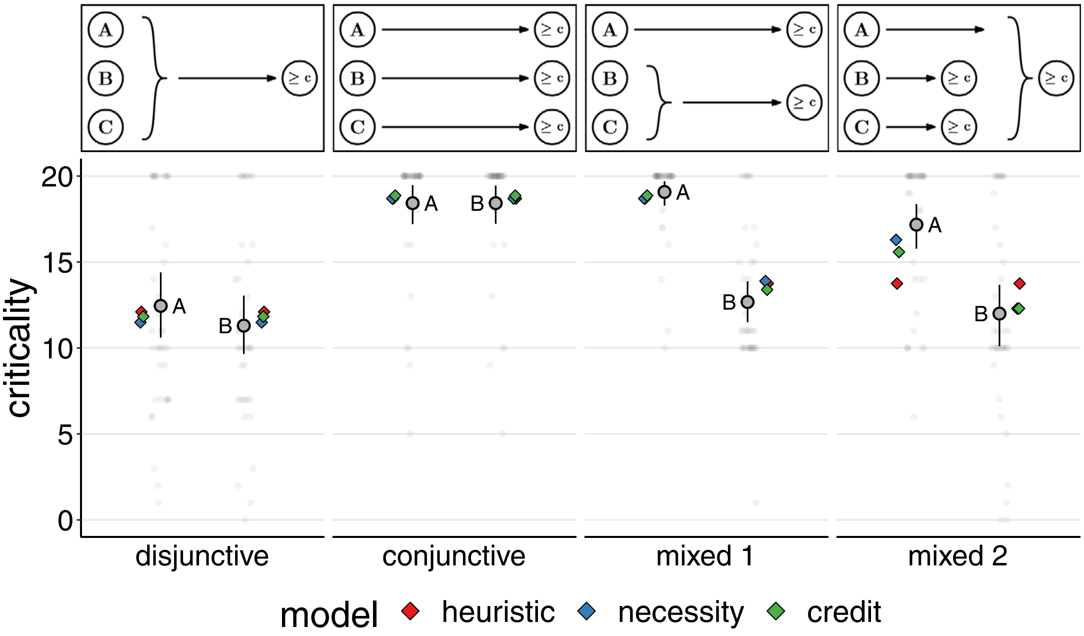


Fig. 5. **Experiment 1**: Criticality judgments (gray points) and model predictions (colored diamonds). Large points show means with 95% confidence intervals. Small points show individual judgments. The model predictions are jittered along the x-axis for visibility. Participants were asked to judge both player A and player B in each of the situations. While the *heuristic model* (red) predicts the same judgments for player A and player B in the mixed 2 challenge, the *necessity model* (blue) and the *credit model* (purple) correctly predict that player A will be seen as more critical than player B in this challenge. *Note*: The model predictions were scaled via a linear regression from the raw predictions shown in Fig. 1 to participants' mean criticality judgments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Both of these models predict that player A would be judged more critical than player B in the mixed 2 challenge. The *heuristic model*, in contrast, predicts that both players would be judged equally critical. In line with the predictions of the *necessity model* and the *credit model*, player A was judged more critical.

In the mixed 1 challenge, all three models correctly predicted that player A would be judged more critical ($\mathrm{M} = 19.08, \mathrm{SD} = 2.38$) than player B ($\mathrm{M} = 12.68, \mathrm{SD} = 4.10$; mean and 95% credible interval of the posterior distribution of the difference $= 6.41$ $[4.97, 7.81]$).[7] 31 out of 40 participants gave a higher rating to player A than player B.

In the mixed 2 challenge, only the *necessity model* and the *heuristic model* correctly predicted that player A would be judged more critical ($\mathrm{M} = 17.18, \mathrm{SD} = 4.16$) than player B ($\mathrm{M} = 12.00, \mathrm{SD} = 5.91$, difference $= 5.2$ $[2.94, 7.37]$). 24 out of 40 participants gave a higher rating to player A than player B.

We can also compare judgments for the same player between challenges. Here, all three models correctly predict that player A would be judged more critical in the mixed 1 challenge than in the mixed 2 challenge (difference $= 1.91$ $[0.44, 3.34]$). Finally, the models correctly predicted that that player B was judged similarly critical in the mixed 1 challenge and the mixed 2 challenge (difference $= 0.7$ $[-1.43, 2.83]$).

Table 2 shows how well each of the models captures participants' mean judgments, and how many participants are best fitted by each model. The *necessity model* and the *credit model* better account for participants' mean judgments (both in terms of correlation and root mean squared error) than the *heuristic model*. The *necessity model* best accounts for the largest number of individual participants, followed by the *heuristic model*, and the *credit model*.

4.3. Discussion

Experiment 1 set out to test the *heuristic model* against the *necessity model* and *credit model*. To pit the models against one another, we introduced a new challenge where the team only wins if either player A, or both players B and C succeed. According to the *heuristic model*, player A and player B are equally critical in that challenge. According to the other two models, player A is more critical than player B. The results clearly show that participants viewed player A as more critical. The *necessity model* and the *credit model* accurately capture participants' criticality judgments across all of the challenges. None of the other models, including the *heuristic model*, were able to account for participants' judgments as well.

5. Experiment 2: Criticality with probabilities

Based on the re-analysis of Lagnado et al. (2013), we found that only the *heuristic model*, the *necessity model*, and the *credit model* adequately capture participants' criticality judgments. Experiment 1 presented evidence against the *heuristic model*. Experiment 2 pits the *necessity model* against the *credit model*.

To do so, we manipulated information about the probability with which each player was likely to succeed in their task. When computing criticality, the *necessity model* conditions separately on the success and failure of the player of interest. This means that, according to this model, a player's probability of success does not affect their criticality. How critical a player is only depends on the causal structure, and the success probability of their teammates. In contrast, according to the *credit model*, a player's success probability affects their criticality. Specifically, if a player forms part of a disjunctive (sub-)structure, than their criticality increases the more likely they are to succeed.

For example, let us consider the mixed 1 challenge where in order for the team to win, player A needs to succeed and at least one out of players B and C. Let us assume that player A has a 70% chance of succeeding and player B has a 30% chance. Now let us compare two situations: one in which player C's likelihood of success is 10% and another in which it is 90%. The *necessity model* predicts that player C's criticality is the same in both situations. The *credit model*, in contrast, predicts that player C's criticality is greater when their chance of success if 90% compared to 10%. To understand why this is the case, remember that the *credit model* considers the probability that a player

[7] We follow the convention of calling something an effect when the 95% credible interval of the parameter of interest excludes 0.

is pivotal for the team's win. Player C is pivotal for the success in a situation in which both player A and player C succeed, but player B fails. Because this situation becomes more likely as player C's chances of success go up, their criticality increases. By manipulating both the causal structure of the challenge and the prior information about each player's probability of success, Experiment 2 allows us to tease apart the predictions of the *necessity model* and the *credit model*.

5.1. Methods

5.1.1. Participants

Seventy (*age*: M = 19, SD = 1; *gender*: 53 female, 17 male) University College London first-year undergraduate students participated in the experiment for course credit.[8]

5.1.2. Materials and procedure

The experiment was programmed using Adobe Flash and run as part of a lab class. Participants were seated in individual cubicles in a computer cluster room. All instructions appeared on screen. Participants learned that their role in the experiment would be that of an external observer. Their tasks were (i) to make predictions about whether different teams in a hypothetical game show would succeed in their team challenge and (ii) to indicate how critical they perceived each player to be for the team's success in the challenge (see Fig. 6). We made it clear to participants that their criticality evaluations would be made *before* the event has taken place, and that they will not be seeing the teams' outcomes in the challenge.

The game that each player in the game show played was called the "dot-clicking game". In this game, a dot reappears in a random location on the screen each time after it is been clicked on. A player succeeds in the game if they manage to click the dot often enough within a limited time period. To get a better sense of how the game works, participants played it themselves for four trials. Afterwards, we told participants that the players in the game show played a different version of the game in which the players had more time, the dot was smaller, and the criterion for a successful trial was 20 clicks.[9]

The hypothetical game show had two parts. In the first part, players practiced the dot-clicking game individually for ten trials. We manipulated participants' expectations about each player's success by varying their practice performance. While some players reached the criterion of 20 clicks in most of their ten practice trials, others only succeeded rarely. Fig. 6 shows an example where the three players succeeded in 30% of their practice trials. We told participants that a player's performance in the practice reflects how likely they will succeed in the upcoming team challenge. In the second part, players were randomly assigned to teams and positions in one of three challenges. Each challenge featured three players and was either disjunctive, conjunctive, or mixed 1 as shown in Fig. 3.

In each trial of the main task, participants first received information about the percentage of trials in which each team player had succeeded in the practice (shown in the top right of each screenshot in Fig. 6). Players were indicated by initials, and no initials were repeated to make it clear that a new set of players participated in each challenge. Participants then clicked to reveal which of the three challenges the team had been assigned to (shown in the top left in Fig. 6). Afterwards, participants answered the question "How high are the chances that the team is going to win the challenge?" by typing in a number between 0 and 100% in a text box (see Fig. 6a). The inclusion of this prediction task made sure that participants processed both the information about the players' prior performance as well as what the team challenge they had been assigned to. Lastly, participants used sliders to answer the question "How critical is each player for the team's result in this challenge?" (see Fig. 6b). There was a separate slider for each player on the team. The slider endpoints were labeled "none" (0) and "very much" (10).[10] Sliders were initialized at 0 and each slider had to be clicked on (or moved) to be able to continue. Participants were able to remind themselves about how everything worked by moving the mouse over an info button on the top right of the screen which remained available throughout the experiment.

Participants first completed one practice trial and then answered six forced-choice comprehension questions. They received immediate feedback about whether their answer was correct or wrong. The correct answer was provided in written form. Overall, participants answered 88% of the questions correctly. Participants then completed 24 trials that were presented in random order. After this main part of the experiment, participants answered a few more questions that are not relevant here (e.g. we asked participants to assign responsibility to individual players based on whether or not they succeeded in their task). At the end of the experiment, participants were prompted to provide their demographic information and asked to write in a few sentences about how they assessed the players' criticality in the experiment. It took participants on average 7.6 (SD = 1.87) minutes to go through the instructions and 23.4 (SD = 4.89) minutes to complete the entire experiment.

5.1.3. Design

Both (i) team structures and (ii) information about previous performance of the players were varied within participants. The x-axis of Fig. 7 shows the 8 different patterns of performance in the practice trials. There was a set of patterns in which all three players had identical prior performance (3 3 3 and 7 7 7). In two more sets, the priors for players A and B were held constant while the prior of player C was varied (e.g. the prior for player A was 7, the prior for player B was 3, and the prior for player C was either 1, 5, or 9).[11] As dependent variables, we assessed participants' estimates of the winning chances for each team as well as the criticality ratings for each player in the team.

5.2. Results

For each challenge, we first asked participants to evaluate how high the chances are that the team is going to win the challenge (see Fig. 6a). We asked this question to ensure that participants paid close attention to the team challenge, as well as to the information about the performance of each player on the practice trials. Participants' probability judgments are shown in Fig. C.1 in the appendix. Probability judgments changed as expected with the manipulated skill of the players. And, as prior work has shown (e.g. Bar-Hillel, 1973), participants tended to underpredict a team's probability of success for disjunctive challenges, and overpredict for conjunctive challenges.

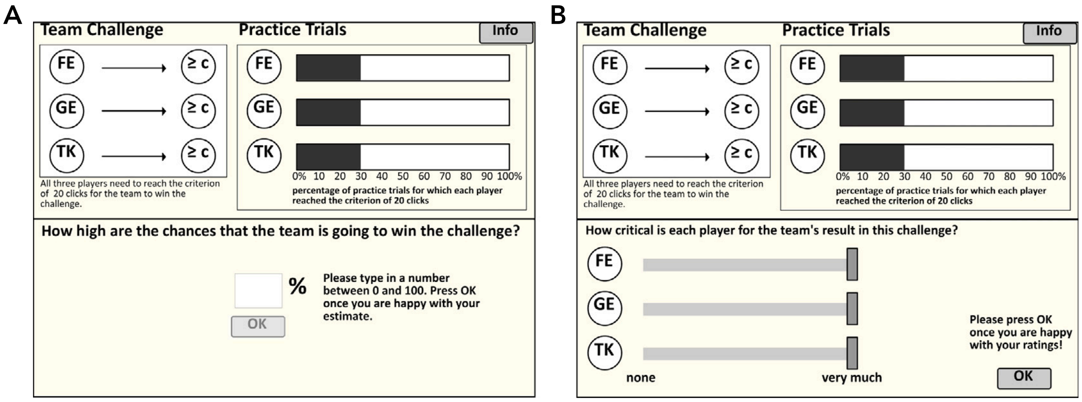
Fig. 7 shows participants' criticality judgments for the three different challenges and the different player skill levels together with the predictions of the *necessity model* and the *credit model*. The results show that participants judgments were sensitive both to the structure as well as the information about how likely each player would succeed in their individual task.

[8] The class size of the students who participated in the experiment was $N = 131$. 70 participated in the criticality experiment that we focus on here, and 61 participated in the effort experiment that we mention in the General Discussion. The results of the effort experiment are shown in Fig. E.1 in the appendix.

[9] This was done to render participants' performance in their trials uninformative about the difficulty of the task in the game show.

[10] The labeling of one of the endpoint as "none" instead of "not at all" was due to a programming error. We do not believe that this affected participants' responses.

[11] The priors were randomly assigned to the 3 players in the disjunctive and conjunctive team challenges. The matching of priors to players was fixed for the mixed challenge in the same way as shown in the table.



A

Team Challenge

All three players need to reach the criterion of 20 clicks for the team to win the challenge.

Practice Trials

percentage of practice trials for which each player reached the criterion of 20 clicks

How high are the chances that the team is going to win the challenge?

Please type in a number between 0 and 100. Press OK once you are happy with your estimate.

B

Team Challenge

All three players need to reach the criterion of 20 clicks for the team to win the challenge.

Practice Trials

percentage of practice trials for which each player reached the criterion of 20 clicks

How critical is each player for the team's result in this challenge?

Please press OK once you are happy with your ratings!

Fig. 6. Experiment 2. Screenshots of the main task: (a) probability judgment phase, (b) criticality judgment phase. For each team challenge, participants were first asked to estimate how high the chances are that the team is going to win the challenge. This probability judgment phase made sure that participants pay attention to the structure of the team challenge, and the performance of each player on the practice trials. In the criticality judgment phase, participants judged how critical each player is for the team's result in this challenge.

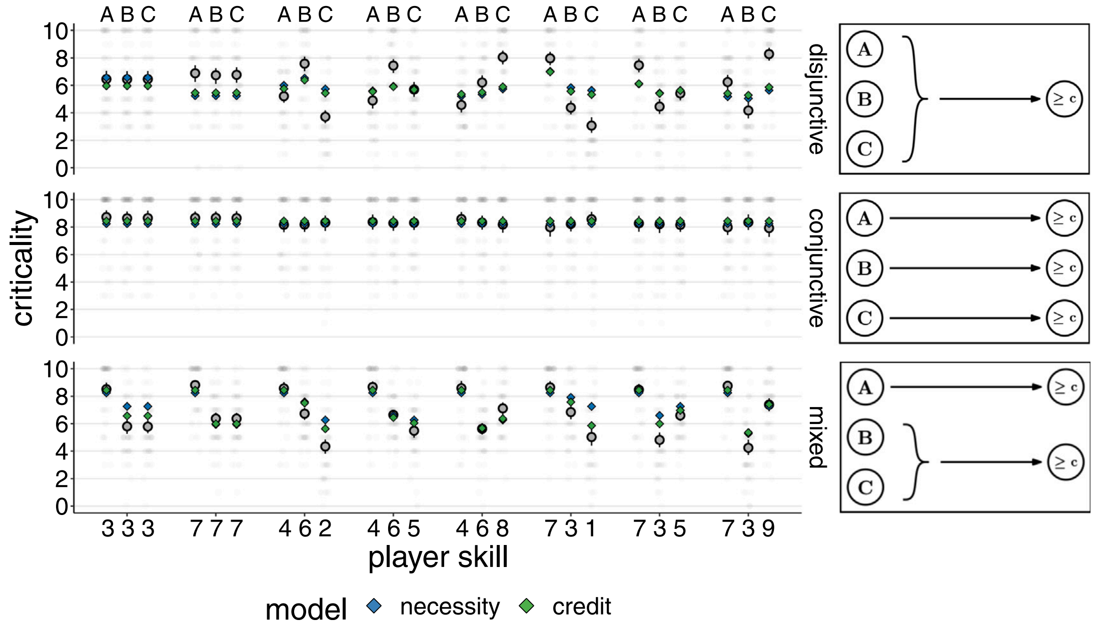


Fig. 7. Experiment 2: Criticality judgments (gray points) and model predictions (colored diamonds). Large points show means with 95% confidence intervals. Small points show individual judgments. Participants were asked to judge how critical each player A, B, and C were for each of the three different challenges. We also manipulated each player's skill (i.e. how many out of 10 times they succeeded in the practice phase). Player skills are shown on the x-axis. The *necessity model* predicts that a player's criticality depends on the structure and the skill of the other players (but not on the skill of the player themselves). For example, consider player C in the last three trials of the mixed challenge where player A's skill is 7 and player B's skill is 3. Here, the *necessity model* predicts that player C's criticality is the same no matter whether their skill is 1, 3, or 9. In contrast, the *credit model* correctly predicts that player C is judged more critical as their skill increases. *Note*: The model predictions were scaled via a linear regression from the values shown in Table B.1 to the mean criticality judgments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the disjunctive challenge, participants generally judged players more critical who were more likely to succeed. In the conjunctive challenge, each player was judged to be highly critical no matter what their skill level was. Finally, in the mixed challenge, player A was judged as highly critical, and for player B and player C, their criticality increased with their skill level.

Two further trends are apparent for players in disjunctive challenges. First, as the skill of one player increases, the criticality of the other players goes down. For example, consider the criticality of player A in the disjunctive challenges in which their skill is 7. As player C's skill increases from 1 to 5 to 9, the judged criticality of player A goes down. Second, a player's criticality increases with their skill level. For example, in the mixed challenge, player C's judged criticality increases as their skill level increases from 1 to 5 to 9. To confirm this observation, we ran a Bayesian linear mixed effects model with player C's skill levels as predictor. For this analysis, we restricted the set of situations to ones in which player C forms part of a disjunction (so we excluded the conjunctive challenge), and for which the skill level of player C changes while the skill of the other two players was held constant (so we excluded the "3 3 3" and "7 7 7" situations). As expected, there was a positive effect of player C's skill on how critical the player was judged (posterior mean = 0.43, 95% credible interval = $[0.36, 0.49]$).

5.2.1. Model comparison

Table 2 shows how well each of the different models captured participants' criticality judgments. Overall, the *credit model* performed best. It achieved the highest correlation, lowest error, and accounted best for a large majority of individual participants (50 out of 70). The *necessity model* only accounted best for 3 out of 70 participants. Qualitatively, the *credit model* captures almost all of the trends in the data. It correctly predicts that criticality goes up with skill in the disjunctive challenge, and is high irrespective of skill in the conjunctive challenge. It further predicts that as one player's skill goes up in disjunction, the other players' criticality goes down. Both of these effects are also predicted by the *necessity model*. However, only the *credit model* correctly predicts that a player's criticality goes up in disjunction as their own skill increases. As we discussed earlier, the *necessity model* does not predict this effect because it conditions on the target player's individual outcome.

Quantitatively, both the *necessity model* and the *credit model* underpredict how much participants' criticality judgments vary with the players' skills in the disjunctive challenge. Both models also incorrectly predict that each players' criticality would be judged lower in the "7 7 7" compared to the "3 3 3" situation. In contrast, participants' assign roughly the same degree of criticality to each player in both of these situations.

Even though the *responsibility model* does a relatively poor job of capturing participants judgments overall, there were 13 out of 70 participants whose judgments were best predicted by this model. The *responsibility model* predicts that a player's criticality in conjunctive challenges decreases with their skill level, and that it increases as the skill level of the other players increases (see Fig. D.1 in the appendix, as well as Table B.1 for the predictions of all the models).

5.3. Discussion

In this experiment, we manipulated both causal structure as well as information about how likely each player in the group was to succeed. Manipulating success probabilities served two objectives. First, we show how the models can incorporate probabilistic information and make predictions that align with human responses. Second, by manipulating each player's probability of success, we were able to create situations for which the *necessity model* and the *credit model* make different predictions. Only the *credit model* correctly predicts that a person in a disjunctive (sub)structure is judged more critical as their probability of success increases. Because the necessity model conditions both on the failure and success of the player of interest (see Eq. (2)), it is insensitive to the probability of success.

While the *credit model* accounted for much of the variance in participants' judgments, it underpredicted how strongly criticality judgments were affected by players' likelihood of success in disjunctive structures. It also makes the incorrect prediction that each player will be judged as less critical in disjunctive structures when all of the players' probability of success is high rather than low.

6. General discussion

We started this paper with a motivating scenario in which three psychologists aim to submit a paper in time to make the journal deadline. One is responsible for writing the theoretical model, and the other two are running separate experiments. A successful submission requires a theoretical model plus at least one experiment. How critical is each psychologist for making the deadline? We developed a number of computational models that make different predictions about how criticality is evaluated, and tested the models' predictions against participants' judgments. Across experiments, we manipulated the causal structure of the situation that dictates how individual contributions combine to determine the group outcome, as well as the probability with which different players in a team were likely to succeed.

In a re-analysis of the data from Lagnado et al. (2013), we found that participants' criticality judgments were consistent with three of the five models we considered: a *heuristic model* that assigns full criticality to players whose contributions are necessary and divides up the criticality between players whose contributions combine disjunctively; a *necessity model*, which considers how a player's individual's success contributes to the collective outcome; and the *credit model*, which computes the probability that a player's contribution is pivotal for a positive outcome. Models based on anticipated blame that a player would receive for a negative outcome or on anticipated pivotality for any outcome did not predict participants' judgments well. While prior work conceptualized the notion of criticality in terms of the probability that the agent's action would make a difference to the outcome (e.g. Au et al., 1998; Engl, 2022; Rapoport, 1987), we found that this model does not account well for participants' judgments. Instead of considering how pivotal a player would be for positive or negative outcomes, participants' specifically care about what role each player plays in bringing about a positive outcome.

A simple example for where the two accounts come apart is a scenario in which two players form a team and where each player has a 50% of succeeding in their task. Accounts that construe criticality as the probability of being pivotal predict that a player is equally critical no matter whether individual contributions combine disjunctively or conjunctively. In a disjunctive scenario, a player is pivotal when the other player failed (and there is a 50% of that happening). In a conjunctive scenario, a player is pivotal if the other player succeeded (which, again, will happen with 50% probability). However, people judge that players are more critical in conjunctive compared to disjunctive scenarios. This effect of causal structure is predicted by the *credit model* of criticality. The *credit model* looks at pivotality for positive outcomes only. In a conjunctive scenario, each player is pivotal when the outcome is positive (irrespective of the group size and each player's probability of success). In a disjunctive scenario, a player's criticality increases the greater their probability of success is compared to their partner's. This is because in a disjunctive scenario, a player is only then pivotal for the win if all the other players failed their task.

Experiment 1 replicated the results from the Lagnado et al. (2013), and introduced a novel group structure for which the *heuristic model* makes different predictions from those of the necessity model and the anticipated credit model. The results lined up with the predictions of the latter two models, and provided evidence against the *heuristic model*. In Experiment 2, we manipulated each player's likelihood of success. The necessity model predicts that a player's criticality depends on the causal structure of the situation, the likelihood of their team members' success, but *not* on the likelihood of success of the player under consideration. In contrast, the anticipated responsibility model predicts that a player's success probability matters, too. Specifically, if a player's contribution combines disjunctively, then criticality is predicted to increase with the player's likelihood of success. Participants' criticality judgments in Experiment 2 were better accounted for by the *credit model*. While the model accurately captured most of the qualitative patterns in participants' judgments, it incorrectly predicted that in a purely disjunctive structure, participants would see each player as more critical when their likelihood of success was low compared to when it was high. In contrast, participants made very similar criticality judgments in both cases.

Much research in the past has demonstrated how responsibility serves a *backward-looking* function by identifying targets for blame and praise (Alicke, 2000; Shaver, 1985; Weiner, 1995). Here, we provide evidence that responsibility serves a *forward-looking* function as well. Considerations about how much prospective responsibility individuals have for their group's success are closely related to perceptions of criticality.

6.1. Criticality and causal judgments

Our results show that people's criticality judgments are sensitive to the causal structure of the situation. Recent work has shown that information about the causal structure of a situation and how likely each event will happen not only affects people's judgments of responsibility (Gerstenberg & Lagnado, 2010, 2012; Koskuba et al., 2018; Lagnado & Gerstenberg, 2015; Lagnado et al., 2013; Zultan et al., 2012) but also their causal judgments (Gerstenberg & Icard, 2020; Henne, Kulesza, Perez, & Houcek, 2021; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Quillien & Lucas, 2022). Earlier work showed that when multiple causes contribute to some outcome, people often tend to give higher causal ratings to abnormal or unexpected events compared to normal or expected ones (Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; McGill & Tenbrunsel, 2000). More recently, a number of studies have demonstrated an intriguing interaction between causal structure and normality on people's causal judgments. While for conjunctive structures, people give higher causal ratings to abnormal events, for disjunctive structures, people judge normal events as more causal (Gerstenberg & Icard, 2020; Icard et al., 2017).

Icard et al. (2017) developed a model that accounts for this pattern of results by assuming that people consider both whether an event was necessary in the actual situation, and whether that event is generally sufficient for bringing about the outcome. Quillien and Lucas (2022) show that this pattern of causal judgments can also be explained by assuming that people imagine counterfactual possibilities that are both a-priori likely and similar to what actually happened, and then check to what extent the cause and effect are correlated across these counterfactual possibilities. While these two accounts explain the pattern of causal judgments as arising from people's tendency to consider certain counterfactuals, Kirfel, Icard, and Gerstenberg (2022) argue that the results are also consistent with the idea that people favor causes that would make for good interventions (see also Fazelpour, 2021; Lombrozo, 2010; Morris et al., 2018; Woodward, 2006). In a similar vein, considerations of criticality are important for making good decisions, such as assigning players with different abilities to different roles on a team.

6.2. From criticality to effort

There is a tight link between perceived criticality and the motivation to act (Weiner, 1986, 2011). Generally, people are motivated when they expect their action to make a difference (Bandura, 1977). In another condition of Experiment 2, we had asked a separate group of participants how much effort they expect each player in the team to exert. The results are shown in Fig. E.1 in the appendix. The effort judgments closely mirrored the criticality judgments. Generally, the more critical a player was judged to be, the more effort they were expected to exert. However, there was one subtle difference between the judgments: when judging criticality, participants gave very high ratings in the conjunctive structure irrespective of the players' likelihood of success. In contrast, when judging effort for conjunctive challenges, participants expected lower skilled players to put in more effort than higher skilled players. This subtle effect is consistent with the idea that when judging how much effort they expect an agent to exert, participants consider the probability that their action will make a difference to the outcome more generally, instead of only considering whether their action would be pivotal for a positive outcome.

This effect is in line with prior work that has found that people modulate their effort as a function of the group structure and their partner's likelihood of success (Kerr & Bruun, 1983). In disjunctive tasks, participants with *low* ability reduced their efforts whereas in conjunctive tasks, participants with *high* ability tried less hard. Kerr and Bruun (1983) hypothesized that motivation was linked to how dispensable one's efforts was perceived to be. When participants were asked after the experiment to rate how much they thought the group success depended on their own performance, their ratings qualitatively matched how much effort they exerted. More able participants felt more indispensable in disjunctive tasks compared to conjunctive tasks. Conversely, weaker participants rated themselves more indispensable in conjunctive as opposed to disjunctive tasks.

The close link between people's perception of criticality and their willingness to exert effort has implications for the effective design of teams and institutional structures. On the one hand, disjunctive redundancies are important because they increase robustness. A positive outcome can still be reached even if not everyone in the group did a great job. On the other hand, such redundancies are likely to lead people to think of their contributions as not being critical, thereby diminishing their motivation to work hard (Falk et al., 2020; Falk & Szech, 2013; Wu & Gerstenberg, 2023). In fact, many problems we face as society today, such as the problem of global warming, are problems of perceived criticality. Even though the problem is caused by the collective of people, individuals do not believe that they can make a difference to the outcome, and so are not motivated to put in the effort that is needed to change things (Friedenberg & Halpern, 2019).

6.3. Why the focus on positive outcomes?

While prior work has defined criticality as the probability that a person's action will be pivotal for the outcome (see, e.g., Eq. (1)), we found that participants' criticality judgments are best explained by a model that assumes that people focus on the pivotality for positive outcomes specifically. Why this focus on positive outcomes?

One possibility is that participants interpreted our experimental question to be about positive outcomes. In our experiments, we made sure not to bias the way in which we asked participants about each player's criticality. We asked "How critical is player A for the group outcome?" instead of asking "How critical is player A for the team's success?". However, because we focused on achievement settings, it is possible that participants interpreted the question in this way. That said, for the prior research that looked at individual decisions in public goods games, there is also a clear sense in which the provision of the public good is a positive outcome. It is possible that in a setting in which actions and outcomes are more symmetric (e.g. deciding between two outcomes that each have their pros and cons), judgments of criticality align more closely with anticipated responsibility. When it is not clear whether an outcome is positive or negative, the notions of credit or blame do not apply.

Another potential justification for judging criticality the way our participants did comes from the fact that attributions of responsibility and criticality are subject to uncertainty. For example, two external observers who differ in their expectations about the group members' performance or in their understanding of the way in which the individual contributions combine to determine the group outcome will likely arrive at different judgments of criticality. Although in our experiments we provided people with all the relevant information, matters are often much more complex and uncertain in the real world. Consider a situation in which we know that the positive outcome happened, but we do not know what each person in the team did. From the positive outcome, we can infer that each conjunctive player succeeded. We can also infer that for disjunctive players, more skilled players were more likely successful than less skilled players. Participants' criticality judgments in our experiments are thus consistent with the idea that people consider which players most likely would have succeeded in their task (and were thus deserving of credit) if all they knew was that the outcome was positive.

6.4. Limitations and future directions

We explored participants' criticality judgments in a relatively small set of possible situations. For example, our group sizes only varied between two and four members, and we only tested a subset of possible causal structures that dictate how individual contributions translate into group outcomes. Future work needs to look at how well the *credit model* generalizes to other situations. Furthermore, we focused on situations in which each person only takes a single action. It would be interesting to explore how people assess the criticality of different agents when they take multiple actions (see Engl, 2022).

In our setting, each player's actions are independent from one another. While settings like this exist in the real world (e.g. when different judges evaluate sports performances such as in athletics), group members more commonly affect one another directly. What one person does will make a difference to how others will act subsequently. Again, future work is required to better understand how people judge criticality in more dynamic contexts like these.

The steps that our models go through to determine criticality are relatively complex. For example, the *credit model* assesses the probability of the different possible situations, and computes criticality by considering the chances that a person's action would be pivotal for a positive outcome. We do not see our model as a process model detailing the underlying cognitive computations that people carry out when making their judgments (Marr, 1982). Rather, we see our model as an as-if model that adequately captures what factors are important for people's intuitions about criticality (Berg & Gigerenzer, 2010). Namely, the performance expectations of the judged player, the expectations of the other players, and the task structure. Our main claim is that responsibility and criticality are closely intertwined, and that responsibility may be the more foundational concept which can be used to define what it means to be critical. More work is required to better understand the underlying cognitive processes by which people arrive at their criticality judgments.

7. Conclusion

People want to make a positive difference. Much prior work in psychology has focused on how people assign responsibility for events that already happened. However, there is a also a sense in which people have prospective responsibilities: depending on their causal role and ability, their actions can be more or less critical for making a difference. Looking back, making a difference makes you responsible. Looking ahead, the potential to make a difference is no less important — especially the potential to make a difference for the better. We believe that both perspectives on responsibility are important and encourage more work on the forward-looking function of responsibility.

CRediT authorship contribution statement

Tobias Gerstenberg: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **David A. Lagnado:** Conceptualization, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Ro'i Zultan:** Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Table A.1
Re-analysis.

Term	Heuristic	Necessity	Credit	Blame	Responsibility
Heuristic		1.00	1.00	−.29	.53
Necessity	1.00		.99	−.30	.54
Credit	1.00	.99		−.28	.53
Blame	−.29	−.30	−.28		.45
Responsibility	.53	.54	.53	.45	

Data availability

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

Acknowledgments

TG and RZ were supported by a grant from the Binational Science Foundation (BSF), Israel #2020212. DL and RZ were supported by the Economic and Social Research Council (ESRC) #RES-062-33-0004.

Appendix A. Correlations between model predictions

See Tables A.1–A.4.

Appendix B. Experiment 2 model predictions

See Table B.1.

Appendix C. Experiment 2 win probabilities

See Fig. C.1.

Appendix D. Experiment 2 criticality predictions of all models

See Fig. D.1.

Appendix E. Experiment 2 effort judgments

See Fig. E.1.

Table A.2
Experiment 1.

Term	Heuristic	Necessity	Credit	Blame	Responsibility
Heuristic		.94	.96	−.64	.21
Necessity	.94		.99	−.48	.43
Credit	.96	.99		−.48	.38
Blame	−.64	−.48	−.48		.38
Responsibility	.21	.43	.38	.38	

Table A.3
Experiment 2.

Term	Heuristic	Necessity	Credit	Blame	Responsibility
Heuristic					
Necessity			.97	−.64	.41
Credit		.97		−.63	.36
Blame		−.64	−.63		.21
Responsibility		.41	.36	.21	

Table A.4
All experiments combined.

Term	Heuristic	Necessity	Credit	Blame	Responsibility
Heuristic		.97	.98	−.46	.39
Necessity	.97		.98	−.60	.43
Credit	.98	.98		−.59	.37
Blame	−.46	−.60	−.59		.25
Responsibility	.39	.43	.37	.25	

Table B.1
Experiment 2: Raw model predictions for player A, player B, and player C with differing structures and skill levels. The skill level captures how many out of 10 times a player succeeded in the practice.

Index	Trial	Structure	Player	Skill	Necessity	Blame	Credit	Responsibility
1	1	conjunctive	A	3	1.00	0.06	1.00	0.09
2	1	conjunctive	B	3	1.00	0.06	1.00	0.09
3	1	conjunctive	C	3	1.00	0.06	1.00	0.09
4	2	conjunctive	A	7	1.00	0.22	1.00	0.49
5	2	conjunctive	B	7	1.00	0.22	1.00	0.49
6	2	conjunctive	C	7	1.00	0.22	1.00	0.49
7	3	conjunctive	A	4	1.00	0.08	1.00	0.12
8	3	conjunctive	B	6	1.00	0.03	1.00	0.08
9	3	conjunctive	C	2	1.00	0.20	1.00	0.24
10	4	conjunctive	A	4	1.00	0.20	1.00	0.30
11	4	conjunctive	B	6	1.00	0.09	1.00	0.20
12	4	conjunctive	C	5	1.00	0.14	1.00	0.24
13	5	conjunctive	A	4	1.00	0.36	1.00	0.48
14	5	conjunctive	B	6	1.00	0.16	1.00	0.32
15	5	conjunctive	C	8	1.00	0.06	1.00	0.24
16	6	conjunctive	A	7	1.00	0.01	1.00	0.03
17	6	conjunctive	B	3	1.00	0.05	1.00	0.07
18	6	conjunctive	C	1	1.00	0.19	1.00	0.21
19	7	conjunctive	A	7	1.00	0.05	1.00	0.15
20	7	conjunctive	B	3	1.00	0.27	1.00	0.35
21	7	conjunctive	C	5	1.00	0.12	1.00	0.21
22	8	conjunctive	A	7	1.00	0.10	1.00	0.27
23	8	conjunctive	B	3	1.00	0.54	1.00	0.63
24	8	conjunctive	C	9	1.00	0.03	1.00	0.21
25	9	disjunctive	A	3	0.49	1.00	0.22	0.49
26	9	disjunctive	B	3	0.49	1.00	0.22	0.49
27	9	disjunctive	C	3	0.49	1.00	0.22	0.49
28	10	disjunctive	A	7	0.09	1.00	0.06	0.09
29	10	disjunctive	B	7	0.09	1.00	0.06	0.09
30	10	disjunctive	C	7	0.09	1.00	0.06	0.09
31	11	disjunctive	A	4	0.32	1.00	0.16	0.32
32	11	disjunctive	B	6	0.48	1.00	0.36	0.48
33	11	disjunctive	C	2	0.24	1.00	0.06	0.24
34	12	disjunctive	A	4	0.20	1.00	0.09	0.20
35	12	disjunctive	B	6	0.30	1.00	0.20	0.30
36	12	disjunctive	C	5	0.24	1.00	0.14	0.24
37	13	disjunctive	A	4	0.08	1.00	0.03	0.08
38	13	disjunctive	B	6	0.12	1.00	0.08	0.12
39	13	disjunctive	C	8	0.24	1.00	0.20	0.24
40	14	disjunctive	A	7	0.63	1.00	0.54	0.63
41	14	disjunctive	B	3	0.27	1.00	0.10	0.27
42	14	disjunctive	C	1	0.21	1.00	0.03	0.21
43	15	disjunctive	A	7	0.35	1.00	0.27	0.35
44	15	disjunctive	B	3	0.15	1.00	0.05	0.15
45	15	disjunctive	C	5	0.21	1.00	0.12	0.21
46	16	disjunctive	A	7	0.07	1.00	0.05	0.07
47	16	disjunctive	B	3	0.03	1.00	0.01	0.03
48	16	disjunctive	C	9	0.21	1.00	0.19	0.21
49	17	mixed	A	3	1.00	0.42	1.00	0.51
50	17	mixed	B	3	0.70	0.17	0.41	0.21
51	17	mixed	C	3	0.70	0.17	0.41	0.21
52	18	mixed	A	7	1.00	0.75	1.00	0.91
53	18	mixed	B	7	0.30	0.17	0.23	0.21
54	18	mixed	C	7	0.30	0.17	0.23	0.21
55	19	mixed	A	4	1.00	0.56	1.00	0.68
56	19	mixed	B	6	0.80	0.18	0.71	0.32
57	19	mixed	C	2	0.40	0.18	0.12	0.16
58	20	mixed	A	4	1.00	0.71	1.00	0.80
59	20	mixed	B	6	0.50	0.12	0.38	0.20
60	20	mixed	C	5	0.40	0.12	0.25	0.16
61	21	mixed	A	4	1.00	0.87	1.00	0.92
62	21	mixed	B	6	0.20	0.05	0.13	0.08
63	21	mixed	C	8	0.40	0.05	0.35	0.16
64	22	mixed	A	7	1.00	0.15	1.00	0.37
65	22	mixed	B	3	0.90	0.60	0.73	0.63
66	22	mixed	C	1	0.70	0.60	0.19	0.49
67	23	mixed	A	7	1.00	0.36	1.00	0.65
68	23	mixed	B	3	0.50	0.45	0.23	0.35
69	23	mixed	C	5	0.70	0.45	0.54	0.49
70	24	mixed	A	7	1.00	0.80	1.00	0.93
71	24	mixed	B	3	0.10	0.14	0.03	0.07
72	24	mixed	C	9	0.70	0.14	0.68	0.49

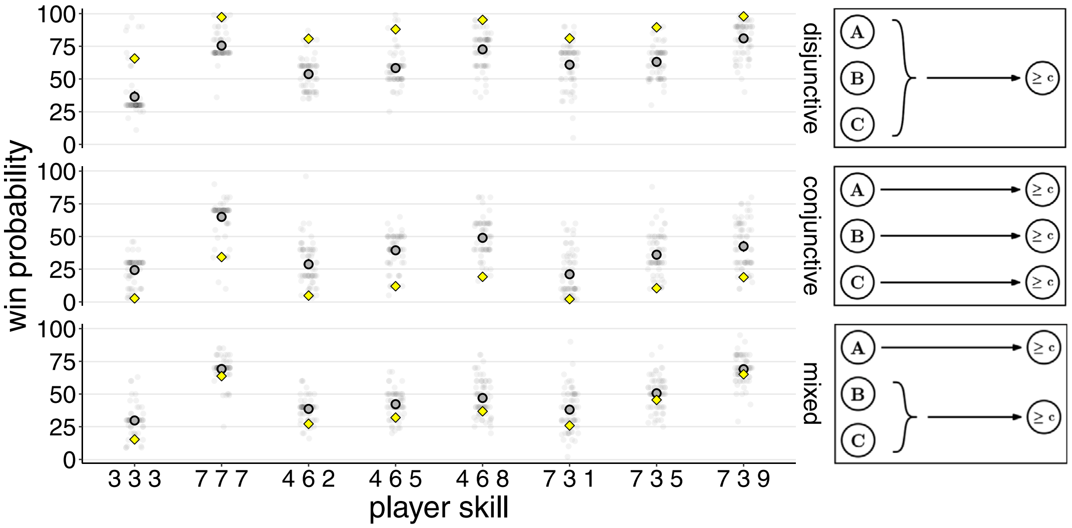


Fig. C.1. **Experiment 2**: Judged win probability (gray points) and predicted ground truth win probability (yellow diamonds). Large points show means with 95% confidence intervals. Small points show individual judgments. Participants were asked to predict how high the chances are that the team is going to win the challenge. Participants underestimated the probability of the team winning in the disjunctive challenge, and overestimated it in the conjunctive challenge. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

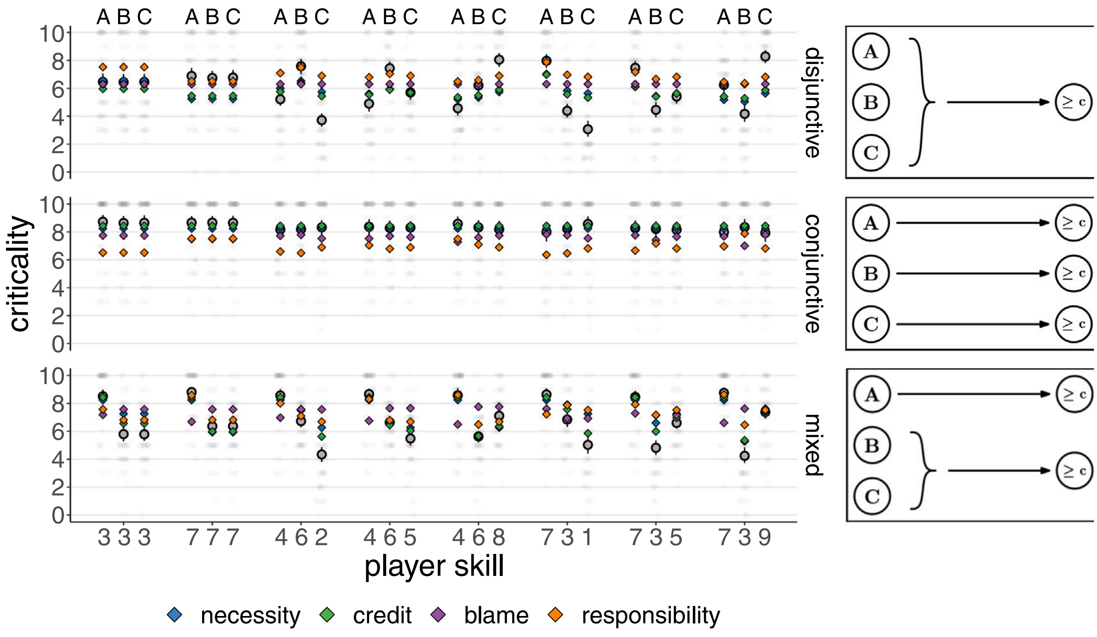


Fig. D.1. **Experiment 2**: Criticality judgments (gray points) and model predictions (colored diamonds). Large points show means with 95% confidence intervals. Small points show individual judgments. *Note*: The model predictions were scaled via a linear regression from the values shown in Table B.1 to the mean criticality judgments. We do not show the predictions of the *heuristic model* here because it cannot incorporate probabilistic information about the player's likelihood of success (as manipulated via player skill). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

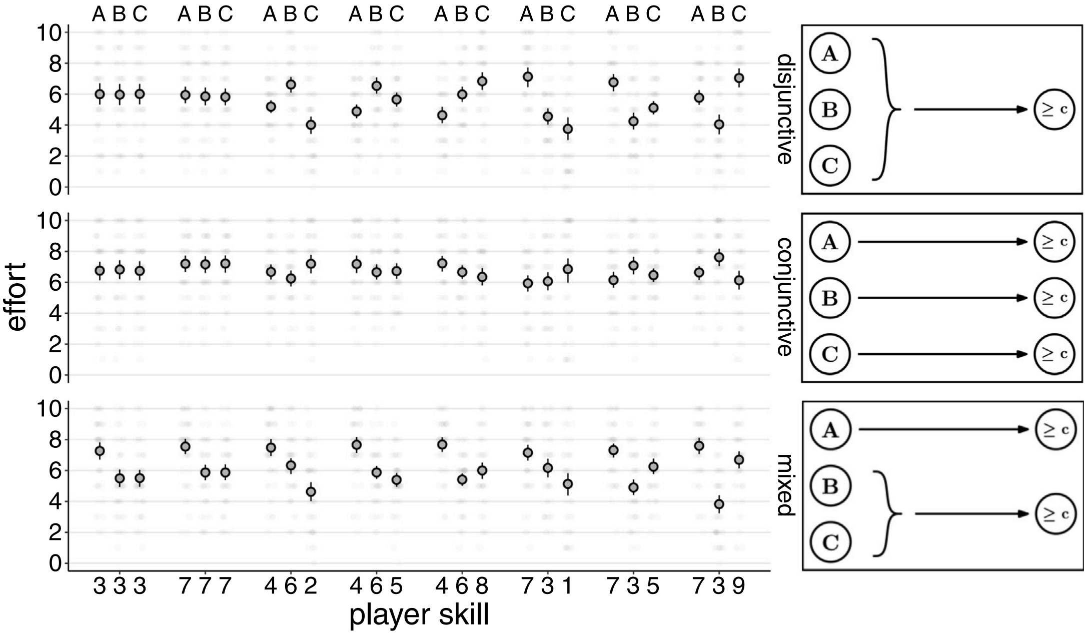


Fig. E.1. Experiment 2: Effort judgments (gray points). Large points show means with 95% confidence intervals. Small points show individual judgments. In the experiment, participants were asked to answer the following question: "How much effort do you expect each player to exert in this challenge?" Participants' provided their answers on sliding scales with the endpoints labeled "none" and "very much".

References

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

Anselm, R., Bhatia, D., Fischbacher, U., Hausfeld, J., et al. (2022). *Blame and Praise: Responsibility Attribution Patterns in Decision Chains*: *Technical report*, Thurgauer Wirtschaftsinstitut, Universität Konstanz.

Au, W. T. (2004). Criticality and environmental uncertainty in step-level public goods dilemmas. *Group Dynamics: Theory, Research, and Practice*, *8*(1), 40–61.

Au, W. T., Chen, X. P., & Komorita, S. S. (1998). A probabilistic model of criticality in a sequential public good dilemma. *Organizational Behavior and Human Decision Processes*, *75*(3), 274–293.

Au, W. T., & Chung, M. W. (2007). Effect of order of contribution in a sequential public goods dilemma. *Group Decision and Negotiation*, *16*(5), 437–449.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

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

Bartling, B., Fischbacher, U., & Schudy, S. (2015). Pivotality and responsibility attribution in sequential voting. *Journal of Public Economics*, *128*, 133–139.

Berg, N., & Gigerenzer, G. (2010). As-if behavioral economics: Neoclassical economics in disguise? *As-if Behavioral Economics*, 1000–1033.

Braham, M., & Hees, M. (2009). Degrees of causation. *Erkenntnis*, *71*(3), 323–344. http://dx.doi.org/10.1007/s10670-009-9184-8.

Braham, M., & van Hees, M. (2013). An anatomy of moral responsibility. *Mind*, *121*(483), 601–634. http://dx.doi.org/10.1093/mind/fzs081.

Brickman, P., Ryan, K., & Wortman, C. B. (1975). Causal chains: Attribution of responsibility as a function of immediate and prior causes.. *Journal of Personality and Social Psychology*, *32*(6), 1060–1067.

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

Dannenberg, A., Löschel, A., Paolacci, G., Reif, C., & Tavoni, A. (2015). On the provision of public goods with probabilistic and ambiguous thresholds. *Environmental and Resource Economics*, *61*(3), 365–383.

De Cremer, D. (2007). When the rich contribute more in public good dilemmas: The role of provision point level. *European Journal of Social Psychology*, *37*(3), 536–546.

De Cremer, D., & van Dijk, E. (2002). Perceived criticality and contributions in public good dilemmas: A matter of feeling responsible to all? *Group Processes & Intergroup Relations*, *5*(4), 319–332.

Douer, N., & Meyer, J. (2022). Judging one's own or another person's responsibility in interactions with automation. *Human Factors*, *64*(2), 359–371.

Duch, R., Przepiorka, W., & Stevenson, R. (2015). Responsibility attribution for collective decision makers. *American Journal of Political Science*, *59*(2), 372–389.

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour*, *3*(6), 554–559.

Engl, F. (2022). Causal responsibility attribution: Theory and experimental evidence.

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.

Fazelpour, S. (2021). Norms in counterfactual selection. *Philosophy and Phenomenological Research*, *103*(1), 114–139.

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

Felsenthal, D. S., & Machover, M. (2009). A note on measuring voters' responsibility. *Homo Oeconomicus*, *26*(2), 259–271.

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

Friedenberg, M., & Halpern, J. Y. (2019). Blameworthiness in multi-agent settings. In *Proceedings of the thirty-third AAAI conference on artificial intelligence (AAAI-19)*.

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.

Gelman, A., Katz, J., & Tuerlinckx, F. (2002). The mathematics and statistics of voting power. *Statistical Science*, *17*(4), 420–435.

Gerstenberg, T. (2022). What would have happened? Counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *377*(1866), 20210339.

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual conference of the cognitive science society* (pp. 720–725). Austin, TX: Cognitive Science Society.

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., & Kareev, Y. (2010). The dice are cast: The role of intended versus actual contributions in responsibility attribution. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual conference of the cognitive science society* (pp. 1697–1702). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, [ISSN: 00100277] *177*, 122–141. http://dx.doi.org/10.1016/j.cognition.2018.03.019.

Goldman, A. I. (1999). Why citizens should vote: A causal responsibility approach. *Social Philosophy and Policy*, *16*(2), 201–217.

Hardin, G. (1968). The tragedy of the commons. *Science*, *126*, 1243–1248.

Hart, H. L. A. (2008). *Punishment and responsibility*. Oxford: Oxford University Press.

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. New York: Oxford University Press.

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

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

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.

Isaac, R. M., Walker, J. M., & Williams, A. W. (1994). Group size and the voluntary provision of public goods:: Experimental evidence utilizing large groups. *Journal of Public Economics*, *54*(1), 1–36.

Kerr, N. L. (1989). Illusions of efficacy: The effects of group size on perceived efficacy in social dilemmas. *Journal of Experimental Social Psychology*, *25*(4), 287–313.

Kerr, N. L. (1992). Efficacy as a causal and moderating variable in social dilemmas. In *Social dilemmas: theoretical issues and research findings* (pp. 59–80).

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.

Kerr, N. L., & Kaufman-Gilliland, C. M. (1997). "... And besides, I probably couldn't have made a difference anyway": Justification of social dilemma defection via perceived self-inefficacy. *Journal of Experimental Social Psychology*, *33*(3), 211–230.

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Social Psychology*, *151*(7), 1481–1501.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 183–214.

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., & 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.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, *37*(5), 879–901.

McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of Personality and Social Psychology*, *79*(5), 677–689. http://dx.doi.org/10.1037/0022-3514.79.5.677.

Meehl, P. E. (1977). The selfish voter paradox and the thrown-away vote argument. *American Political Science Review*, *71*(1), 11–30.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. PsyArXiv URL https://psyarxiv.com/nq53z.

Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, *121*(1–2), 93–149.

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

Quillien, T., & Lucas, C. G. (2022). Counterfactuals and the logic of causal selection.

Rapoport, A. (1987). Research paradigms and expected utility models for the provision of step-level public goods. *Psychological Review*, *94*(1), 74–83.

Rapoport, A. (1988). Provision of step-level public goods: Effects of inequality in resources. *Journal of Personality and Social Psychology*, *54*(3), 432–440.

Rapoport, A., Bornstein, G., & Erev, I. (1989). Intergroup competition for public goods: Effects of unequal resources and relative group size. *Journal of Personality and Social Psychology*, *56*(5), 748–756.

Riker, W. H., & Ordeshook, P. C. (1968). A theory of the calculus of voting. *American Political Science Review*, *62*(1), 25–42.

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

Sousa, P. (2009). A cognitive approach to moral responsibility: The case of a failed attempt to kill. *Journal of Cognition and Culture*, *9*(3), 171–194.

Spadaro, G., Graf, C., Jin, S., Arai, S., Inoue, Y., Lieberman, E., et al. (2022). Cross-cultural variation in cooperation: A meta-analysis. *Journal of Personality and Social Psychology*.

Steiner, I. D. (1972). *Group process and productivity*. Academic Press.

Teigen, K. H., & Brun, W. (2011). Responsibility is divisible by two, but not by three or four: Judgments of responsibility in dyads and groups. *Social Cognition*, *29*(1), 15–42.

Weiner, B. (1986). *An attribution theory of motivation and emotion*. New York: Springer.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: The Guilford Press.

Weiner, B. (2011). An attribution theory of motivation. In P. A. M. V. Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology: Volume 1* (pp. 135–155). London: Sage Publications Ltd.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Wu, S. A., & Gerstenberg, T. (2023). If not me, then who? Responsibility and replacement. PsyArXiv URL https://psyarxiv.com/m2rcj/.

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