Cognitive Psychology 129 (2021) 101412

Contents lists available at ScienceDirect

Cognitive Psychology

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

Predicting responsibility judgments from dispositional inferences and causal attributions

Antonia F. Langenhoff [a], Alex Wiegmann [b], Joseph Y. Halpern [c], Joshua B. Tenenbaum [d], Tobias Gerstenberg [e,*]

[a] University of California, Berkeley, United States
[b] Ruhr University Bochum, Germany
[c] Cornell University, United States
[d] Massachusetts Institute of Technology, United States
[e] Stanford University, United States

ARTICLE INFO

Keywords:
Responsibility
Causality
Counterfactuals
Pivotality
Normality
Voting
Expectations

ABSTRACT

The question of how people hold others responsible has motivated decades of theorizing and empirical work. In this paper, we develop and test a computational model that bridges the gap between broad but qualitative framework theories, and quantitative but narrow models. In our model, responsibility judgments are the result of two cognitive processes: a dispositional inference about a person's character from their action, and a causal attribution about the person's role in bringing about the outcome. We test the model in a group setting in which political committee members vote on whether or not a policy should be passed. We assessed participants' dispositional inferences and causal attributions by asking how surprising and important a committee member's vote was. Participants' answers to these questions in Experiment 1 accurately predicted responsibility judgments in Experiment 2. In Experiments 3 and 4, we show that the model also predicts moral responsibility judgments, and that importance matters more for responsibility, while surprise matters more for judgments of wrongfulness.

1. Introduction

Shortly before the 2016 presidential election, Christopher Suprun, a Texas state elector for the Republican party, signaled that he would refuse to vote for Donald Trump, even if Trump won the popular vote in his state. Trump did indeed win the popular vote in Texas and on election day, as announced, Suprun voted for a different candidate. His decision caused turmoil among Republicans. Both Suprun's party colleagues and the voters vociferously proclaimed their anger in newspapers, blogs, and social networks. Despite Suprun's attempt, Trump won the electoral vote – and thus, the presidential election. Imagine that Hilary Clinton had become the next president of the United States. Certainly, Suprun's party colleagues would have held him responsible for contributing to Clinton's victory and Trump's loss in that case. But to what extent? Intuitively, Suprun would have been blamed more than a Democratic state elector who also voted against Trump. And suppose that Clinton's victory margin was only a couple of votes, as some projections had suggested before the election. Presumably Republicans would have blamed Suprun even more.

Judgments of responsibility are ubiquitous in our everyday lives. When something goes wrong – for example, when our favored

* Corresponding author at: 450 Jane Stanford Way, Building 420, Stanford, CA 94305, United States.
E-mail address: gerstenberg@stanford.edu (T. Gerstenberg).

https://doi.org/10.1016/j.cogpsych.2021.101412
Received 22 September 2019; Received in revised form 28 March 2021; Accepted 28 June 2021
0010-0285/

candidate lost an election – we want to know who is to responsible. The concept of responsibility has intrigued researchers in psychology (Alicke, 2000; Hilton, McClure, & Slugoski, 2005; Lagnado & Harvey, 2008; Shaver, 1985; Cushman, 2008; Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000; Malle, 2021), philosophy (Sartorio, 2007; Feinberg, 1968; Strawson, 1962; Shoemaker, 2015), and the legal sciences (Hart & Honoré, 1959/1985; Moore, 2009; Tobia, 2018; Summers, 2018; Lagnado, Fenton, & Neil, 2013; Bayles, 1982) for decades. In this paper, we further develop and test a computational model for responsibility judgments that was originally introduced by Gerstenberg et al. (2018).

The model predicts that responsibility judgments are influenced by two processes. The first process is a *dispositional inference* that captures what can be learned about a person's character from observing their action (Heider, 1946; Ajzen, 1971; Weiner & Kukla, 1970; Fishbein & Ajzen, 1973). The idea is that, in a given situation, the observer forms an expectation about how another person will act, based on their knowledge about the person and the situation. The more the person's actual behavior diverts from what was expected, the more likely the observer infers that the person's action must have been determined by an unobserved aspect of their disposition. This dispositional inference, in turn, translates into a responsibility judgment: Another person is held responsible to the extent that their action was determined by their disposition, goal, or desire, rather than by determinants that were outside of their control (e.g. Alicke, 2000; Uttich & Lombrozo, 2010).

In Suprun's case, his party affiliation and the outcome of the popular vote all spoke in favor of him voting for Trump. Given the gap between their expectations about how he would vote and Suprun's actual vote, the model predicts that Republicans would assign a high level of blame to him for contributing to Trump's (hypothetical) loss. Critically, the model predicts that Republicans would blame Suprun more than, for example, a Democratic state elector who also voted against Trump, but for whom voting for a candidate other than Trump was to be expected.

The second process is a *causal attribution* that determines what role the person's action played in bringing about the outcome. The model predicts that a person is held more responsible for an outcome the closer their action was to having made a difference to the outcome (see Chockler & Halpern, 2004; Lagnado, Gerstenberg, & Zultan, 2013). In the version of our hypothetical scenario above, in which Clinton and Trump were almost on a par and Clinton won the election by a margin of only a couple of votes, Suprun's vote for a different candidate was closer to having made a difference to the outcome than in a scenario in which the vast majority of electoral college members voted for candidates other than Trump. In the first case, had Suprun voted for Trump, he might have just tipped the balance in Trump's favor, while in the latter case, Trump would have lost the election even if Suprun had decided to vote for him. The model predicts that Republicans would blame Suprun more in the close call compared to the clear loss.

In what follows, we first review prior work that has looked at how people draw dispositional inferences and make causal attributions. Next, we discuss existing frameworks and models of responsibility attribution, and then explain how we implemented dispositional inferences and causal attributions in our computational model. We test the predictions of our model in four experiments. We conclude by highlighting what we see as the key contributions of our work, and by discussing some remaining challenges.

1.1. Prior work on responsibility judgments

Both dispositional inferences and causal attributions are important for assigning responsibility (e.g. Shaver, 1985; Weiner, 1995; Malle, Guglielmo, & Monroe, 2014; Alicke, 2000; Schlenker, Britt, Pennington, Murphy, & Doherty, 1994; Mao & Gratch, 2012; Chockler & Halpern, 2004). The two components are also reflected in the law as central elements for determining criminal liability (Duff, 1993; Lagnado & Gerstenberg, 2017; Summers, 2018): *mens rea* (a guilty mind) and *actus reus* (a guilty act). In this section, we discuss prior research that has influenced how we formalize dispositional inferences and causal attributions in our model.

1.1.1. Dispositional inferences

What does a person's action reveal about the kind of person they are? In most situations, there are several possible explanations for any observed behavior. For example, an action might be primarily influenced by external factors, such as the situation the agent was in, or by internal factors, such as the agent's abilities, dispositions, beliefs, and desires. To what extent external and internal factors influenced a person's action is critical for attributions of responsibility (Alicke, 2000; Malle et al., 2014; Shaver, 1985; Weiner & Kukla, 1970; Woolfolk, Doris, & Darley, 2006).

Generally, people consider how both internal and external factors shape behavior. However, when explaining others' behavior (as compared to our own behavior), we tend to emphasize dispositional or character-based explanations, and neglect the influence of situational factors (Ross, Amabile, & Steinmetz, 1977; Jones & Harris, 1967). Moreover, when considering the morality of other people's actions (Waldmann, Nagel, & Wiegmann, 2012), we often focus on those features of an action that are diagnostic about a person's character, rather than on its consequences, or on whether a moral rule was broken (Hursthouse, 1999; McIntyre, 2019; Bartels & Pizarro, 2011; Bayles, 1982; Pizarro, Uhlmann, & Salovey, 2003; Uhlmann, Pizarro, & Diermeier, 2015; Reeder, 2009). Based on these findings, researchers have argued that people are inherently motivated to determine the moral character of others (Uhlmann et al., 2015). We evaluate whether others are caring, fair, and trustworthy, and based on these evaluations, decide whether to cooperate, compete, or avoid them in the future. For judgments of *moral* responsibility, what the action reveals about the person might matter more than what difference it made to the outcome (Cushman, 2008).

Early attribution theorists suggested Bayesian inference as a normative framework for studying how people understand others' behavior (Ajzen, 1971; Ajzen & Fishbein, 1975; Fischhoff & Beyth-Marom, 1983; Morris & Larrick, 1995; Fishbein & Ajzen, 1973; Trope & Burnstein, 1975; Trope, 1974). From this perspective, people make sense of others' behaviors by considering different hypotheses for a given action, favoring those with higher prior probability that explain the behavior well. Recent work has modeled behavioral attribution processes computationally within this framework (Baker, Saxe, & Tenenbaum, 2009; Baker, Jara-Ettinger, Saxe,

& Tenenbaum, 2017; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Assuming that agents act approximately rationally (Dennett, 1987), one can infer the underlying beliefs and desires from their actions. For example, Suprun's fellow Republicans had clear expectations about how he *should* vote: in line with his party affiliation. When Suprun defied this expectation, his fellow Republicans' could infer that Suprun's desire to support their party wasn't as strong as they would have liked it to be.

1.1.2. Causal attribution

What causal role did the person's action play in bringing about the outcome? One way of capturing whether a person's action was causally connected to an outcome is to consider whether the outcome would have been different in a counterfactual situation in which the person hadn't acted (Lewis, 1973; Woodward, 2003; Halpern & Pearl, 2005; Pearl, 2000; Yablo, 2002). While this counterfactual test works well in simple situations, it fails in more complex situations that involve multiple causes. Imagine the following scenario: A political committee, consisting of Ms. A, Ms. B, and Ms. C, votes on a policy. The policy is passed if at least two of the committee members vote for it. When all three members end up voting in favor of the policy, none of the three committee members qualify as a cause of the outcome according to the simple counterfactual test. Even if any one of the members had voted against the policy it would still have passed.

To solve this problem Halpern and Pearl (2005) introduced a test of counterfactual dependence that considers not only whether an action would have made a difference in the actual situation, but also whether it would have made a difference in another possible situation that could have arisen. According to this criterion, each committee member qualifies as a cause in our scenario. For example, even though Ms. A's vote didn't make a difference in the actual situation, it would have made a difference if Ms. B (or Ms. C) had voted differently.

Inspired by Halpern and Pearl's (2005) model, Chockler and Halpern (2004) defined a graded notion of causal responsibility based on how close a person's action was to making a difference to the outcome. The fewer changes would have been required to make the person's action pivotal, the more causal responsibility the person bears for the outcome. Gerstenberg and colleagues tested this model in a number of experiments, finding that indeed, individuals are held more responsible the closer their action was to having been pivotal (Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Zultan, Gerstenberg, & Lagnado, 2012; Lagnado et al., 2013).

How close a person's action was to making a difference to an outcome is not the only thing that matters. Responsibility judgments are also sensitive to how critical a person's contribution was for a positive outcome (Lagnado et al., 2013), whether their action was sufficient for bringing about the outcome (Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015), whether there was a physical connection between the cause and the outcome (Lombrozo, 2010; Dowe, 2000; Walsh & Sloman, 2011; Wolff, 2007), whether the action was optimal (Johnson & Rips, 2015), whether the causal relationship between action and outcome was robust to possible changes in the background conditions (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Vasilyeva, Blanchard, & Lombrozo, 2018), and whether the action was normal or abnormal (Gerstenberg & Icard, 2019; Hitchcock & Knobe, 2009; Halpern & Hitchcock, 2015; Samland & Waldmann, 2016; Morris, Phillips, Gerstenberg, & Cushman, 2019; Quillien, 2020; Kirfel, Icard, & Gerstenberg, 2020; Sytsma, Livengood, & Rose, 2012; Hilton & Slugoski, 1986; Knobe & Fraser, 2008).

1.2. Theoretical frameworks and computational models of responsibility judgments

In this section, we discuss the theoretical frameworks and computational models that have motivated our account.

1.2.1. Theoretical frameworks

The first comprehensive frameworks of responsibility judgments were decision-stage models (Heider, 1958; Weiner, 1995; Shaver, 1985). Shaver (1985), for example, differentiated between causal attribution, responsibility judgments, and blame, whereby each subsequent stage requires that the conditions for the previous stage are met. For example, an agent can only be held responsible if they were a cause of the outcome. To be held responsible, the agent must have acted knowingly and voluntarily. Whether blame is warranted depends on whether the agent had acceptable excuses or justifications for their actions. Weiner (1995) proposed a similar stage-like framework in which an agent is to be held responsible when they caused the outcome and were in control. More recently, Malle et al. (2014) proposed the *Path Model of Blame* in which the agent's intention plays a central role. If an agent intentionally caused an outcome, the observer considers the agent's reasons for acting. Whereas if the outcome was brought about unintentionally, the observer considers the agent's obligation and capacity for preventing the outcome from happening (see also, Monroe and Malle, 2017).

The role that person inferences play in judgments of responsibility is highlighted in Alicke's (2000) *Culpable Control Model*. Alicke (2000) posits that when strong negative emotions are evoked, people experience an immediate "desire to blame" which may bias their construal of what happened. So when two people acted identically and caused a negative outcome, observers might feel a stronger desire to blame the person who has a dubious moral character than the morally virtuous person. Therefore, the observer might view the person with the dubious moral character as having played a more important causal role in bringing about the outcome than the person with the good character. Similarly, Schlenker et al.'s (1994) *Triangle Model* emphasizes the role that person information plays for judgments of responsibility. People are responsible by virtue of the norms that apply to them in a given situation. When a person drowns, for example, the lifeguard is to blame because based on their role, it was their responsibility to prevent this outcome from happening (see also Hamilton, 1978).

What all of these theoretical frameworks have in common is that they start with a causal analysis of what happened, and then postulate additional factors that are relevant to assigning responsibility or blame, such as the agent's character. However, as Malle et al. (2014, p. 177) noted, "a major limitation of […] all extant models […] is that they do not generate any quantitative predictions." They cannot tell us how *much* a person was responsible for a particular outcome. Generating quantitative predictions, however, is

critical for rigorous empirical tests.

1.2.2. Computational models

Computational models of responsibility judgments have relied on a variety of formal tools such as logic, probability, and counterfactuals. For example, Shaver's (1985) framework (discussed above) has a logical structure in that it proposes an entailment relationship between the concepts of causation, responsibility, and blame, where each subsequent concept entails the former but requires additional conditions to be met in order to apply (see also Hewstone & Jaspars, 1987).

Some computational models explain responsibility judgments in terms of the difference that an action made to the observer's degree of belief that the outcome would happen (Spellman, 1997; Brewer, 1977; Fincham & Jaspars, 1983). The more an action changed the subjective probability in the outcome, the more responsibility is attributed to that action (but see McClure, Hilton, & Sutton, 2007; Gerstenberg & Lagnado, 2012). Relatedly, Johnson and Rips (2015) have shown that decision-makers are held more responsible for (positive) outcomes that resulted from optimal rather than suboptimal choices.

Counterfactuals have also played an important role in computational models of responsibility (Engl, 2018; Naumov & Tao, 2018; Chockler & Halpern, 2004; Felsenthal & Machover, 2009; Lagnado et al., 2013). For example, the structural model of responsibility by Chockler and Halpern (2004) mentioned above defines responsibility as closeness to pivotality, where actions that were close to being pivotal are held more responsible for the outcome.

What these computational models have in common is that they focus on the causal role that the action played in bringing about the outcome. More recently, Mao and Gratch (2012) have developed a computational model of responsibility that builds on Shaver's (1985) logical entailment model and Chockler and Halpern's (2004) extended counterfactual model. Their model considers both the agent's mental states and their action when assigning responsibility (see also Halpern & Kleiman-Weiner, 2018; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015, which include formalizations of intentions). However, since the model uses logical inference rules, it doesn't yield graded predictions.

Our work presented here follows most closely from a computational model developed in Gerstenberg et al. (2018). Their model considers both the causal role that an action played for an outcome, as well as what the action reveals about the person. Gerstenberg et al. (2018) used Chockler and Halpern's (2004) model to capture what causal role the action played. Further, they developed a Bayesian model that infers what kind of person a person is from having observed their action.

1.3. Towards a comprehensive computational model of responsibility attributions

In this paper, we extend Gerstenberg et al.'s (2018) model and test its predictions in a voting paradigm that allows us to quantitatively manipulate information relevant to the dispositional inferences and causal attributions. In our experiments, we presented participants with scenarios in which different political committees voted on whether or not a policy should be passed. This new paradigm allows us to ask a number of new questions.

First, does the model pass a more direct test of its two key components: dispositional inferences and causal attributions? In previous tests, Gerstenberg et al. (2018) manipulated how expected an agent's action was, and whether the action made a difference to the outcome. While participants' responsibility judgments were consistent with the model's predictions, Gerstenberg et al. (2018) didn't test the components of their model directly. Here, we assess participants' dispositional inferences and causal attributions by asking them to evaluate (a) how surprising an agent's action was and (b) how important the action was for bringing about the outcome. We then investigate whether these judgments predict responsibility judgments as predicted by the computational model.

Second, do the model's predictions hold in more complex causal settings? Gerstenberg et al.'s (2018) previous tests of the model focused on situations in which a single agent brought about an outcome. However, several people often jointly contribute to an outcome, as in our hypothetical example where Clinton won the presidential election. In this scenario, Suprun was one among many electoral college members who voted for a candidate other than Trump, and thereby contributed to Clinton's victory. Gerstenberg and colleagues have investigated how people distribute responsibility in situations in which the contributions of several individuals combine to yield a group outcome (Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Lagnado et al., 2013; Zultan et al., 2012; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018; Lagnado & Gerstenberg, 2015; Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015). Here, we connect this research on responsibility judgments in group settings with Gerstenberg et al.'s (2018) work by manipulating both the causal structure of the situation as well as action expectations in graded ways.

Third, does the model also capture people's judgments of *moral* responsibility? We predict that in the moral domain, inferences about a person's character become more important than causal attributions, reflecting the fundamental human motivation to determine the moral character of others (Uhlmann et al., 2015).

We report the results of four experiments designed to tackle these questions. Experiment 1 tests how participants make dispositional inferences (asking how surprising a particular action was in a given situation) and causal attributions (asking how important an action was for the outcome). Experiment 2 asks participants to make responsibility judgments in a large variety of situations that manipulated action expectations and the causal structure in graded ways. Experiment 3 applies the model to the moral domain by manipulating the moral valence of the outcome. Finally, Experiment 4 brings together key elements of the three previous studies by assessing judgments of surprise, importance, and responsibility for a variety of voting scenarios in moral contexts. Experiments 3 and 4 also explore differences between evaluations of responsibility and moral wrongfulness (see Cushman, 2008).

Policy information
Number: # 109383
Supported by: The Democratic party
Votes in favor of policy required: 5

Votes

	Party affiliation	Voted "yes"
Allie	Democrat	
Bridget	Democrat	✓
Christie	Democrat	✓
Dalia	Republican	
Emma	Republican	✓

Outcome: The policy was **not passed**. 3 out of 5 committee members voted in favor of the policy and 5 votes were required for the policy to pass.

Fig. 1. Exemplary voting scenario. Here, the policy was generally supported by the Democratic party and required five votes in favor to pass. Two of the three Democrats in the committee voted for the policy, and one of the two Republicans. The policy didn't was because only three voted in favor and five votes were required. Between scenarios, we manipulated which party generally supported the policy, how many votes were required for a policy to pass, the party affiliations of the committee members, and their votes.

a)

$$\mathbf{Party}_{\mathbf{same}} \sim \mathbf{beta}(a, b)$$
$$\mathbf{Party}_{\mathbf{other}} \sim \mathbf{beta}(b, a)$$
$$\mathbf{Policy} \sim \mathbf{beta}\left(\frac{a+b}{2}, \frac{a+b}{2}\right)$$
$$\mathbf{Vote}_{\mathbf{same}} = \frac{\mathbf{Party}_{\mathbf{same}} + \mathbf{Policy}}{2}$$
$$\mathbf{Vote}_{\mathbf{other}} = \frac{\mathbf{Party}_{\mathbf{other}} + \mathbf{Policy}}{2}$$

b)

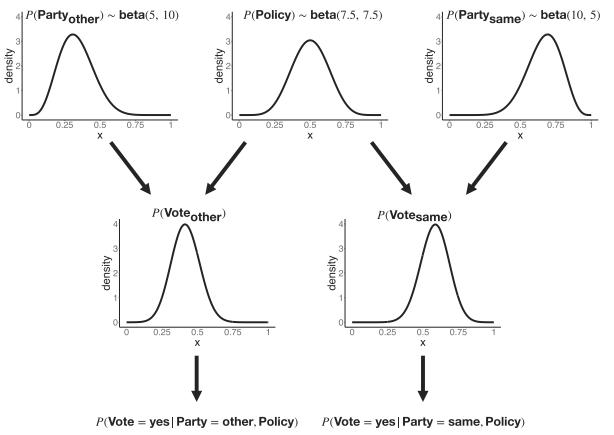

Fig. 2. Generative voting model: (a) Mathematical form. (b) Graphical illustration. When deciding how to vote, a voter takes into account his party affiliation (same or other), and the quality of the policy, weighing each factor equally. We fit a and b to the data with the constraint that $a > b$, reflecting the assumption that committee members affiliated with the party that supports the policy are*a priorimore likely to vote in favor of the policy than committee members from the other party. The diagram shows the shape of the prior distributions for* $a = 10$*, and* $b = 5$*. See Fig. A1 for a sensitivity analysis of the parameter space, and Table C1 for detailed model predictions. Note:* $\sim$ *indicates "distributed as".*

2. Overview of the experimental paradigm

In our experiments, we presented participants with scenarios in which different political committees voted on whether or not a policy should be passed. For each scenario, participants saw how many votes in favor were required for the policy to pass, how each of the committee members voted, and what the outcome of the vote was. In Experiments 1 and 2, participants also saw each committee member's party affiliation and which party supported the policy: the Republican or the Democratic party.

Fig. 1 shows a voting scenario similar to the ones used in Experiments 1 and 2. Policy #109383 was up for vote. There were five people on the committee: Allie, Bridget, Christie, Dalia and Emma. The policy was supported by the Democratic party. Five votes in favor of the policy were required in order for the policy to be passed. As it turned out, the Democrats Bridget and Christie, as well as the Republican Emma voted in favor of the policy, whereas Allie and Dalia voted against it. The policy was not passed since only three committee members voted in favor but all five votes were required for the policy to pass.

Experiments 3 and 4 didn't include information about party affiliation. Instead, we told participants about the content of the policy that was up for vote. In Experiment 3, one group of participants made their judgments in a context where the content and the consequences of the policy were "morally neutral" (changing documents into a certain font) while another group made their judgments in a context where the content and the consequences of the vote were "morally negative"(introducing corporal punishment in schools). In Experiment 4, participants saw a range of policies with different contents, and they rated how "morally bad" each policy was.

3. Model

We now discuss how we implemented the computational model's dispositional inference component and the causal attribution component for the experiments reported here.

3.1. Dispositional inferences

The model predicts that in a given situation, people draw inferences about a person's character from their action, and that these inferences, in turn, affect the extent to which a person is held responsible for an outcome. In our paradigm, the question is how much an observer learns about a committee member from how they voted. We assume that there are three driving forces that affect a committee member's vote: (1) their belief about how strongly their party supports the policy, (2) their belief about the quality of the policy, and (3) their individual preference. An observer infers (1) and (2) based on how the committee members voted. A committee member's individual preference remains unobserved. We assume that a vote is attributed to a committee member's individual preference to the extent that it is surprising, given how the other committee members voted. We expect that a committee member will be held more responsible for the outcome if their vote was surprising and thus indicative of a strong individual preference.

Specifically, we assume the generative voting model illustrated in Fig. 2. Committee members who are affiliated with the party that is stated to support the policy start with a prior belief that their party supports the policy ($\text{Party}_{\text{same}}$), whereas committee members from the other party believe that their party doesn't support the policy ($\text{Party}_{\text{other}}$). A committee member votes by equally taking into account their belief about their party's support as well as their belief about the quality of the policy (Policy). We assume that beliefs about the quality of a policy are initially unbiased – that is, policies are just as likely to be good or bad. We model these prior beliefs using beta distributions which have support between 0 and 1. We then assume that a committee member makes their choice about how to vote ($\text{Vote}_{\text{same}}$ or $\text{Vote}_{\text{other}}$) by equally weighting their belief about the party's support as well as the quality of the policy.

Our model performs Bayesian inference by conditioning on the observed evidence (the votes) to go from prior distributions over the party and policy factors to posterior distributions over these factors (see Eq. 1).

$$\begin{aligned} &p\left(\text{Party}_{\text{same}}, \text{Party}_{\text{other}}, \text{Policy} | \overrightarrow{\text{Votes}}\right) \propto \\ &p\left(\overrightarrow{\text{Votes}} | \text{Party}_{\text{same}}, \text{Party}_{\text{other}}, \text{Policy}\right) \cdot p(\text{Party}_{\text{same}}) \cdot p(\text{Party}_{\text{other}}) \cdot p\left(\text{Policy}\right) \end{aligned} \tag{1}$$

We assume that the vector of votes $\overrightarrow{\text{Votes}}$ is generated from a binomial distribution with the probability of each vote determined by party membership and policy as shown in Fig. 2. Based on the posteriors over $\text{Party}_{\text{same}}$, $\text{Party}_{\text{other}}$ and Policy, the model then forms an expectation about how the committee member of interest will vote. We implemented the dispositional inference model in `R` (R Core Team, 2019) using the `greta` package (Golding, 2018). We modelled the prior distributions over Party and Policy as beta distributions, and the likelihood function for the pattern of votes as a binomial distribution.

For an example, consider committee member Allie in the voting scenario shown in Fig. 1. Allie is a Democrat, and thus affiliated with the party that supports the policy. Accordingly, she is a priori more likely to vote for rather than against the policy. The model then updates this prior distribution based on how the other committee members voted. The two other Democrats, Bridget and Christie, voted for the policy, and one out of the two Republicans voted for the policy. Based on this evidence, the model now believes Allie is even more likely than before to vote in favor of the policy.

We define the extent to which a committee member's vote is surprising as the difference between the actual vote (coding a vote against the policy as 0 and a vote for the policy as 1) and the expected vote (where we use the mean of the posterior over the committee member's vote ($\text{Vote}_{\text{same}}$ or $\text{Vote}_{\text{other}}$ depending on the committee member's party) as our measure of expectation; see Fig. 2). Given that an observer would have expected Allie to vote for the policy, her actual vote against the policy is surprising. The inference that Allie's vote must have been affected by her individual preference (since it's not well-explained by how the others voted) is then predicted to lead to an increased judgment of responsibility.

3.2. Causal attributions

The model predicts that a person is held more responsible to the extent that their action was important for bringing about the outcome. The model construes importance by taking into account how close the person's action was to having been pivotal for the outcome, and by considering the number of causes that contributed to the outcome.

We define the pivotality of a person's action A for an outcome E in a particular scenario S as

$$\text{Pivotality}\left(A, S, E\right) = \frac{1}{C+1}, \tag{2}$$

where C is the minimal number of changes that are required to make A pivotal for E. S describes the causal structure of the situation and what actually happened. In our voting scenarios, S includes the number of votes needed for a policy to be passed (the threshold) and how each committee member voted. In the voting scenarios that we consider, C simply represents the number of other voters who would have needed to vote differently in order for the person under consideration to become pivotal. For example, Allie's pivotality in our example above is $\frac{1}{2}\left(\frac{1}{1+1}\right)$, since one vote needs to be changed to make Allie's vote pivotal (Dalia would have needed to vote in favor

of the policy, rather than against it).

In addition to pivotality, the model also considers the number of causes that contributed to the outcome. Different lines of research suggest that people assign more responsibility to an action for an outcome if fewer causes contributed to the outcome (White, 2014; Darley & Latané, 1968; Latané, 1981). To see how this notion differs from that of pivotality, consider the well-known "diffusion of responsibility" phenomenon: In situations where multiple people would be capable of helping another person in an emergency, people often have a reduced sense of responsibility (Darley & Latané, 1968).[1] In such a situation, each "bystander" is pivotal – if anyone intervened, the victim would be helped, but nevertheless individuals have a reduced sense of responsibility as the number of people who could help increases.

Thus, in addition to how close a person's action was to making a difference to the outcome (as measured by pivotality), we also predict that the number of causes that contributed to the outcome affects how important an individual contribution is perceived. The more causes contributed to an outcome, the less important each individual cause is perceived to be. In our voting setting, this means that the more committee members voted in line with the outcome of the vote, the less important each vote is perceived to be.

Overall, we predict that both *pivotality* and *the number of causes* affect participants' causal attributions. We assume that both factors affect causal attributions in an additive way (with *number of causes* being a negative predictor).

$$\text{Causal attribution} = \alpha + \beta_1 \cdot \text{Pivotality} + \beta_2 \cdot \text{Number of causes}, \tag{3}$$

whereby β_1 and β_2 determine how much emphasis is put on pivotality and the number of causes when making causal attributions, and α is a constant used for mapping from the predictors to participants' response scale in a regression.

3.3. Bringing it together: The computational model

We predict that judgments of responsibility are sensitive to what the observer learned about the person from their action ('dispositional inference'), and how important the person's action was perceived for the outcome ('causal attribution'). We assume that both factors of the model combine additively to affect judgments of responsibility,

$$\text{Responsibility} = \alpha + \beta_1 \cdot \text{Dispositional inference} + \beta_2 \cdot \text{Causal attribution}, \tag{4}$$

whereby β_1 and β_2 capture how much influence each factor has on the overall judgment, and α is the intercept in a regression that links the predictors with the responsibility judgments.

In the remainder of this paper, we report the results of four empirical studies, designed to answer the following questions: How do dispositional inferences and causal attributions relate to responsibility judgments of individuals in groups (Experiments 1 and 2)? How well does the model capture judgments of moral responsibility (Experiments 3 and 4)?.

4. Experiment 1: Testing dispositional inferences and causal attributions directly

In Experiment 1, participants' task was to judge to what extent a politician's vote on whether a new policy should be passed was (1) surprising and (2) important for the outcome. We predicted that participants' judgments of how surprising the vote of an individual committee member was would increase the greater the difference was between how the committee member was expected to vote (based on his party membership and on how the other committee members voted) and how the committee member actually voted. We predicted that judgments of how important an individual vote was would increase the closer the vote was to having been pivotal for the outcome, and the fewer committee members contributed to the outcome.

4.1. Methods

4.1.1. Participants

40 participants (M_{age} = 35, SD_{age} = 11, 10 female, 30 male) were recruited via Amazon Mechanical Turk. Participation was restricted to workers based in the US with a prior approval rate greater than 95% (see Mason & Suri, 2012, for details about how Amazon Mechanical Turk works).

4.1.2. Design

Experiment 1 included 27 voting scenarios. Each scenario featured a different political committee comprised of five members.[2] Between scenarios, we manipulated how each committee member voted, how many votes in favor of the policy were required for the policy to pass (1–5), the outcome of the vote (passed/ didn't pass), which political party supported the policy (Democrats/ Republicans) and the party affiliation of each committee member. Fig. 1 shows one scenario. For each scenario, we assessed importance and surprise judgments for one out of the five committee members. We selected 27 scenarios that elicit a range of predictions from our

[1] Note, however, that a recent study of real-life bystander intervention found that in most actual public conflicts, at least one person did something to help the victim (Philpot, Liebst, Levine, Bernasco, & Lindegaard, 2019).

[2] Note that unlike the example in Fig. 1, we used only male first names for the politicians within our actual experiments, in order to eliminate possible gender effects.

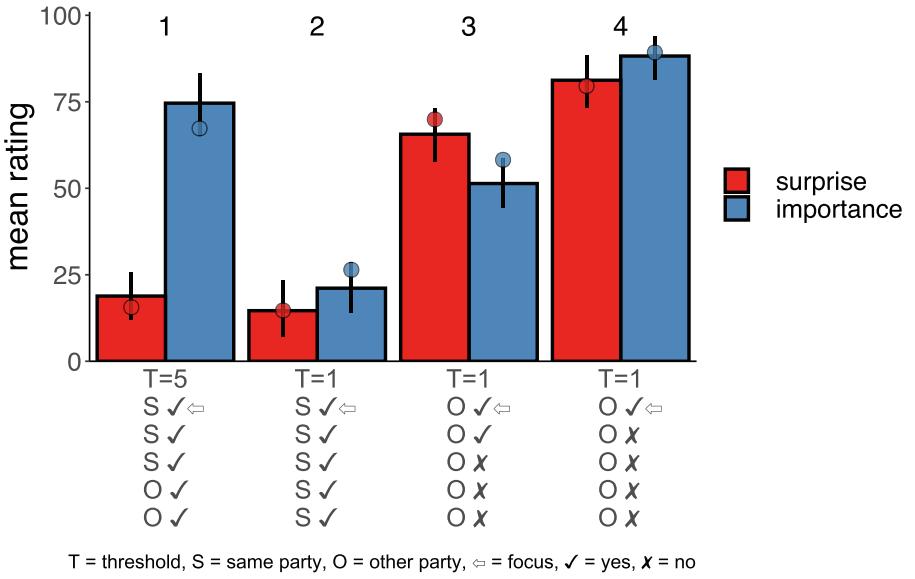

Fig. 3. Experiment 1: Importance and surprise judgments for Scenarios 1 to 4. Bars indicate mean judgments, error bars indicate bootstrapped 95% confidence intervals, and points indicate model predictions. The text on the x-axis shows what happened in each scenario. For example, in Scenario 3, the threshold T for the vote to pass was 1, the focus person (indicated by the arrow) was from the other party (O) that doesn't support the policy, and voted in favor of the policy. All other four committee members were also from the other party. One of them voted in favor, and three voted against the policy.

surprise and importance model.[3]

4.2. Procedure

The experiment was administered via *Qualtrics*. After receiving instructions, participants answered a set of comprehension check questions. Participants were redirected to the beginning of the survey in case they didn't correctly answer all of the comprehension check questions. Participants were then presented with the 27 voting scenarios in randomized order. For each scenario, participants were asked to judge the extent to which they considered one of the committee members' votes important and surprising. For example, when the committee member John had been described as having voted in favor of the policy and the policy passed, participants were asked: (1) "How important was John's vote for the policy passing?" and (2) "How surprising was John's vote?". Participants responded using continuous sliders whose endpoints were labeled with "not important at all" (0) and "very important" (100), as well as "not surprising at all" (0) and "very surprising" (100). On average, it took participants 13.67 min ($SD = 9.29$) to complete the experiment. All materials including data, experiments, model code and analysis scripts are available here: https://github.com/cicl-stanford/voting_responsibility

4.3. Results

We first describe participants' judgments for a selection of cases in detail, and then report their overall judgments.

4.3.1. Detailed analysis of a selection of scenarios

Fig. 3 shows the results of four of the voting scenarios. We chose these four scenarios with the goal of providing illustrative examples for how the manipulated variables affected participants' judgments. The figure shows participants' mean judgments together with the predictions of the surprise and importance model described above. In all four scenarios, the policy was passed because the number of votes in favor met or exceeded the threshold (T).

In all four scenarios, the committee member for whom ratings were assessed (the "focus person", indicated by the arrow in Fig. 3) voted in favor of the policy. In Scenario 1 and 2, the focus person was affiliated with the party that supported the policy, whereas in Scenario 3 and 4, the focus person was affiliated with the other party.

Participants were more surprised when a person voted in favor of a policy despite being from the opposite party. However, surprise

[3] See Table B1 in the Appendix for a full list of the scenarios.

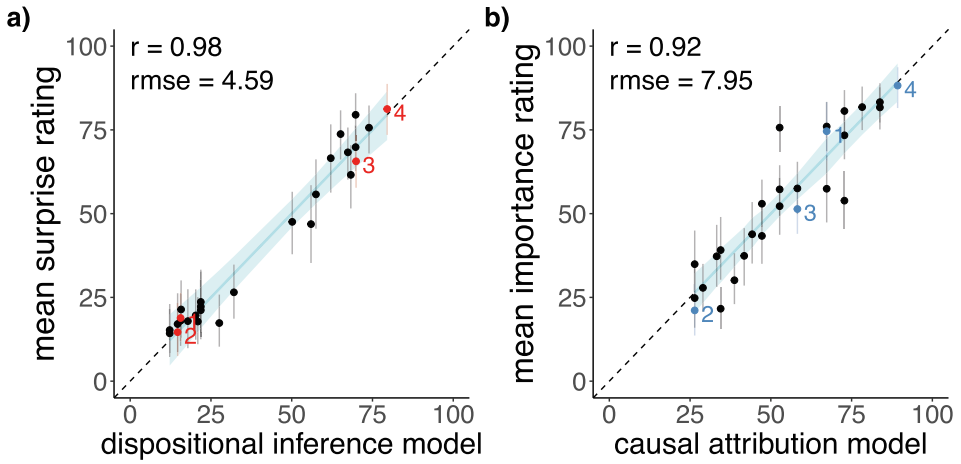

Fig. 4. Experiment 1: (a) Surprise judgments. (b) Importance judgments. Data points show mean judgments. The colored data points correspond to the four scenarios shown in Fig. 3. The blue ribbons show the 95% highest-density interval (HDI) for the model fit. The error bars indicate bootstrapped 95% confidence intervals. *Note*: r = Pearson's correlation, RMSE = root mean squared error.

judgments were not solely determined by whether a person's vote was consistent with their party affiliation. Participants were more surprised about the person's vote in Scenario 4 than in Scenario 3 (9.63 [7.83, 11.36]).[4] The model accurately captures this difference. The model assumes that a person's voting decision is determined not only by their party membership but also by the quality of the policy. A policy's quality can be inferred from how other committee members voted. While in Scenario 4 all others voted against the policy, in Scenario 3, one of the other committee members also voted in favor of the policy. As predicted by the model, participants were sensitive to this subtle difference in their surprise judgments.

We now consider participants' importance judgments. In Scenarios 1 and 4, the focus person's vote is pivotal. In both scenarios, had the focus person voted against the policy, the policy would not have passed. In Scenarios 2 and 3, the outcome is overdetermined. However, whereas in Scenario 2, all other committee members would have needed to vote differently in order for the focus person to become pivotal for the outcome, in Scenario 3, the focus person is only "one step away" from being pivotal. The focus person would have been pivotal if the second committee member had also voted against the policy. As predicted, participants judged the focus person's vote as more important the closer it was to being pivotal for the outcome. Participants' importance judgments are greater in Scenarios 1 and 4 than in Scenario 3 (20.05 [15.02, 25.15]), and greater in Scenario 3 than in Scenario 2 (31.79 [26.76, 36.83]).

If pivotality was the only factor that influenced people's judgments of importance, then varying the threshold while keeping pivotality fixed should not make a difference. That is, we should expect no difference in importance judgments between Scenario 1 and 4 since in both scenarios, the focus person's vote was pivotal for the outcome. However, participants considered the person's vote more important in Scenario 4 than in Scenario 1 (21.96 [16.89, 27.93]). This shows that participants' importance judgments are not solely determined by how close a person's vote was to having been pivotal for the outcome, but that it also matters how many causes contributed to the outcome. In Scenario 4, there was only a single cause for the policy passing – the focus member's vote. In contrast, in Scenario 1, there were five causes for the policy passing – all of the committee members' votes were required. A vote is seen as more important when it is the only cause versus just one of several causes. Our model of causal attribution which considers both pivotality and number of causes adequately captures participants' importance judgments.

4.3.2. Overall results and model comparison

Fig. 4 shows scatter plots of the model's predictions and participants' mean surprise and importance ratings for all 27 scenarios. We fitted the model to individual participants' responses by specifying a Bayesian linear mixed effects model with random intercepts and slopes for each predictor.

Our *dispositional inference model* captures participants' average surprise judgments very well with $r = .98$ and RMSE = 4.59 (Fig. 4a). A model that considers only whether the committee member voted in line with their party affiliation also correlates well with participants' judgments $r = .95$ and RMSE = 7.66. We compared the models using approximate leave-one-out crossvalidation as model selection criterion (PSIS-LOO; cf. Vehtari, Gelman, & Gabry (2017)). According to this criterion, the Bayesian surprise model performs better than the model that only considers party affiliation (difference in expected log predictive density (elpd) = 38.4, with a standard error of 16.1).[5]

[4] For any statistical claim, we report the mean of the posterior distribution together with the 95% highest-density interval (HDI). Here, for example, the posterior over the difference between Scenario 4 and 3 has a mean of 9.63, and the 95% HDI ranges from 7.83 to 11.36. The Bayesian models were written in Stan (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019).

[5] As a rule of thumb, a model is considered superior when the difference in expected log predictive density is greater than twice the standard error of that difference (for details, see Vehtari et al., 2017).

Fig. 4b shows that the *causal attribution model* accounts well for participants' mean importance judgments with $r = .92$ and RMSE $= 7.95$. The causal attribution model considers both the extent to which a person's action was pivotal for the outcome, as well as the number of causes that contributed to the outcome. This model compares favorably with lesioned models that only consider a subset of the predictors, such as just pivotality ($r = .88$ and RMSE $= 10.06$; elpd $= 37.4$, standard error $= 8.8$) or just the number of causes that contributed to the outcome ($r = .54$ and RMSE $= 17.54$; elpd $= 233.3$, standard error $= 23.7$).

4.4. Discussion

In this experiment, we presented participants with a number of different voting scenarios that manipulated how many votes were required for a particular policy to pass, the political affiliation of the committee members, how each committee member voted, and whether the policy passed (see Fig. 1). This information affected participants' judgments of how surprising and important a committee member's vote was. To explain participants' surprise judgments, we developed a dispositional inference model that forms an expectation about how a committee member would vote based on the committee members' party affiliations as well as how they voted. This model captures participants' surprise judgments well, and better than an alternative model that only considers a committee member's party affiliation.

Participants' judgments about how important a committee member's vote was for the outcome are well-explained by our causal attribution model. This model considers both how close a person's vote was to being pivotal for the outcome, as well as how many other committee members voted alike. A vote is seen as more important the closer it was to being pivotal (i.e., when the outcome of the overall vote would have been different had the committee member voted differently) and the fewer causes contributed to the outcome.

5. Experiment 2: Responsibility judgments in voting scenarios

In Experiment 1, we manipulated the extent to which a vote was surprising and its importance for the outcome, and assessed how this affected participants' dispositional inferences and their causal attributions. Our model predicts that both components contribute additively to responsibility judgments. Voter should be judged more responsible when their vote was suprising, when it was close to being pivotal, and when there were few causes that contributed to the outcome. To test these predictions, we presented participants with voting scenarios like those in Experiment 1, and asked to what extent committee members were *responsible* for the outcome of the vote.

5.1. Methods

5.1.1. Participants

208 participants ($M_{age} = 36$, $SD_{age} = 14$, 86 female, 122 male) were recruited via Amazon Mechanical Turk using Psiturk (Gureckis et al., 2016). Participation was again restricted to workers based in the US with a prior approval rate greater than 95%.

5.1.2. Design

We manipulated the size of the committee ($N = 3$ vs. $N = 5$), the political affiliations of the committee members (M_{p_i}), how each committee member voted (v_i), and the threshold for the policy to be passed (T). We aimed to test as many possible combinations of these different factors as possible. In principle, there would have been $2^3 \times 2^3 \times 3 + \times 2^5 \times 2^5 \times 5 = 5312$ different possible scenarios, taking into account the political affiliations, pattern of votes, and the different thresholds for committees of size 3 and 5. However, since the votes are being cast simultaneously, there are many scenarios that are symmetrical for our purposes. For example, if all of the committee members were Democrats, and two voted for the policy while one voted against it, we don't care which of the three voted against the policy. Taking into account these symmetries reduces the number of scenarios to 340.

We further reduced the number of scenarios by removing all scenarios for which the pattern of votes was unusual. A scenario is unusual if a majority of the committee voted against their political affiliation. For example, consider a scenario in which the policy is supported by the Democrats but all committee member are Republicans. Here, we removed all the scenarios in which more than 2 of the Republicans voted in favor of the policy. Removing all unusual scenarios reduces the number of scenarios to 170 (30 scenarios for committees of size 3, and 140 scenarios for committees of size 5).

We split the 170 scenarios into 10 different conditions with 17 scenarios each. Each condition included 3 scenarios with $N_{committee} = 3$, and 14 scenarios with $N_{committee} = 5$. This selection of scenarios included 24 of the 27 voting scenarios used in Experiment 1. The three scenarios from Experiment 1 that were dropped in Experiment 2 were Scenario 25, 26, and 27 (see Table B1 in the Appendix). In these scenarios, a majority of committee members voted against their party line.

5.2. Procedure

Participants were randomly assigned to one of 10 conditions. After receiving instructions, each participant made responsibility judgments for a set of 17 scenarios. Participants judged to what extent a particular committee member was responsible for the policy passing or not passing. Participants made their judgments on sliding scales ranging from "not at all responsible" (0) to "very much responsible" (100).

Participants assigned responsibility to committee members whose vote was in line with the outcome. Depending on the scenario,

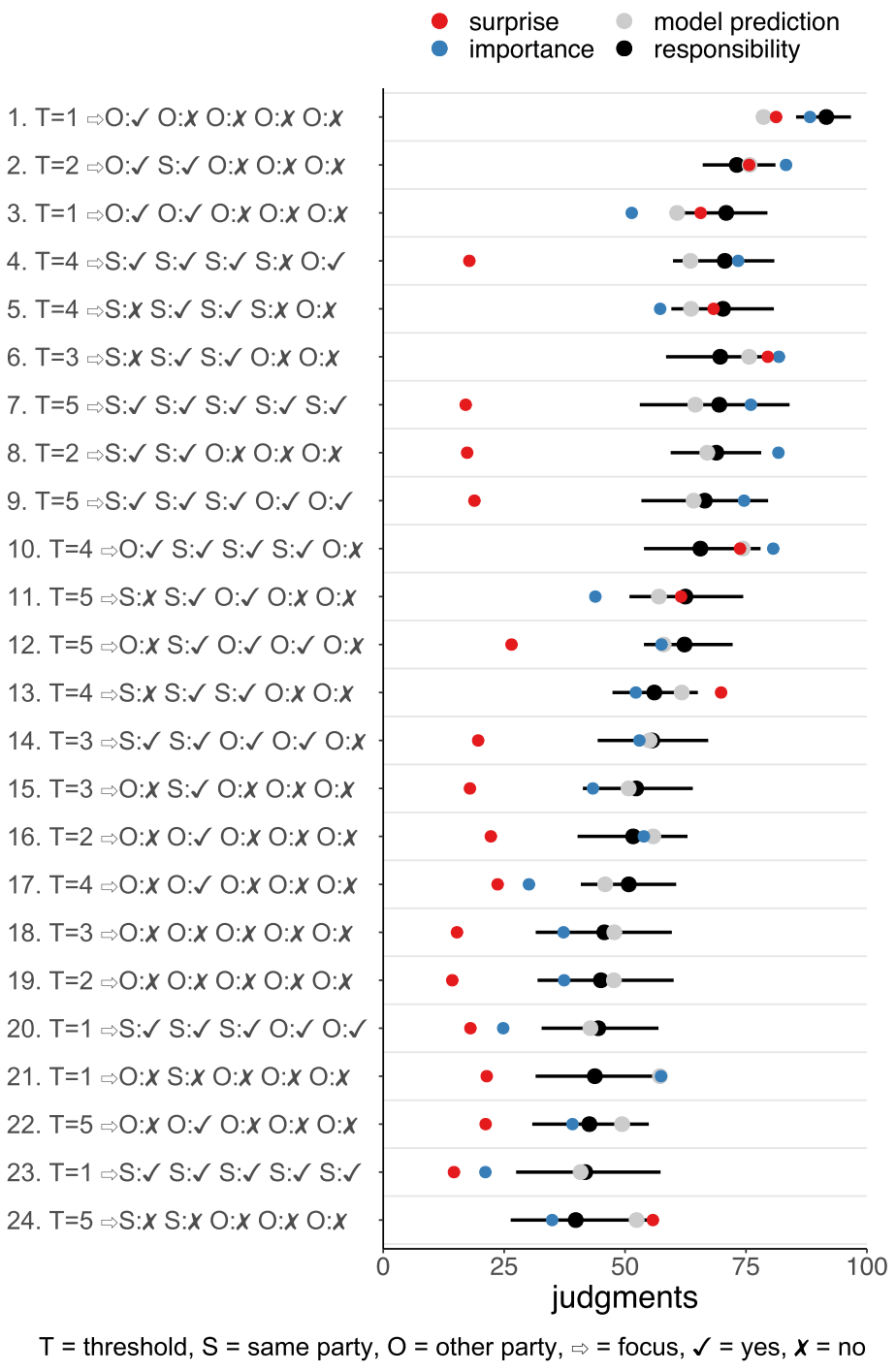

Fig. 5. Mean responsibility judgments (black dots) together with the mean surprise (red dots) and importance (blue dots) judgments based on Experiment 1, as well as the model prediction (gray dots) that combines surprise and importance judgments. We numbered the cases here in decreasing order of participants' mean responsibility judgments. *Note:* The error bars indicate bootstrapped 95% confidence intervals.

participants were either asked to make one or two judgments. When all committee members whose vote was in line with the outcome shared the same party affiliation, participants made only one judgment. When two of the committee members whose vote was in line with the outcome came from different political parties, then participants were asked to judge the responsibility for one of the Democrats and one of the Republicans. Out of the set of 170 scenarios, there were 90 scenarios in which participants were asked to make a single judgment, and 80 scenarios in which they made responsibility judgments for two committee members. Thus, we have a total of

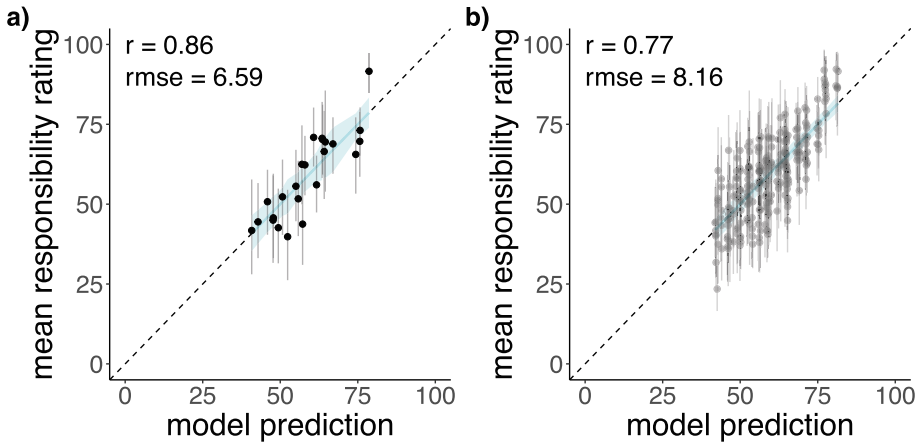

Fig. 6. Experiment 2: (a) Model predictions for a *selection of cases* based on participants' surprise and importance judgments in Experiment 1. (b) Model predictions for the *full set of cases* based on considering surprise, pivotality, and the number of causes as predictors. The scatter plots show model predictions (x-axis) and mean responsibility judgments (y-axis). The blue ribbons indicate the 95% HDI for the regression lines. The error bars indicate bootstrapped 95% confidence intervals. (*Note*: r = Pearson's correlation, RMSE = root mean squared error.).

250 data points. In our example scenario depicted in Fig. 1, two voters voted in line with the outcome of the vote (policy not passed): Allie and Dalia. Because Allie and Dalia came from different political parties, we assessed responsibility judgments for both of them. On average, it took participants 5.96 min ($SD = 5.17$) to complete the experiment.

5.3. Results

We first discuss a selection of cases before examining the data on a higher level of aggregation to see whether, and to what extent, participants' responsibility judgments were influenced by dispositional inferences and causal attributions.

5.3.1. Detailed analysis of a selection of cases

Fig. 5 shows participants' judgments for 24 of the 170 scenarios. These 24 scenarios are the ones that we also used in Experiment 1. The figure shows participants' mean responsibility judgments in addition to the mean surprise and importance judgments from Experiment 1, as well as the predictions of a model that uses participants' surprise and importance judgments from Experiment 1 to predict participants responsibility judgments in the current experiment. For example, in the first scenario, the threshold for the policy passing was one ($T = 1$), and all the committee members were from the party other than the one that supported the policy (O). The policy passed because one of the committee members voted in favor of the policy. We see that in this case, participants in Experiment 1 considered the committee member's action very surprising, and also judged that the vote was very important. Here, in Experiment 2, participants judged the responsibility of the committee member to be very high which is captured by the model.

In Scenario 24, the threshold was 5, but all committee members voted against the policy. Two members were affiliated with the party that supported the policy, and three were affiliated with the other party. Participants in Experiment 1 found it somewhat surprising that the focus person didn't vote for the policy even though he was from the party that supported the policy. Note, however, that they found this less surprising than what the focus person did in Scenario 1 (who also voted against the party affiliation). In Scenario 1, all other committee members voted against the policy, and the focus member was the only one voting in favor. In Scenario 24, all of the committee members voted against the policy, thus making the action of the focus member less surprising.

Participants in Experiment 1 judged that the focus person's action was not particularly important in Scenario 24. His vote was far from being pivotal (all of the other four votes would have needed to change), and there were four other causes of the outcome. Participants in Experiment 2 judged that the focus person in Scenario 24 was not very responsible for the outcome. Again, the model captures this case quite well.

To derive the model predictions for the 24 scenarios used in both Experiment 1 and 2, we used participants' mean surprise and importance judgments from Experiment 1 as predictors in a Bayesian linear mixed effects model of participants' responsibility judgments in Experiment 2, with both random intercepts and slopes. The model accounts well for the responsibility judgments across the 24 scenarios, as shown in Fig. 5 with $r = .86$ and RMSE $= 6.59$. The 95% HDI of the posterior for the surprise predictor ($\beta_{\text{surprise}} = 0.14$ [0.01, 0.26]) and the importance predictor ($\beta_{\text{importance}} = 0.43$ [0.27, 0.57]) both exclude 0. Fig. 6a shows a scatter plot of the model predictions and participants' responsibility judgments. Using the surprise and importance models that were fitted to participants' judgments in Experiment 1 as predictors yields a similar fit to participants' responsibility judgments, with $r = .85$ and RMSE $= 6.76$ (posterior estimates for the surprise ($\beta_{\text{surprise}} = 0.20$, 95% HDI [0.08, 0.32]) and importance predictor ($\beta_{\text{importance}} = 0.41$ [0.27, 0.55]).

Overall, we see that participants' responsibility judgments for this selection of 24 scenarios were both affected by how surprising a committee member's vote was, and how important the vote was for the outcome. We now consider how well the model captures participants' responsibility judgments across the whole range of scenarios.

Table 1
Estimates of the posterior mean, standard error, and 95% HDIs of the different predictors in the Bayesian mixed effects model. *Note*: `n_causes` = number of causes. `responsibility ~ 1 + surprise + pivotality + n_causes + (1 + surprise + pivotality + n_causes | participant)`.

term	estimate	std.error	lower 95% HDI	upper 95% HDI
intercept	59.94	3.25	54.70	65.22
surprise	21.68	4.57	14.17	29.23
pivotality	13.52	1.82	10.47	16.53
n_causes	−5.72	0.50	−6.55	−4.90

5.3.2. Overall results and model comparison

In order to apply the model to the full set of cases, we took the predictor that is relevant for the dispositional inference component of the model (i.e., the surprise model) and those that are relevant for the causal attribution component (i.e., pivotality and the number of causes). We then computed a Bayesian mixed effects model with random intercepts and slopes to predict participants' responsibility judgments (see Table 1). Fig. 6b shows a scatter plot of the model predictions and participants' responsibility judgments for the full set of 170 scenarios (with 250 judgments). Overall, the model predicts participants' responsibility judgments well with $r = .77$ and RMSE $= 8.16$. Table 1 shows the estimates of the different predictors. None of the predictors' 95% HDIs overlap with 0.

To investigate further whether the different model components are needed to capture participants' judgments, we constructed two lesioned models: one that considers only the dispositional inference part (i.e., using only `surprise` as a predictor), and one that considers only the causal attribution part (i.e., using only `pivotality` and `n_causes` as predictors).

The model that considers only surprise as a predictor performs markedly worse, with $r = .29$ and RMSE $= 12.36$. A model that considers only pivotality and the number of causes performs relatively well, with $r = .76$ and RMSE $= 8.37$. Comparing models with approximate leave-one-out crossvalidation as the model-selection criterion shows that a model that includes `surprise` as a predictor performs better than a model that considers only `pivotality` and `n_causes` as predictors (difference in expected log predictive density (elpd) = 89.4, with a standard error of 16.1). A model that, in addition to the predictors discussed here, also considers whether the outcome was positive or negative (i.e., whether the policy passed), does an even better job at predicting participants' responsibility judgments with $r = .81$ and RMSE = 7.60 (difference in elpd = 70.7, with a standard error of 14.6, compared to the model without the `outcome` predictor). Participants assigned more responsibility when the outcome was positive (i.e., when a committee member voted in favor of a policy) than when the outcome was negative (and a committee member voted against a policy).

5.4. Discussion

In this experiment, we asked participants for responsibility judgments about individual committee members across a large set of voting scenarios. Our computational model captured participants' judgments well. While previous work has shown that the model accounts well for responsibility judgments about individual decision-makers in achievement contexts (Gerstenberg et al., 2018), the results of this experiment demonstrate that the model also captures responsibility judgments in more complex group settings. While the responsibility judgments obtained in previous work were consistent with the key processes that the model postulates (dispositional inference and causal attribution), the results of Experiments 1 and 2 together provide a much stronger test of this proposal. Participants' surprise and importance judgments in Experiment 1 predict the responsibility judgments in Experiment 2.

In addition to the factors that our model considers, we found that participants' responsibility judgments were affected by whether the outcome was positive (i.e., the policy was passed) or negative (i.e., the policy was not passed). Participants assigned more responsibility to committee members who had voted for rather than against a policy. This finding is in line with literature showing that people tend to judge actions to be worse than omissions, even if the latter lead to similar consequences (see, e.g., Ritov & Baron, 1992; Baron & Ritov, 2004; Kahneman, Slovic, & Tversky, 1982; Byrne & McEleney, 2000). This notion is also reflected in legal decision making; in many countries, for example, passive euthanasia is tolerated while active euthanasia is heavily sentenced (Rachels, 2007). In our voting setting, participants may have assumed that committee members vote "yes" only if they really agree with the policy, while voting "no" is compatible both with being against the policy and with having no strong opinion. This effect was not predicted by our model, but could in principle be accommodated by it by shifting the prior on how likely people are to vote for versus against a policy.

The results showed that while both components of the model are important, participants' responsibility judgments were most strongly influenced by the causal attribution component of our model. However, as we discussed earlier, the extent to which dispositional inferences play a role for responsibility judgments might differ between domains. In Experiment 3, we test the idea that in the moral domain, people may place more emphasis on dispositional inferences when judging responsibility.

6. Experiment 3: Responsibility and moral wrongfulness judgments in moral contexts

Gerstenberg et al. (2018) previously tested the computational model in achievement contexts, where the outcome critically depended on an individual's skill. Achievement contexts naturally elicit judgments of responsibility, as one can witness in any sports bar. However, judgments of responsibility are also particularly relevant in the moral domain. Research in moral psychology has shown that when people make moral judgments, they often assign more weight to those features of a behavior that seem most informative of

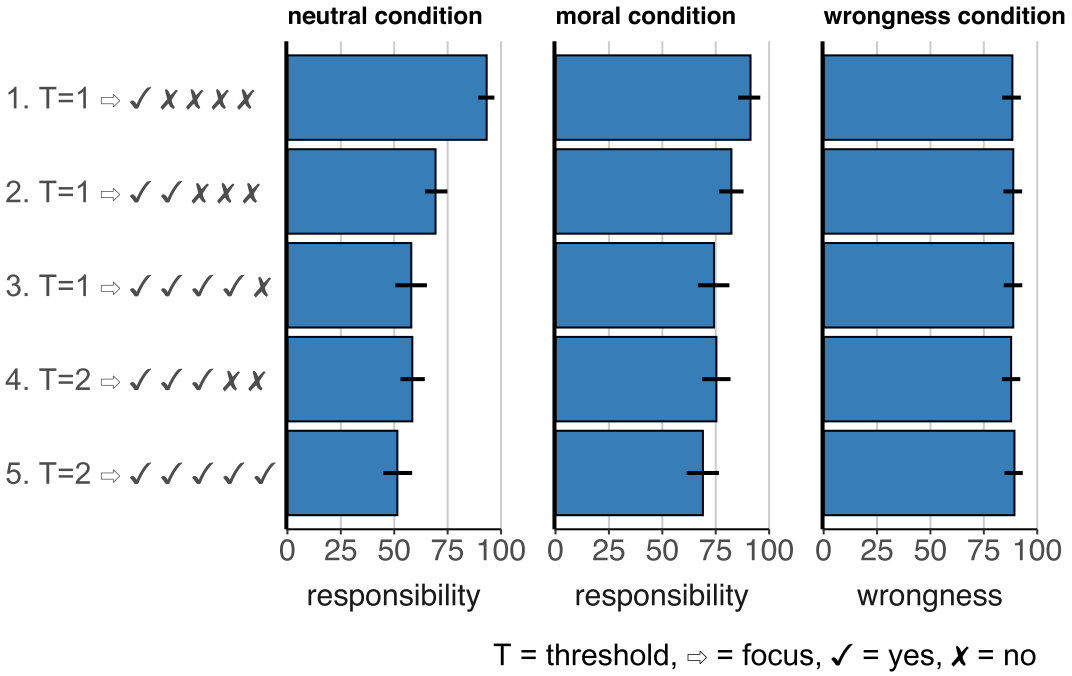

Fig. 7. Experiment 3: Participants' mean responsibility ratings in the *neutral* and *moral* condition, as well as their mean wrongfulness ratings in the wrongfulness condition. *Note*: The error bars indicate bootstrapped 95% confidence intervals.

an agent's character; arguably because it helps us to identify partners for future collaboration (Bartels & Pizarro, 2011; Bayles, 1982; Cushman, 2008; Gerstenberg, Lagnado, & Kareev, 2010; Schächtele, Gerstenberg, & Lagnado, 2011; Pizarro et al., 2003; Uhlmann et al., 2015; Waldmann et al., 2012). With regard to our computational model, this means that when people make moral responsibility judgments, the relative weights they assign to dispositional inferences versus causal attributions may shift such that dispositional inferences play a relatively larger role.

Moreover, in the moral domain, people are not only concerned with judging responsibility (see Malle, 2021, for a recent review of different moral judgments). They are also motivated to determine whether the person's action was generally right or wrong (e.g. Haidt, 2001) and if so, whether the person should be punished for her wrongdoing (e.g. Darley, 2009). Cushman (2008) demonstrated that judgments of blame and punishment rely jointly on the agent's mental states and the causal connection of an agent to a harmful consequence, while judgments of the wrongfulness or permissibility of an action rely predominantly on the mental states of an agent. Regarding our computational model, this suggests that when people evaluate to what extent an agent's action was morally wrong, the focus might be on dispositional inferences, while causal attributions might matter relatively less.

In Experiment 3, we manipulated the moral valence of the policy that the committees voted on. One group of participants assigned responsibility in a "morally neutral context", in which the committee members voted on a policy to change the font of all government documents to *Arial*. A second group of participants assigned responsibility in a "morally negative context". Here, the policy was a request to reintroduce corporal punishment, such as spanking or paddling, in schools. We hypothesized that causal attributions would affect participants' responsibility judgments in both conditions, but that they would play a smaller role in the morally negative condition than in the morally neutral condition. Further, we expected that since people generally assume that others have positive desires and intentions, a vote for a morally negative policy would be considered more surprising. Specifically, we hypothesized that the committee member's immoral vote in the morally negative condition would be more surprising for participants than the committee member's vote in favor of a certain font in the morally neutral condition, and thus that the impact of dispositional inferences in the morally negative condition would be larger.

To test how dispositional inferences and causal attributions affect judgments about the moral wrongfulness of an action, Experiment 3 included a third condition. In this condition, the policy was also a request to reintroduce corporal punishment in schools. However, instead of asking for responsibility judgments, we asked participants for the extent to which they considered the votes of particular committee members morally wrong. We predicted that when people evaluate the extent to which an individual's action was morally wrong, their judgment should be largely unaffected by what causal role the action played.

To sum up, we hypothesized that, first, for judgments of responsibility, the importance of causal attribution would be smaller in the morally negative than in the morally neutral condition. Second, we predicted that for judgments of moral wrongfulness, the causal attribution component of the model would not matter.

6.1. Methods

6.1.1. Participants

314 participants were recruited via Prolific (https://www.prolific.co). Inclusion criteria were English as native language and an approval rate of at least 90%. Experiment 3 involved an attention check and a manipulation-check question. Participants who answered either of these questions incorrectly were removed from the analysis, leaving 236 participants (159 female, 74 male, 3 unspecified, $M_{age} = 31$, $SD_{age} = 9$).

6.1.2. Design

Participants saw scenarios in which a political committee voted on whether or not a motion should be passed. We manipulated how each committee member voted and how many votes in favor of the policy were required for the policy to pass (1–5). We constructed five different voting scenarios whose structure is illustrated in Fig. 7. The size of the committee (5 members) and the outcome of the vote (policy passed) were held constant. The focus person always voted in favor of the policy. The focus person's causal contribution to the outcome varied based on how the remaining committee members voted. For example, in Scenario 2 in Fig. 7, the threshold for the policy to pass is 1. Since, in addition to the focus person, one other committee member ended up voting in favor of the policy, the focus person's pivotality is 0.5 and the outcome was caused by two votes.

In our previous experiments, we manipulated participants' expectations about how a committee member would vote by giving them information about the committee members' party affiliation and about which party supported the policy. Here, we manipulated participants' expectations about how a committee member would vote via information about the moral context of the vote. We also varied the test question. One group of participants assessed responsibility and another group judged the extent to which they considered a vote as morally wrong. Thus, Experiment 3 had three conditions: *neutral* (morally neutral context, responsibility judgments), *moral* (morally negative context, responsibility judgments) and *wrongfulness* (morally negative context, moral wrongfulness judgments).

6.2. Procedure

Experiment 3 was administered in *Unipark*, a German online survey platform. Participants were randomly allocated to one of three conditions. After receiving instructions, they were presented with five voting scenarios in randomized order. Participants gave responsibility or moral wrongfulness judgments, depending on the condition, for those members that had been described as having voted in favor of the motion (1–5 judgments, depending on the scenario). For example, participants in the morally neutral context condition read "To what extent is Dallas responsible for the font of all government documents being changed to Arial". Participants in the morally negative context condition read "To what extent is Dallas responsible for corporal punishment being introduced in schools". Participants in the moral wrongfulness condition read "To what extent is it morally wrong that Dallas voted in favor of introducing corporal punishment in schools?". Ratings were provided on a sliding scale ranging from "not at all responsible" or "not at all morally wrong" (0) to "very much responsible" or "very much morally wrong" (100) with an invisible starting point.

After having completed all five scenarios, participants answered a manipulation-check question that assessed whether the context manipulation was successful. Participants in the morally neutral context condition were asked "How do you morally judge voting in favor of changing government documents into Arial". Participants in the morally negative context condition and the moral wrongfulness condition were asked "How do you morally judge voting in favor of introducing corporal punishment in schools?" They could choose between the answer options "bad", "good" and "neutral". At the end of the survey, participants responded to an attention check question and reported their demographics.[6] On average, it took participants 5.63 min ($SD = 2.52$) to complete this experiment.

6.3. Results

Fig. 7 shows participants' responsibility judgments across the five different scenarios separately for the *neutral* and *moral* condition, and their moral wrongfulness judgments in the *wrongfulness* condition. Qualitatively, we can see that in the *neutral* condition, participants' responsibility judgments differentiate more between the different scenarios than in the *moral* condition. In the wrongfulness condition, participants' judgments were very high and didn't vary between the scenarios. Overall, responsibility judgments were higher in the moral condition ($M = 78.42, SD = 28.53$) than in the neutral condition ($M = 66.11, SD = 31.69$), and wrongfulness judgments were even higher ($M = 88.60, SD = 20.37$).

To test our prediction that the different experimental conditions affect the extent to which the causal attribution component of the model matters, we first compared a Bayesian mixed effects model with the two predictors that capture the causal attribution part of the model (`pivotality` and `n_causes`), and one model that additionally contains `condition` as predictor (as well as interactions between condition and the other predictors). Because we had only five observations for each participant, we included only random

[6] As expected, the majority of participants (91%) in the morally neutral context condition considered changing the font of government documents into Arial as neutral and the majority of participants (74% in *moral* and 79% in *wrongfulness*) in the morally negative context conditions judged introducing corporal punishment in schools as bad. 8 participants in the morally neutral context condition indicated that they considered changing the font of government documents into Arial as bad or good. 25 participants in the *moral* condition and 12 participants in *wrongfulness* condition answered that introducing corporal punishment would be neutral or good. These participants were excluded from subsequent analyses.

intercepts and no random slopes. A model that includes condition as a predictor performs better illustrating that judgments differed between conditions (difference in expected log predictive density(elpd) = 109.0, with a standard error of 15.9).

As predicted, the extent to which causal attributions affected participants' judgments differed between conditions. In particular, what role pivotality played in participants' judgments differed between conditions. To further explore what role causal attributions played in the different conditions, we ran separate Bayesian regressions for each condition with `pivotality` and `n_causes` as predictors, and random intercepts for participants. In the neutral condition, the estimates for the `pivotality` and `n_causes` predictor were $38.40[22.76, 52.98]$ and $-2.77[-5.75, 0.21]$, respectively. In the moral condition, they were $9.05[-6.16, 23.00]$ and $-3.71[-6.44, -0.91]$. Finally, in the wrongfulness condition, the estimates were $-0.10[-4.46, 3.94]$ and $0.17[-0.69, 1.00]$. Consistent with our hypothesis, pivotality had a less strong effect on participants' responsibility judgments in the moral condition than in the neutral condition. Further, it didn't affect participants' wrongfulness judgments at all in the wrongfulness condition.

6.4. Discussion

In this experiment, we tested the predictions of our computational model in the moral domain. We used a similar setup as in Experiments 1 and 2, with individuals in a group voting for an outcome. Instead of manipulating voting expectations via information about the committee members' party affiliation, this time we manipulated information about the moral content and the consequences of the policy, as well as the question that participants were asked to evaluate. We predicted that in a morally negative context, judgments of responsibility would be less sensitive to the causal role that a person's action had for bringing about the outcome, and more strongly affected by what dispositional inference is licensed based on observing the action. We further predicted that judgments of moral wrongfulness would not be at all affected by the causal role of the person.

How pivotal a person's vote was affected participants' responsibility judgments in the neutral condition and in moral condition, albeit to a lesser extent. We didn't assess dispositional inferences about the committee members directly in this study (via surprise ratings). However, the fact that pivotality influenced responsibility judgments less in the moral compared to the neutral condition suggests a shift of emphasis in what participants cared about depending on the domain. In the moral condition, dispositional inferences are particularly important as they reveal whether someone is "a bad person" (cf. Uhlmann et al., 2015). Our results are also in line with research showing that both adults and children place a heavier emphasis on an actor's mental states when evaluating moral rule transgressions than when evaluating conventional rule transgressions (Giffin & Lombrozo, 2015; Josephs, Kushnir, Gräfenhain, & Rakoczy, 2016).

In the *wrongfulness* condition, pivotality didn't influence participants judgments (see Darley, 2009; Cushman, 2008). The extent to which a person's action was judged as morally wrong doesn't depend on what causal role it played. While people might try to excuse their behavior by pointing out that other people behaved equally badly, our results indicate that this might not be the most efficient strategy (Green, 1991; Falk & Szech, 2013).

7. Experiment 4: Assessing importance, surprise, responsibility, and moral wrongfulness within participants in moral contexts

In Experiment 4, we draw together key elements from the prior experiments for a comprehensive test of the model. Like in Experiment 1, we directly assessed the model's two main components, dispositional inferences and causal attributions, by having participants rate the surprisingness and importance of individual votes, respectively. Like in Experiment 2, we investigated whether and how these factors affect responsibility judgments. Further, we tested our model in a moral context as in Experiment 3, assessing both responsibility and judgments of wrongfulness. In the previous experiments, we evaluated the model on aggregate responses. This time, we assessed all of the questions within participants which allowed us to evaluate the model on the individual participant level.

7.1. Methods

7.1.1. Participants

50 participants ($M_{age} = 32$, $SD_{age} = 11$, 29 female) were recruited via Prolific. As in Experiment 3, inclusion criteria were English as native language and an approval rate of 90% or higher.

7.1.2. Design

Each participant viewed 24 voting scenarios in randomized order. In each of these scenarios, at least two votes in favor of the policy were required in order for a policy to pass. The outcome of the vote was always that the policy passed. Between scenarios, we varied the *content of the policy* that was up for vote, and the *number of committee members who voted in favor* of the policy. We arranged the 24 voting scenarios into four sets of six based on our intuition, such that each set could be thought of as spanning a scale from "not morally bad at all" (0) to "very morally bad" (6). We varied whether two, three, four, or all five of the committee members voted in favor of the policy, resulting in 4 different pivotality values; 1, $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$, respectively. We crossed content of the policy with the number of votes, yielding a *policy* (6 levels) $\times$ votes (4 levels) design.

7.2. Procedure

Experiment 4 was administered in *Qualtrics*. After providing consent and receiving instructions, participants answered three comprehension check questions. Participants were redirected to the beginning of the survey in case they didn't answer all three questions correctly. If participants failed any of the comprehension check questions a second time, they were not able to participate in the experiment. After passing the comprehension check questions, participants were presented with a practice scenario, followed by the 24 test scenarios in randomized order. For each scenario, participants first learned about the content of the policy. For example, participants read "Policy up for vote next: Ban children who are not vaccinated against Varicella from going to school.".

For each scenario participants were asked the following five questions:

1. ***Badness***: How morally bad is it to ban children who are not vaccinated against Varicella from going to school? [not morally bad at all – very morally bad]
2. ***Importance***: How important was Tim's vote for the policy passing? [not important at all – very important]
3. ***Surprise***: How surprising was Tim's vote? [not surprising at all – very surprising]
4. ***Responsibility***: To what extent was Tim responsible for the policy passing? [not responsible at all – very responsible]
5. ***Wrongfulness***: To what extent is it morally wrong that Tim voted in favor of banning children who are not vaccinated against Varicella from going to school? [not morally wrong at all – very morally wrong].

Participants made the 'badness' judgment after having learned about the content of the policy. Participants then received information about how each of the five committee members voted on the policy, and answered the remaining four questions. All four questions and sliders were shown on the same screen underneath a table that showed the votes. The order of the 'importance' and the 'surprise' question, and of the 'responsibility' and 'wrongfulness' question was counterbalanced between participants. The 'surprise' and 'importance' questions were always asked before the other two questions. Between scenarios, the questions were adapted depending on who participants were asked to evaluate, and on what the policy was. Like in previous experiments, the sliders ranged from 0 to 100 and the slider handle only appeared once a participant clicked on the slider track. On average, it took participants 28.95 min ($SD = 14.73$) to complete the experiment. Participants were compensated with $5.50.[7]

7.3. Hypotheses

We pre-registered the following five hypotheses on the Open Science Framwork (OSF) (see pre-registration here https://osf.io/qdz54). We also pre-registered which statistical models we would run to test each hypothesis as detailed in the results section below.

7.3.1. Hypothesis 1: The fewer politicians voted in favor of the passed policy, the more important a politician's vote who voted in favor is judged to be

The fewer politicians voted in favor of the policy, the closer each individual politician's vote was to having been pivotal for the policy's passing. We predicted that the judged importance of a committee member's vote increases with how close it was to being pivotal (see Eq. 2).

7.3.2. Hypothesis 2: The more morally negative a policy is perceived, the more surprising a vote in favor of that policy is judged to be

People generally assume that others have positive moral desires and intentions (De Freitas, Cikara, Grossmann, & Schlegel, 2017). Accordingly, we predicted that the more morally bad a policy is perceived to be, the more surprising a vote in favor of that policy will be judged.

7.3.3. Hypothesis 3: Both judgments of importance and surprise predict judgments of responsibility

In Experiments 2 and 3, we found that both importance and surprise predicted judgments of responsibility. In Experiment 4, we predicted that both of these factors would affect responsibility judgments, too.

7.3.4. Hypothesis 4: When judging responsibility, importance matters more for policies that are perceived as morally neutral (compared to morally negative policies), and surprise matters more for policies that are perceived as morally negative (compared to morally neutral policies)

In Experiment 3, we found that importance mattered more for responsibility judgments about a morally neutral policy compared to a morally negative policy (see Fig. 7). Accordingly, we predicted that importance would matter more for morally neutral compared to morally negative policies. Conversely, we predicted that surprise would matter more for negative compared to neutral policies.

7.3.5. Hypothesis 5: Perceived importance matters more for judgments of responsibility compared to wrongfulness judgments.

Based on previous work (Cushman, 2008), we predicted that importance matters more for responsibility compared to wrongfulness judgments. Cushman (2008) found the causal role that an agent's action played in bringing about the outcome mattered more for judgments of blame compared to judgments of wrongfulness.

[7] Tables E1 and E2 in the Appendix show the 24 different scenarios as well as the mean ratings for each question in each scenario.

Table 2
Estimates of the posterior mean, standard error, and 95% HDIs of the different predictors in the Bayesian mixed effects model. *Note*: Pivotality is on a scale from 1 = pivotal in the actual situation to 0 = not pivotal at all (see Eq. 2). `importance ~ 1 + pivotality + (1 + pivotality | participant)`.

term	estimate	std.error	lower 95% HDI	upper 95% HDI
intercept	22.48	4.85	13.22	32.11
pivotality	65.68	5.72	54.32	76.67

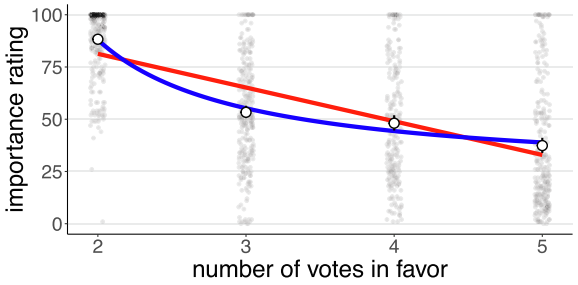

Fig. 8. Experiment 4: Judgments of how important a committee member's vote in favor of a policy was, as a function of how many committee members voted in favor. Two votes were required for a policy to pass. *Note*: White circles show mean judgments, with 95% bootstrapped confidence intervals. Small black circles show individual judgments (jittered along the x-axis for visibility). The red line shows a linear fit to the data using the number of votes as the predictor (`importance ~ 1 + n_votes`). The blue line shows a fit that uses pivotality as a predictor ($\texttt{importance} \sim 1 + \frac{1}{\texttt{n\_votes}-1}$). For example, because two votes are required for a policy to pass, pivotality is 1 if `n_votes` = 2, or 1/4 if `n_votes` = 5.

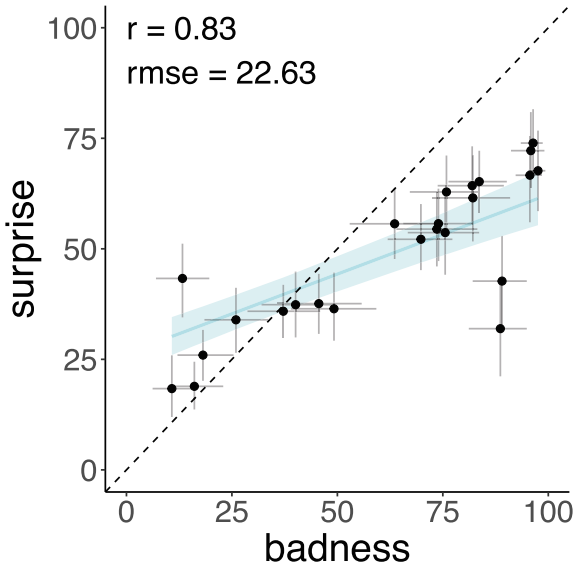

Fig. 9. Experiment 4: Judgments about how surprising a vote in favor of a policy was as a function of how bad a policy was judged. The points show mean judgments across the 24 scenarios. Error bars are 95% bootstrapped confidence intervals. The regression line shows the predictions of the Bayesian mixed effects model as described in Table 3, and the blue ribbon shows the 95% highest-density interval (HDI) o.f the model fit.

7.4. Results

We separate the discussion of our results into a confirmatory analysis section in which we report the pre-registered analysis, and an exploratory analysis in which we discuss additional analysis that we didn't pre-register.

7.4.1. Confirmatory analysis

In this section, we report statistical tests of our pre-registered hypotheses.

Hypothesis 1: The fewer politicians voted in favor of the policy, the more important a politician's vote who voted in favor is judged to be

Table 3
Estimates of the posterior mean, standard error, and 95% HDIs of the different predictors in the Bayesian mixed effects model. `surprise ~ 1 + badness + (1 + badness | participant)`.

term	estimate	std.error	lower 95% HDI	upper 95% HDI
intercept	26.30	2.42	21.67	31.13
badness	0.36	0.04	0.28	0.43

Table 4
Pairwise correlations between variables. *Note*: Pivotality was manipulated experimentally. The remaining variables were assessed empirically. The values show Pearson correlations calculated on the mean judgments for each trial.

	responsibility	wrongfulness	surprise	importance	badness
wrongfulness	.04				
surprise	.41	.78			
importance	.99	-.03	.35		
badness	.11	.96	.83	.04	
pivotality	.96	-.18	.21	.97	-.10

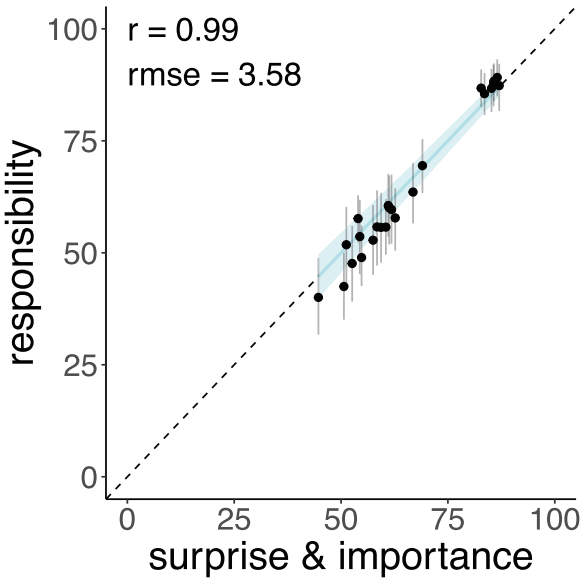

Fig. 10. Experiment 4: Judgments about how responsible a vote was for the policy passing as a function of how surprising and how important that vote was judged to be. The points show mean judgments across the 24 scenarios. Error bars are 95% bootstrapped confidence intervals. The regression line shows the predictions of the Bayesian mixed effects model as described in Table 5, and the blue ribbon shows the 95% highest-density interval (HDI) fo.r the model fit.

Table 5
Estimates of the posterior mean, standard error, and 95% HDIs of the different predictors in the Bayesian mixed effects model. `responsibility ~ 1 + importance + surprise + (1 + importance + surprise | participant)`.

term	estimate	std.error	lower 95% HDI	upper 95% HDI
intercept	25.12	3.63	18.17	32.62
importance	0.65	0.04	0.57	0.72
surprise	0.05	0.01	0.03	0.08

As predicted, participants' importance judgments decreased with pivotality (see Table 2). Fig. 8 shows importance ratings as a function of how many committee members voted in favor of the policy. The more committee members voted in favor of a policy, the less important each vote was perceived to be. Importance ratings didn't decrease linearly with the number of votes (the red line), but rather as a function of how close a person's vote was to having been pivotal for the outcome (the blue line, see Eq. 2). Because the threshold for a policy to pass was 2 votes, each committee member's vote was pivotal when only two votes were cast in favor of the policy.

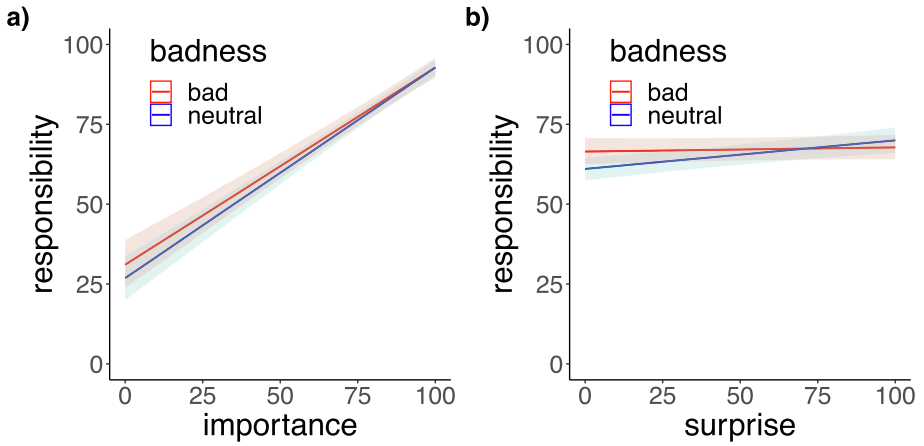

Fig. 11. Experiment 4: (a) Predicted relationship between importance and responsibility. (b) Predicted relationship between surprise and responsibility. The predictions are shown separately for morally bad and morally neutral policies. Table 6 shows the u.nderlying model.

Table 6
Estimates of the posterior mean, standard error, and 95% HDIs of the different predictors in the Bayesian mixed effects model. *Note*: The `badness_dummy` variable was coded as neutral $= -0.5$, and bad $= 0.5$. `responsibility ~ 1 + badness_dummy * (importance + surprise) + (1 + badness_dummy * (importance + surprise) | participant)`.

term	estimate	std.error	lower 95% HDI	upper 95% HDI
intercept	26.45	3.59	19.45	33.52
badness_dummy	−7.92	2.09	−12.12	−4.07
importance	0.64	0.04	0.56	0.71
surprise	0.05	0.02	0.02	0.08
badness_dummy:importance	0.04	0.03	−0.01	0.10
badness_dummy:surprise	0.08	0.03	0.02	0.13

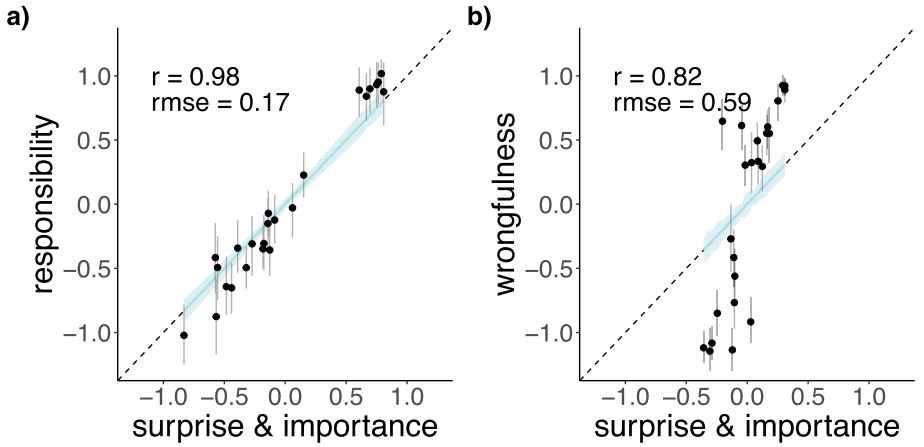

Fig. 12. Experiment 4: Relationship between surprise and importance judgments with (a) responsibility judgments, and (b) wrongfulness judgments. *Note*: Each variable was z-scored on the individual participant level. Points show mean ratings for each of the 24 scenarios. Error bars show 95% bootstrapped confidence intervals. The regression lines shows the prediction of the two Bayesian mixed effects models as described in Table 7, and the blue ribbons show the 95% highest-density interval (HDI) for. each model fit.

Hypothesis 2: The more morally negative a policy is perceived, the more surprising a vote in favor of that policy is judged to be

As predicted, a committee member's vote in favor of a policy was judged to be more surprising the more that policy was judged to be morally bad (see Fig. 9 and Table 3). Table 4 shows the pairwise correlations between the different ratings, including the correlation between judged 'badness' and 'surprise' ($r = 0.83$).

Hypothesis 3: Both judgments of importance and surprise predict judgments of responsibility.

Table 7
Estimates of the posterior mean, standard error, and 95% HDIs of the different predictors in the Bayesian mixed effects model. *Note*: To compare the estimates in the different models, we z-scored judgments of all the variables on the participant level. `responsibility ~ 1 + importance + surprise + (1 + importance + surprise | participant) wrongfulness ~ 1 + importance + surprise + (1 + importance + surprise | participant)`.

response	term	estimate	std.error	lower 95% HDI	upper 95% HDI
responsibility	importance	0.72	0.04	0.63	0.80
responsibility	surprise	0.09	0.02	0.04	0.13
wrongfulness	importance	−0.06	0.04	−0.15	0.01
wrongfulness	surprise	0.43	0.05	0.33	0.53

Table 8
Experiment 4: Summaries of fitting different Bayesian regression models to individual participants' responsibility (top) and wrongfulness judgments (bottom). *Note*: n = number of participants best fit by each model. The r and rmse columns show medians with 25% and 75% quantiles of the correlation and root mean squared error between judgments and model predictions. For example, 31 participants' responsibility judgments were best explained by a model only considering 'importance' as predictor. The median correlation between the model's predictions and these participants' judgments was $r = 0.86$, with the 25% quantile being 0.64 and the 75% quantile being 0.92.

model	formula	n	r	rmse
baseline	`responsibility ~ 1`	5	—	52.47 [49.78, 57.8]
importance	`responsibility ~ 1 + importance`	31	0.86 [0.64, 0.92]	9.87 [7.89, 13.22]
surprise	`responsibility ~ 1 + surprise`	2	0.45 [0.35, 0.54]	9.8 [9.03, 10.57]
importance_surprise	`responsibility ~ 1 + importance + surprise`	11	0.9 [0.86, 0.97]	9.1 [6.59, 13.91]
baseline	`wrongfulness ~ 1`	12	—	39.42 [32.97, 44.28]
importance	`wrongfulness ~ 1 + importance`	4	0.45 [0.35, 0.68]	33.22 [10.56, 37.99]
surprise	`wrongfulness ~ 1 + surprise`	32	0.61 [0.36, 0.88]	32.08 [18, 40.97]
importance_surprise	`wrongfulness ~ 1 + importance + surprise`	1	0.53 [0.53, 0.53]	8.87 [8.87, 8.87]

As predicted, both judgments of importance and surprise were associated with judgments of responsibility. Fig. 10 shows the predicted responsibility judgments of a model that considers both surprise and importance. As Table 5 shows, both surprise and importance affected participants' responsibility judgments. As the estimates for each predictor reveal, importance mattered more than surprise for predicting responsibility. This is also reflected in the high pairwise correlations between importance, surprise, and responsibility (see Table 4).

Hypothesis 4: When judging responsibility, importance matters more for policies that are perceived as morally neutral, and surprise matters more for policies that are perceived as morally negative.

For this analysis, we did a median split on the mean badness ratings across the 24 scenarios, separating policies into 'bad' and 'neutral'. We then used this categorical predictor to estimate the effect of surprise and importance on the judged responsibility for bad versus neutral policies.

Fig. 11a shows that the was a positive relationship between importance and responsibility for both bad and neutral policies. Unlike predicted, the relationship between importance and responsibility was not stronger for neutral compared to bad policies (the 95% HDI of the interaction between badness and importance includes 0, see Table 6).

Fig. 11b shows the relationship between surprise and responsibility for neutral and bad policies. We predicted that surprise would matter more when judging responsibility of voting for bad compared to neutral policies (i.e. a negative interaction between badness and surprise). This was not the case (see Table 6).

Hypothesis 5: Perceived importance matters more for judgments of responsibility compared to wrongfulness judgments.

Fig. 12 shows how well a model that considers surprise and importance accounts for participants' responsibility and wrongfulness judgments. As predicted, importance mattered more for responsibility than for wrongfulness (see Table 7). We correctly predicted that the 95% HDI of importance when predicting responsibility would not overlap with the 95% HDI of importance when predicting wrongfulness. While importance matters more for responsibility than for wrongfulness, surprise matters more for wrongfulness than for responsibility judgments.

7.4.2. Exploratory analysis

In Experiment 4, we assessed the different components of the model within participants. This means we can test on the level of individual participants, whether importance and surprise predict responsibility judgments. For each participant, we compared four different models to see which one accounted best for their responses: A baseline model that just predicts the mean. A model with only importance as a predictor. A model with only surprise as a predictor, and a model that considers both importance and surprise. For each participant, we used their own judgments of surprise and importance as predictors for their responsibility judgments. We placed a strictly positive prior on the regression weights of all the models. This means that solutions in which a negative weight was assigned to any of the predictors weren't possible.

Table 8 shows how well the different models captured individual participants' responsibility judgments. We performed approximate leave-one-out crossvalidation to determine which model best accounted for each participants' judgments. Overall, 31

participants' judgments were best accounted for by the model that only considers importance as a predictor, 11 participants by the model with both importance and surprise as predictors, 5 participants by the baseline model, and 2 participants by the model with only surprise as a predictor. While the model comparison results only reveal which model performed better compared to the others, Table 8 also shows that participants' judgments were well accounted for by the models. For example, for the 11 participants whose judgments were best accounted for by the model that considers both importance and surprise as predictors, the median correlation between model predictions and judgments was $r = .9$ with a median RMSE $= 9.1$.

We performed the same analysis on participants' wrongfulness judgments as well. As Table 8 shows, 32 participants' judgments were best accounted for by a model that only considers surprise as a predictor, 12 participants by the baseline model, 4 participants by the model with only importance, and 1 participant by the model with both importance and surprise. For wrongfulness judgments, the model fits weren't quite as high with a median correlation of $r = 0.61$, RMSE $= 32.08$ for 32 participants who were best fitted by the surprise only model.

7.5. Discussion

Experiment 4 provided a comprehensive test of our model. As predicted, the results show that judgments of responsibility are influenced by how surprising and important a vote was. Votes in support of policies that are morally bad are perceived as more surprising. The closer a vote was to having been pivotal for the outcome, the more important it was judged. The results of Experiment 2 showed that both pivotality as well as the number of causes mattered. The experiment untangled these two factors by manipulating the number of votes required for a policy to pass. In Experiment 4, we kept this threshold constant. Thus, the number of causes and their pivotality were confounded. However, the way in which participants' importance judgments changed as a function of the number of votes in favor was consistent with the pivotality model (see Fig. 8).

In contrast to what we predicted, how morally bad a policy was perceived didn't affect the extent to which importance and surprise influenced participants' responsibility judgments. Our prediction was based on previous work (Giffin & Lombrozo, 2015; Josephs et al., 2016) and on our results from Experiment 3, in which we found that the causal role of a committee member's vote had a larger impact on participants' responsibility judgments for neutral compared to moral scenarios (see Fig. 7). One reason for the discrepancy between experiments might be how we assessed responsibility. In Experiment 3, the question focused on the *consequences* of the committee member's vote (e.g. "to what extent is Tim responsible for corporal punishment being introduced in schools?"), whereas in Experiment 4 it focused on the passing of the policy itself (e.g., "to what extent is Tim responsible for the policy passing?"). It is possible that when the consequences of a policy were emphasized, participants interpreted the question to be about *moral* responsibility (Experiment 3). In contrast, when the policy's passing was emphasized, participants may have interpreted the question to be about *causal* responsibility (Experiment 4). Responsibility is a broad concept that can be interpreted differently depending on the situation (see Malle, 2011; Samland & Waldmann, 2016).

In Experiment 4 we also assessed what factors influenced participants' judgments for moral wrongfulness. Importance matters more than surprise for judgments of responsibility, but the reverse holds for judgments of wrongfulness (see Table 7). This is consistent with prior work which has found that an agent's causal contribution mattered more for judgments of blame, whereas for judgments of wrongfulness the agent's mental states were more important (Cushman, 2008).

Experiment 4 allowed us to evaluate how well the model captures individual participants' responses. We found that while most participants' responsibility judgments were best captured by a model that only considers importance as a predictor, there was a large group of participants whose responsibility judgments were sensitive to both surprise and importance (see Table 8). For wrongfulness, most participants' judgments were best explained by a model that only considers surprise as a predictor. These analyses show that the model not only captures the postulated relationship between the variables on an aggregate level, but also on the level of individual participants.

8. General Discussion

In this paper, we developed and tested a computational model that predicts responsibility judgments by considering what causal role a person's action played in bringing about the outcome, and what the action reveals about the kind of person they are. We tested the model in voting scenarios in which multiple members of a political committee voted on different policies. This setting allowed us to quantitatively manipulate information relevant to the two components of the model, and systematically investigate how each component affects people's responsibility judgments. We manipulated the causal structure of the situation, the party affiliation of the committee members, how each member voted, whether the policy was passed, and the moral context of the vote.

The party affiliations, voting pattern, and moral context affect the extent to which a particular committee member's vote is *surprising*. For example, an individual committee member's vote is particularly surprising when they voted "yes" even though their party didn't support the policy, and all of the other committee members voted against the policy. The threshold which determines how many votes are required for a policy to pass and the voting pattern affect how *important* an individual committee member's vote was for the outcome. For example, an individual "yes" vote was particularly important when one vote was required and none of the other committee members voted "yes".

In our model, surprise and importance map onto two cognitive processes: surprise is linked to a *dispositional inference* because a surprised observer will update her beliefs about the person – she has learned something about the person that she didn't know before. Importance is linked to *causal attribution* as it expresses an assessment of the structure of the situation and the causal role that a person's action played in bringing about the outcome.

Formalizing precise quantitative predictions about how people assign responsibility has important implications. For example, a currently much debated topic is how to design artificially intelligent agents that behave responsibly in critical situations (Mao & Gratch, 2006; Halpern & Kleiman-Weiner, 2018; Friedenberg & Halpern, 2019; Himmelreich, 2019; Wallach & Allen, 2008). What

should a self-driving car do when it faces the decision between staying on the lane and potentially hitting a child on a bike that suddenly appeared versus switching lanes and potentially colliding with oncoming traffic (Rahwan et al., 2019; Awad et al., 2018; De Freitas et al., 2021)? Our work begins to bridge the gap between extant theoretical frameworks of responsibility judgments that are conceptually rich but don't make quantitative predictions (see Malle et al., 2014), and computational models that generate quantitative predictions but only consider a small subset of the factors that are known to influence responsibility judgments.

The work presented here makes several contributions toward a more comprehensive computational model of responsibility judgments. First, we developed specific computational implementations for the dispositional inference and causal attribution components of our model. While prior computational work had only indirectly tested how these components relate to responsibility judgments (Gerstenberg et al., 2018), we provided a more direct test here. We show that our computational models of dispositional inference and causal attribution accurately predict participants' surprise and importance judgments (Experiment 1) and that these two components, in turn, are critical for capturing participants' responsibility judgments (Experiment 2). The extent to which participants considered the vote of an individual committee member to be surprising was affected by the committee member's party affiliation, and by how the other committee members voted (in Experiments 1 and 2), as well as by the moral content of the policy (in Experiment 4). In addition, the factors that we captured with our causal attribution model predicted importance ratings: votes were judged more important if they were closer to being pivotal, and if fewer causes contributed to the outcome (Experiment 1). The close relationships between surprise, importance, and responsibility held not only on the aggregate level (Experiment 2) but also on the level of individual participants (Experiment 4).

Second, this work connects prior research on how expectations and dispositional inferences affect responsibility judgments to individual decision makers (Gerstenberg et al., 2018) with research on how responsibility is attributed to individuals in groups (Gerstenberg & Lagnado, 2010; Koskuba et al., 2018; Lagnado & Gerstenberg, 2015; Zultan et al., 2012; Lagnado et al., 2013). Our responsibility model accounted well for participants' judgments across a large range of situations.

Third, we show that our model captures responsibility judgments in the moral domain. Recent proposals in the moral psychology literature have emphasized that when evaluating other people's actions, we tend to focus on information that is indicative of a person's character or disposition (Bartels & Pizarro, 2011; Bayles, 1982; Cushman, 2008; Gerstenberg et al., 2010; Schächtele et al., 2011; Pizarro et al., 2003; Uhlmann et al., 2015; Waldmann et al., 2012). Based on this work, we predicted that the causal attribution component of our model would matter more when judging responsibility for policies that were perceived as morally neutral (compared to morally negative policies), and the dispositional inference component would matter more for policies that were perceived as morally negative (compared to morally neutral policies). Experiment 3 provided some evidence that importance matters less in moral contexts. However, how morally bad a policy was didn't affect how much suprise versus importance influenced responsibility judgments in Experiment 4. Both Experiment 3 and 4 showed that when participants judged moral wrongfulness rather than responsibility, what causal role the committee member's action played mattered less. Consistent with prior work (Cushman, 2008), this demonstrates that different moral judgments rely on systematically different inputs.

Our model construes dispositional inferences and causal attributions as independent components that combine additively to yield responsibility judgments (see Eq. 4). However, it is likely that the two components affect one another. For example, as Alicke (2000) has shown, people tend to take a person's moral character into account when considering what causal role their action played. More generally, people's causal judgments have been shown to be affected by whether the event of interest was normal or abnormal (Hitchcock & Knobe, 2009; Kominsky et al., 2015; Icard et al., 2017; Gerstenberg & Icard, 2019; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Hilton & Slugoski, 1986; Kahneman & Miller, 1986; Samland & Waldmann, 2016). The results of Experiment 4 reveal a moderate positive correlation between participants' surprise and importance judgments ($r = .35$). Future work is required to better understand how exactly the different concepts relate to one another.

8.1. *Future directions: dispositional inference*

Dispositional inferences are a key component underlying responsibility judgments. More research is needed on how dispositional inferences affect responsibility judgments. We will discuss three open challenges: the source of prior expectations, how action expectations map onto responsibility judgments, and how different mental states can be incorporated into the model.

8.1.1. *The source of prior expectations*

A central idea in our model is that people compare their expectation about how an agent *should* act with the agent's *actual* behavior. But where do people's initial expectations come from? There are at least two possible sources: expectations about how *this person* will act, or how *a reasonable person* would act (cf. Sytsma et al., 2012; Tobia, 2018). In our paradigm, participants only ever observed a single action from each person that they were asked to evaluate. So our paradigm cannot tease apart these different sources of prior expectations. Future work needs to investigate how expectations resulting from person-specific versus more general standards of comparisons interact with one another to inform judgments of responsibility.

The different possible sources for prior expectations also highlight the complex relationship between dispositional inferences and judgments of responsibility. For example, consider a situation in which we know a person well, we expect them to act in a morally bad manner, and they do. Our model predicts that this person would be held less responsible than a person for whom the morally bad behavior came as a surprise. However, a model that emphasizes the difference between the judged person's disposition, and what the

disposition of a reasonable person should be, would predict no difference here. In both cases, the person acted far worse than a reasonable person would have.[8]

8.1.2. From action expectations to responsibility judgments

In our voting setting, we were able to go directly from action expectations (and whether or not they were violated) to responsibility judgments. This direct mapping from action expectations to responsibility judgments is not generally warranted. While agents are often held more responsible for unexpected actions (Brewer, 1977; Fincham & Jaspars, 1983; Malle, Monroe, & Guglielmo, 2014; Petrocelli, Percy, Sherman, & Tormala, 2011), sometimes it's the case the expected actions elicit higher responsibility judgments (Johnson & Rips, 2015). Gerstenberg et al. (2018) showed that violating expectations in itself doesn't result in more (or less) responsibility, but that dispositional inferences *mediate* the relationship between action expectations and responsibility judgments. Unexpected actions can lead to different dispositional inferences – and thus, differentially affect judgments of responsibility – depending on the context in which they are made. As a goalie in soccer, for example, saving an unexpected shot is diagnostic for skill and good future performance. Thus, the computational model predicts that, in this context, unexpected actions that produce positive outcomes will yield more responsibility. In contrast, in contexts where unexpected actions are indicative of poor decision-making – for example, when a contestant in a game show bets on the color with the lower probability in a two-colored spinner – the model predicts that unexpected actions will be assigned less credit. Future research is needed to figure out how people's understanding of the situation affects the mapping from action expectations to responsibility judgments.

8.1.3. Modeling different mental states

Responsibility is a rich and multifaceted concept (cf. Hart, 2008). A variety of factors influence how people assign responsibility. For example, people take into account whether agents intended the consequences of their actions (Cushman, Knobe, & Sinnott-Armstrong, 2008; Shultz & Wright, 1985), whether the consequences were realized in the intended way (Alicke, Rose, & Bloom, 2012; Guglielmo & Malle, 2010; Pizarro et al., 2003; Schächtele et al., 2011; Gerstenberg et al., 2010), and whether they were able to foresee the consequences of their actions (Lagnado & Channon, 2008; Markman & Tetlock, 2000; Young & Saxe, 2009).

In our computational model, mental states are relevant for the dispositional inference component: an agent's mental state affects an observer's expectations about how the agent will act, and thus the extent to which the observer draws an inference about the agent's character. So far, mental states are only implicitly represented in our model. An important goal for future research is to build on the extensive qualitative work on the role of mental states in responsibility judgments (see, e.g. Malle et al., 2014; Alicke, 2000), and generate testable quantitative predictions. Our voting paradigm could serve as a framework for such experiments. For example, when someone votes in favor of a policy and is fully aware of its morally negative consequences, this licenses a stronger inference about that person's character than when someone voted in favor of a policy without realizing that implementing the policy would have morally negative consequences.

8.2. Future directions: Causal attribution

In this paper, we have extended the simple model of causal contribution used in Gerstenberg et al. (2018) by implementing a graded notion of pivotality, and by considering the number of causes that contributed to the outcome. Our model assumes that observers have no uncertainty about the agent's pivotality and the number of causes that contributed to the outcome. In many situations, however, observers are uncertain about what actually happened, and need to infer each agent's causal contribution. More generally, causal attributions are nuanced and influenced by a variety of factors (e.g., Wolff, 2007; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Lombrozo, 2010; Allen et al., 2015; White, 2014; Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Einhorn & Hogarth, 1986; Kirfel et al., 2020; Morris et al., 2019; Samland & Waldmann, 2016; Kominsky et al., 2015; Vasilyeva et al., 2018).

We will briefly discuss two factors that affect causal attributions: the extent to which the cause is spatiotemporally connected to the outcome, and the function that determines how individual causes contributed to the outcome.

8.2.1. Physical processes

Our computational model is based on counterfactual theories of causation (Lewis, 1973). In philosophy, *process theories of causation* are another framework for thinking about causation. Process theories establish causal relationships by analyzing whether there was a physical connection between candidate cause and effect. Empirical work has shown that participants' causal judgments are sensitive to information about physical connections (Lombrozo, 2010; Dowe, 2000; Walsh & Sloman, 2011; Wolff, 2007), as well as the consideration of counterfactuals (Gerstenberg et al., 2017; Gerstenberg, Goodman, Lagnado, & Tenenbaum, in press; Gerstenberg & Stephan, in press).

Our experimental paradigm didn't manipulate information about physical connections. In our everyday lives, we often experience situations in which the processes by which individuals contribute to an outcome differ, such as in team sports. Based on previous research (Lombrozo, 2010; Dowe, 2000; Walsh & Sloman, 2011; Wolff, 2007; Iliev, Sachdeva, & Medin, 2012), it seems plausible that people would consider this information when judging responsibility. Future research should look at situations where individual causes differ in how physically connected they are to a joint outcome.

[8] We thank an anonymous reviewer for raising this point.

8.2.2. Causal integration functions

Another aspect that has been shown to affect judgments of responsibility is the way in which individual contributions combine to determine a group outcome (Gerstenberg & Lagnado, 2010; Waldmann, 2007; Lagnado et al., 2013; Zultan et al., 2012; Allen et al., 2015; Pearl, 1999; Woodward, 2006; Teigen & Brun, 2011). For example, contributions may combine additively (like in voting), conjunctively (where the weakest link determines the group's performance), or disjunctively (where the group is as good as its strongest member). To better understand how individual contributions in group settings affect responsibility judgments, future research should explore such situations as well as situations in which individuals differ in how much power they have to affect the outcome. For example, the votes of some countries in the United Nations count more than the votes of others, and in presidential elections in the United States, some states cast more votes than others (see Taylor & Zwicker, 1993; Rabinowitz & MacDonald, 1986). Settings like these will help tease apart how much responsibility judgments are affected by how necessary and sufficient an agent's action was for the outcome (Icard et al., 2017).

9. Conclusion

In this paper, we developed and tested a computational model that postulates two key cognitive processes in responsibility judgments: dispositional inferences and causal attributions. We have shown that the model's predictions hold when its two key components are assessed directly, when the model is employed in complex causal settings, and when it is tested in the moral domain. Overall, this work takes us one step closer toward a comprehensive computational model of responsibility judgments.

Acknowledgments

We thank Aaron Beller, Thomas Icard, Max Kleiman-Weiner, and Kevin Smith for feedback and discussion.

TG and JBT were supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216. JYH was supported in part by NSF grants IIS-0911036 and CCF-1214844, AFOSR grant FA9550-08–1-0438, ARO grant W911NF-14–1-0017, and the DoD Multidisciplinary University Research Initiative (MURI) program administered by AFOSR under grant FA9550-12–1-0040.

Part of this research was published in the Proceedings of the Cognitive Science Conference: Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX, 2015 (pp. 788–793). Cognitive Science Society.

Appendix A. Sensitivity analysis of Bayesian surprise model

Fig. A1.

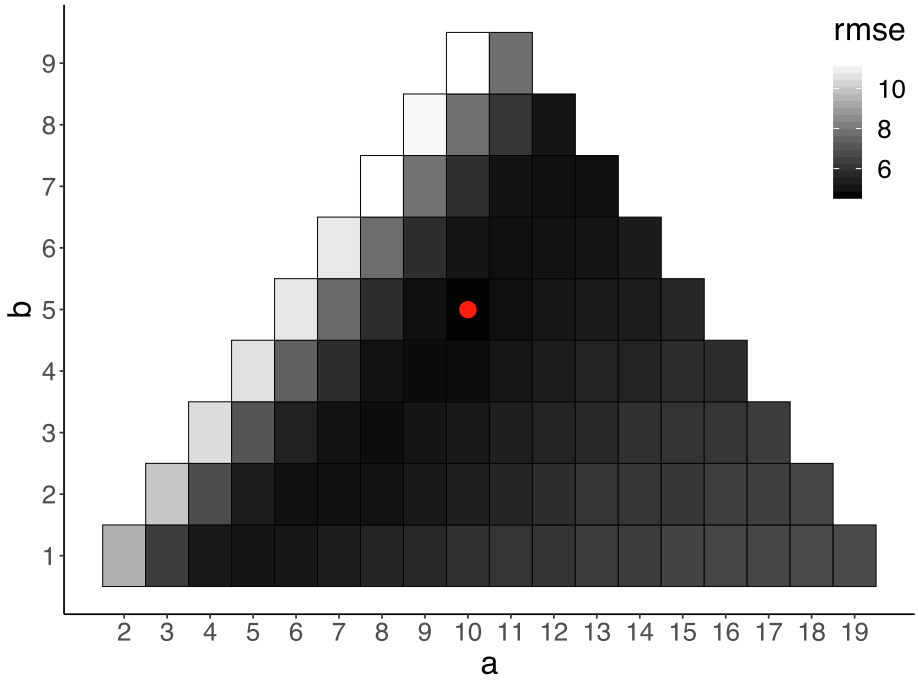

Fig. A1. Sensitivity analysis of the Bayesian surprise model. The tile plot shows the root mean squared error between model predictions and participants' mean surprise judgments for the scenarios presented in Experiment 1 as a function of the two parameters a and b that were fitted to the data (see Fig. 2). The red dot indicates the best fitting set of parameters: $a = 10$ and $b = 5$.

Appendix B. Scenarios presented in Experiment 1

Table B1.

Table B1
List of 27 scenarios presented in Experiment 1.

		party					vote						
scenario	person	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
1	1	1	1	1	0	0	1	1	1	1	1	5	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	0	0	0	0	0	1	1	0	0	0	1	1
4	1	0	0	0	0	0	1	0	0	0	0	1	1
5	1	1	1	1	1	1	1	1	1	1	1	5	1
6	2	1	1	0	0	0	1	0	1	0	0	5	0
7	2	0	0	0	0	0	1	0	0	0	0	5	0
8	3	1	1	1	0	0	1	1	0	0	0	4	0
9	1	1	1	1	0	0	1	1	1	1	1	1	1
10	3	1	1	1	0	0	1	1	0	0	0	3	0
11	2	1	0	0	0	0	0	0	0	0	0	1	0
12	1	1	1	0	0	0	1	1	0	0	0	2	1
13	1	1	1	1	1	0	1	1	1	0	1	4	1
14	4	1	0	0	0	0	1	1	1	0	0	5	0
15	1	0	0	0	0	0	0	0	0	0	0	2	0
16	2	1	0	0	0	0	1	1	0	0	0	2	1
17	4	1	1	1	0	0	1	1	1	1	0	4	1
18	1	1	1	0	0	0	0	0	0	0	0	5	0
19	1	0	0	0	0	0	0	0	0	0	0	3	0
20	1	1	1	0	0	0	1	1	1	1	0	3	1
21	2	0	0	0	0	0	1	0	0	0	0	4	0
22	2	0	0	0	0	0	1	0	0	0	0	2	0
23	2	1	0	0	0	0	1	0	0	0	0	3	0
24	3	1	1	1	1	0	1	1	0	0	0	4	0
25	5	1	0	0	0	0	0	1	1	1	1	1	1
26	5	1	1	0	0	0	0	0	1	1	1	2	1
27	5	0	0	0	0	0	1	1	1	1	1	2	1

Note: person: indicates which person's action in the committee participants were asked to judge; party: 1 = affiliated with the party that supports the policy, 0 = opposite party; vote: 1 = yes, 0 = no; threshold: number of votes required in favor in order for the policy to pass; outcome: 1 = policy passed, 0 = policy didn't pass.

Appendix C. Surprise model predictions

Table C1.

Table C1
Predictions of the Bayesian surprise model. same = committee members who are affiliated with the party that supports the policy, other = committee members from the other party, n = number of people in the committee, yes = number of people who voted yes, party = mean of the party posterior, vote = mean of the vote posterior, surprise = predicted surprise associated with a 'yes' vote, policy = mean of the policy posterior. For example, in Scenario 18, the committee had two members affiliated with the party that supported the policy and two members from the other party (in addition to the fifth committee member whose action we are evaluating). One of the committee members affiliated with the party that supported the policy voted yes, and one of the committee members from the other party voted yes. The posterior indicates that the remaining committee member will vote 'yes' with probability.58 if he is affiliated with the party that supported the policy, and probability.42 if he is from the other party. The surprise column indicates how surprised an observer would be with a 'yes' vote. The surprise predictions were fitted to the data with a Bayesian linear mixed effects model yielding the following parameters for the fixed *intercept* = −76.41 and *slope* = 244.60. (*Note*: Model predictions are based on the model in which $a = 10$, and $b = 5$; see Fig. 2 for details of the model.)

			same					other			
scenario	n	yes	party	vote	surprise	n	yes	party	vote	surprise	policy
1	0	0	0.67	0.57	28.78	2	0	0.31	0.39	72.88	0.47
2	0	0	0.66	0.59	24.67	2	1	0.34	0.42	64.43	0.51
4	0	0	0.66	0.55	33.36	4	0	0.29	0.36	79.55	0.44
5	0	0	0.67	0.58	27.31	4	1	0.32	0.40	69.92	0.49
6	0	0	0.66	0.59	24.70	4	2	0.34	0.43	63.64	0.51
9	1	0	0.65	0.56	32.10	1	0	0.33	0.39	71.64	0.46
11	1	1	0.68	0.59	24.08	1	0	0.32	0.41	67.87	0.50
12	1	1	0.68	0.60	20.63	1	1	0.35	0.44	59.95	0.53

(*continued on next page*)

Table C1 (*continued*)

			same					other			
scenario	n	yes	party	vote	surprise	n	yes	party	vote	surprise	policy
13	1	0	0.65	0.55	34.28	3	0	0.31	0.38	76.02	0.45
14	1	0	0.65	0.56	30.61	3	1	0.32	0.40	70.94	0.47
17	1	1	0.68	0.57	27.59	3	0	0.30	0.39	73.92	0.47
18	1	1	0.68	0.59	23.48	3	1	0.32	0.41	67.11	0.50
19	1	1	0.67	0.60	20.29	3	2	0.35	0.44	59.67	0.53
22	2	1	0.66	0.58	26.37	0	0	0.33	0.42	66.61	0.50
23	2	2	0.69	0.61	19.77	0	0	0.34	0.43	62.40	0.52
24	2	0	0.64	0.54	36.10	2	0	0.31	0.38	76.44	0.44
27	2	1	0.66	0.57	29.43	2	0	0.30	0.39	73.36	0.47
28	2	1	0.66	0.58	26.28	2	1	0.34	0.42	65.67	0.50
30	2	2	0.69	0.60	21.99	2	0	0.31	0.41	68.40	0.50
31	2	2	0.69	0.61	18.15	2	1	0.34	0.43	61.86	0.53
32	2	2	0.69	0.62	15.61	2	2	0.36	0.46	55.81	0.56
35	3	1	0.65	0.55	32.99	1	0	0.33	0.39	71.78	0.46
37	3	2	0.68	0.59	24.40	1	0	0.33	0.41	67.02	0.50
38	3	2	0.67	0.60	20.93	1	1	0.36	0.44	59.77	0.53
39	3	3	0.70	0.61	18.41	1	0	0.32	0.42	65.13	0.53
40	3	3	0.70	0.63	14.58	1	1	0.35	0.46	56.87	0.56
43	4	2	0.65	0.57	28.09	0	0	0.34	0.41	67.05	0.49
44	4	3	0.69	0.60	20.75	0	0	0.34	0.43	63.15	0.52
45	4	4	0.71	0.63	14.68	0	0	0.34	0.44	60.01	0.55

Appendix D. Scenarios presented in Experiment 2

Tables D1 and D2.

Table D1
List of 30 scenarios with three committee members in Experiment 2. Note that if there was a member from each party that supported the outcome, participants were asked to assign responsibility to each member on separate sliders.

	party			vote				
scenario	p1	p2	p3	v1	v2	v3	threshold	outcome
1	0	0	0	0	0	0	1	0
2	0	0	0	1	0	0	1	1
3	1	0	0	0	0	0	1	0
4	1	0	0	1	0	0	1	1
5	1	0	0	1	1	0	1	1
6	1	1	0	1	0	0	1	1
7	1	1	0	1	1	0	1	1
8	1	1	0	1	1	1	1	1
9	1	1	1	1	1	0	1	1
10	1	1	1	1	1	1	1	1
11	0	0	0	0	0	0	2	0
12	0	0	0	1	0	0	2	0
13	1	0	0	0	0	0	2	0
14	1	0	0	1	0	0	2	0
15	1	0	0	1	1	0	2	1
16	1	1	0	1	0	0	2	0
17	1	1	0	1	1	0	2	1
18	1	1	0	1	1	1	2	1
19	1	1	1	1	1	0	2	1
20	1	1	1	1	1	1	2	1
21	0	0	0	0	0	0	3	0
22	0	0	0	1	0	0	3	0
23	1	0	0	0	0	0	3	0
24	1	0	0	1	0	0	3	0
25	1	0	0	1	1	0	3	0
26	1	1	0	1	0	0	3	0
27	1	1	0	1	1	0	3	0
28	1	1	0	1	1	1	3	1
29	1	1	1	1	1	0	3	0
30	1	1	1	1	1	1	3	1

Note: party: 1 = affiliated with the party that supports the policy, 0 = opposite party; vote: 1 = yes, 0 = no; threshold: number of votes required in favor in order for the policy to pass; outcome: 1 = policy passed, 0 = policy didn't pass.

Table D2
List of 140 scenarios with five committee members in Experiment 2. Note that if there was a member from each party that supported the outcome, participants were asked to assign responsibility to each member on separate sliders.

	party					vote						
scenario	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
31	0	0	0	0	0	0	0	0	0	0	1	0
32	0	0	0	0	0	1	0	0	0	0	1	1
33	0	0	0	0	0	1	1	0	0	0	1	1
34	1	0	0	0	0	0	0	0	0	0	1	0
35	1	0	0	0	0	1	0	0	0	0	1	1
36	1	0	0	0	0	0	1	0	0	0	1	1
37	1	0	0	0	0	1	1	0	0	0	1	1
38	1	0	0	0	0	1	1	1	0	0	1	1
39	1	1	0	0	0	0	0	0	0	0	1	0
40	1	1	0	0	0	1	0	0	0	0	1	1
41	1	1	0	0	0	1	1	0	0	0	1	1
42	1	1	0	0	0	1	0	1	0	0	1	1
43	1	1	0	0	0	1	1	1	0	0	1	1
44	1	1	0	0	0	1	1	1	1	0	1	1
45	1	1	1	0	0	1	0	0	0	0	1	1
46	1	1	1	0	0	1	1	0	0	0	1	1
47	1	1	1	0	0	1	1	1	0	0	1	1
48	1	1	1	0	0	1	1	0	1	0	1	1
49	1	1	1	0	0	1	1	1	1	0	1	1
50	1	1	1	0	0	1	1	1	1	1	1	1
51	1	1	1	1	0	1	1	0	0	0	1	1
52	1	1	1	1	0	1	1	1	0	0	1	1
53	1	1	1	1	0	1	1	1	1	0	1	1
54	1	1	1	1	0	1	1	1	0	1	1	1
55	1	1	1	1	0	1	1	1	1	1	1	1
56	1	1	1	1	1	1	1	1	0	0	1	1
57	1	1	1	1	1	1	1	1	1	0	1	1
58	1	1	1	1	1	1	1	1	1	1	1	1
59	0	0	0	0	0	0	0	0	0	0	2	0
60	0	0	0	0	0	1	0	0	0	0	2	0
61	0	0	0	0	0	1	1	0	0	0	2	1
62	1	0	0	0	0	0	0	0	0	0	2	0
63	1	0	0	0	0	1	0	0	0	0	2	0
64	1	0	0	0	0	0	1	0	0	0	2	0
65	1	0	0	0	0	1	1	0	0	0	2	1
66	1	0	0	0	0	1	1	1	0	0	2	1
67	1	1	0	0	0	0	0	0	0	0	2	0
68	1	1	0	0	0	1	0	0	0	0	2	0
69	1	1	0	0	0	1	1	0	0	0	2	1
70	1	1	0	0	0	1	0	1	0	0	2	1
71	1	1	0	0	0	1	1	1	0	0	2	1
72	1	1	0	0	0	1	1	1	1	0	2	1
73	1	1	1	0	0	1	0	0	0	0	2	0
74	1	1	1	0	0	1	1	0	0	0	2	1
75	1	1	1	0	0	1	1	1	0	0	2	1
76	1	1	1	0	0	1	1	0	1	0	2	1
77	1	1	1	0	0	1	1	1	1	0	2	1
78	1	1	1	0	0	1	1	1	1	1	2	1
79	1	1	1	1	0	1	1	0	0	0	2	1
80	1	1	1	1	0	1	1	1	0	0	2	1
81	1	1	1	1	0	1	1	1	1	0	2	1
82	1	1	1	1	0	1	1	1	0	1	2	1
83	1	1	1	1	0	1	1	1	1	1	2	1
84	1	1	1	1	1	1	1	1	0	0	2	1
85	1	1	1	1	1	1	1	1	1	0	2	1
86	1	1	1	1	1	1	1	1	1	1	2	1
87	0	0	0	0	0	0	0	0	0	0	3	0
88	0	0	0	0	0	1	0	0	0	0	3	0
89	0	0	0	0	0	1	1	0	0	0	3	0
90	1	0	0	0	0	0	0	0	0	0	3	0
91	1	0	0	0	0	1	0	0	0	0	3	0
92	1	0	0	0	0	0	1	0	0	0	3	0
93	1	0	0	0	0	1	1	0	0	0	3	0
94	1	0	0	0	0	1	1	1	0	0	3	1
95	1	1	0	0	0	0	0	0	0	0	3	0
96	1	1	0	0	0	1	0	0	0	0	3	0

(*continued on next page*)

Table D2 (*continued*)

	party					vote						
scenario	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
97	1	1	0	0	0	1	1	0	0	0	3	0
98	1	1	0	0	0	1	0	1	0	0	3	0
99	1	1	0	0	0	1	1	1	0	0	3	1
100	1	1	0	0	0	1	1	1	1	0	3	1
101	1	1	1	0	0	1	0	0	0	0	3	0
102	1	1	1	0	0	1	1	0	0	0	3	0
103	1	1	1	0	0	1	1	1	0	0	3	1
104	1	1	1	0	0	1	1	0	1	0	3	1
105	1	1	1	0	0	1	1	1	1	0	3	1
106	1	1	1	0	0	1	1	1	1	1	3	1
107	1	1	1	1	0	1	1	0	0	0	3	0
108	1	1	1	1	0	1	1	1	0	0	3	1
109	1	1	1	1	0	1	1	1	1	0	3	1
110	1	1	1	1	0	1	1	1	0	1	3	1
111	1	1	1	1	0	1	1	1	1	1	3	1
112	1	1	1	1	1	1	1	1	0	0	3	1
113	1	1	1	1	1	1	1	1	1	0	3	1
114	1	1	1	1	1	1	1	1	1	1	3	1
115	0	0	0	0	0	0	0	0	0	0	4	0
116	0	0	0	0	0	1	0	0	0	0	4	0
117	0	0	0	0	0	1	1	0	0	0	4	0
118	1	0	0	0	0	0	0	0	0	0	4	0
119	1	0	0	0	0	1	0	0	0	0	4	0
120	1	0	0	0	0	0	1	0	0	0	4	0
121	1	0	0	0	0	1	1	0	0	0	4	0
122	1	0	0	0	0	1	1	1	0	0	4	0
123	1	1	0	0	0	0	0	0	0	0	4	0
124	1	1	0	0	0	1	0	0	0	0	4	0
125	1	1	0	0	0	1	1	0	0	0	4	0
126	1	1	0	0	0	1	0	1	0	0	4	0
127	1	1	0	0	0	1	1	1	0	0	4	0
128	1	1	0	0	0	1	1	1	1	0	4	1
129	1	1	1	0	0	1	0	0	0	0	4	0
130	1	1	1	0	0	1	1	0	0	0	4	0
131	1	1	1	0	0	1	1	1	0	0	4	0
132	1	1	1	0	0	1	1	0	1	0	4	0
133	1	1	1	0	0	1	1	1	1	0	4	1
134	1	1	1	0	0	1	1	1	1	1	4	1
135	1	1	1	1	0	1	1	0	0	0	4	0
136	1	1	1	1	0	1	1	1	0	0	4	0
137	1	1	1	1	0	1	1	1	1	0	4	1
138	1	1	1	1	0	1	1	1	0	1	4	1
139	1	1	1	1	0	1	1	1	1	1	4	1
140	1	1	1	1	1	1	1	1	0	0	4	0
141	1	1	1	1	1	1	1	1	1	0	4	1
142	1	1	1	1	1	1	1	1	1	1	4	1
143	0	0	0	0	0	0	0	0	0	0	5	0
144	0	0	0	0	0	1	0	0	0	0	5	0
145	0	0	0	0	0	1	1	0	0	0	5	0
146	1	0	0	0	0	0	0	0	0	0	5	0
147	1	0	0	0	0	1	0	0	0	0	5	0
148	1	0	0	0	0	0	1	0	0	0	5	0
149	1	0	0	0	0	1	1	0	0	0	5	0
150	1	0	0	0	0	1	1	1	0	0	5	0
151	1	1	0	0	0	0	0	0	0	0	5	0
152	1	1	0	0	0	1	0	0	0	0	5	0
153	1	1	0	0	0	1	1	0	0	0	5	0
154	1	1	0	0	0	1	0	1	0	0	5	0
155	1	1	0	0	0	1	1	1	0	0	5	0
156	1	1	0	0	0	1	1	1	1	0	5	0
157	1	1	1	0	0	1	0	0	0	0	5	0
158	1	1	1	0	0	1	1	0	0	0	5	0
159	1	1	1	0	0	1	1	1	0	0	5	0
160	1	1	1	0	0	1	1	0	1	0	5	0
161	1	1	1	0	0	1	1	1	1	0	5	0
162	1	1	1	0	0	1	1	1	1	1	5	1
163	1	1	1	1	0	1	1	0	0	0	5	0
164	1	1	1	1	0	1	1	1	0	0	5	0

(*continued on next page*)

Table D2 (*continued*)

	party					vote						
scenario	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
165	1	1	1	1	0	1	1	1	1	0	5	0
166	1	1	1	1	0	1	1	1	0	1	5	0
167	1	1	1	1	0	1	1	1	1	1	5	1
168	1	1	1	1	1	1	1	1	0	0	5	0
169	1	1	1	1	1	1	1	1	1	0	5	0
170	1	1	1	1	1	1	1	1	1	1	5	1

Note: party: 1 = affiliated with the party that supports the policy, 0 = opposite party; vote: 1 = yes, 0 = no; threshold: number of votes required in favor in order for the policy to pass; outcome: 1 = policy passed, 0 = policy didn't pass.

Appendix E. Scenarios presented in Experiment 4

Tables E1 and E2.

Table E1
Scenarios in Experiment 4.

scenario	policy
1	introduce corporal punishment in schools
2	introduce the death penalty for minor crimes (e.g. theft)
3	introduce prison sentences of at least 5 years for doctors who practice abortion
4	increase sales taxes on meat products (e.g. sausages or burgers) by 3 percent
5	exempt all politicians from paying taxes
6	introduce tax breaks for dog owners
7	introducing a law that makes it illegal for people with disabilities to get a job
8	introduce a fine of at least $200 for littering in public places
9	ban children who are not vaccinated against Varicella from going to school
10	change the color of all government buildings to red
11	declare that sexual abuse within marriage does not count as a criminal liability
12	legally require every citizen to get vaccinated against the flu
13	change the legal voting age to 35 years
14	allow all citizens over the age of 21 to purchase a gun without a license
15	criminalize same-sex relationships
16	make it easier for landlords to increase rent
17	decriminalizing corruption for politicians
18	change the standard size of public trashcans from 33 gallons to 60 gallons
19	introduce prison sentences for extramarital affairs
20	changing the font of government docs to Arial
21	legalize gambling in casinos
22	make it mandatory to keep dogs on leash at all times in public
23	lie to citizens of the country about an oil leak that is poisening the country's drinking water
24	requiring every citizen older than 12 to have their own social media profile for official identification

Table E2
Experiment 4. Mean ratings for each scenario together with bootstrapped 95% confidence intervals.

scenario	votes	badness	importance	surprise	responsibility	wrongfulness
1	2	83.64 [76.34, 90.16]	90.14 [86, 93.94]	65.2 [58.22, 72.18]	87.32 [81.7, 92.16]	83.02 [76.26, 89.16]
2	3	96.4 [93.48, 98.68]	61.56 [55.86, 67.06]	73.92 [66.32, 81.58]	69.46 [63.34, 75.4]	95.34 [92.2, 98]
3	3	73.58 [63.98, 83.2]	59.72 [52.16, 66.8]	54.46 [46.02, 62.7]	63.56 [56.56, 70.06]	72.68 [62.2, 83]
4	5	40.1 [32.02, 48.68]	36 [28.66, 44.06]	37.36 [29.96, 44.84]	42.48 [35.08, 50.02]	35.94 [27.08, 45.76]
5	5	88.64 [81.16, 94.9]	37.32 [29.48, 45.36]	31.94 [21.16, 42.26]	51.8 [43.02, 60.24]	86.4 [77.88, 93.7]
6	3	37.18 [28.68, 45.92]	41.26 [34.6, 47.88]	35.88 [29.82, 41.86]	57.64 [52.3, 62.84]	28.92 [21.4, 36.54]
7	5	95.86 [91.16, 99.08]	45.22 [36.02, 54.76]	72.2 [63.74, 80.92]	55.78 [47.14, 63.94]	94.68 [90.04, 98.24]
8	4	16.1 [10.5, 22.94]	43.9 [36.62, 51.48]	18.9 [13.66, 24.46]	48.94 [42.58, 56.04]	13.24 [8.62, 18.52]
9	3	45.6 [35.74, 55.76]	54.62 [48.68, 60.48]	37.6 [30.76, 44.34]	57.78 [50.5, 64.38]	41.56 [32.68, 50.03]
10	2	63.58 [52.92, 74.72]	85.58 [79.16, 90.98]	55.66 [47.7, 63.6]	85.52 [80.74, 90.12]	20.44 [13.68, 28.42]
11	4	95.64 [92.26, 98.5]	50.08 [41.88, 57.98]	66.64 [55.99, 75.52]	60.06 [51.84, 67.22]	92.02 [86.72, 96.1]
12	4	49.2 [38.98, 59.24]	46.66 [39.24, 55.26]	36.44 [29.22, 44.62]	52.82 [45.08, 60.64]	48.58 [38.2, 59.34]
13	5	75.56 [66.68, 83.62]	37.68 [29.28, 46.26]	53.66 [44.14, 63.02]	47.62 [39.1, 56.02]	79.3 [71.7, 86.56]
14	3	73.96 [65.59, 81.98]	50.64 [42.84, 57.95]	55.7 [47.23, 63.6]	60.54 [53.56, 67.62]	72.58 [64, 80.44]

(*continued on next page*)

Table E2 (*continued*)

scenario	votes	badness	importance	surprise	responsibility	wrongfulness
15	5	82.12 [72.48, 90.96]	39.82 [31.58, 48.82]	61.52 [51.66, 71.18]	53.62 [45.22, 61.82]	81.3 [72.12, 90.26]
16	2	69.8 [61.92, 77.28]	90.48 [87.04, 93.7]	52.16 [45.18, 60.13]	89.14 [84.9, 93.18]	71.18 [63.12, 78.8]
17	4	89.04 [82.08, 94.9]	49.06 [41.62, 57.32]	42.7 [32.76, 52.9]	55.68 [47.78, 63.38]	84.22 [75.98, 91.32]
18	5	10.76 [6.2, 15.94]	28.28 [20.92, 36.42]	18.38 [11.96, 25.9]	40.04 [31.74, 48.86]	11.04 [5.58, 17.18]
19	2	75.86 [67.18, 83.4]	87.54 [82.24, 92.02]	62.86 [54.5, 71.1]	86.72 [81.42, 91.04]	71.42 [62.12, 79.76]
20	2	13.28 [7, 19.64]	90 [86.1, 93.8]	43.3 [34.48, 51.16]	88.3 [83.86, 92.26]	11.3 [6.04, 17.22]
21	2	25.96 [18.48, 33.64]	86.12 [81.18, 90.48]	33.94 [26.48, 41.18]	86.76 [82.58, 90.92]	24.06 [17.16, 30.94]
22	3	18.14 [12.1, 25.38]	52.16 [45.18, 59.49]	25.96 [20.14, 31.66]	55.74 [49.58, 61.98]	14.76 [9.48, 20.32]
23	4	97.56 [95.24, 99.3]	50.82 [42.88, 59.22]	67.66 [58.52, 76.74]	59.66 [52.04, 67.43]	96.18 [92.58, 98.76]
24	2	81.96 [73.78, 89.46]	88.16 [83.56, 92.14]	64.28 [55.76, 73.18]	87.58 [82.69, 92]	81.34 [72.72, 88.72]

References

Ajzen, I. (1971). Attribution of dispositions to an actor: Effects of perceived decision freedom and behavioral utilities. *Journal of Personality and Social Psychology, 18*(2), 144–156.

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin, 82*(2), 261–277.

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

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

Alicke, M. D., Rose, D., & Bloom, D. (2012). Causation, norm violation, and culpable control. *The Journal of Philosophy, 108*(12), 670–696.

Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 84–89). Austin, TX: Cognitive Science Society.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 0064.

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

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*(2), 74–85.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161.

Bayles, M. D. (1982). Character, purpose, and criminal responsibility. *Law and Philosophy, 1*(1), 5–20.

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

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

Byrne, R. M., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1318.

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

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

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

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*(1), 281–289.

Darley, J. M. (2009). Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science, 5*, 1–23.

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

De Freitas, J., Censi, A., Walker Smith, B., Di Lillo, L., Anthony, S. E., & Frazzoli, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. *Proceedings of the National Academy of Sciences, 118*(11). e2010202118.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*(9), 634–636.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dowe, P. (2000). *Physical causation*. Cambridge, England: Cambridge University Press.

Duff, R. A. (1993). Choice, character, and criminal liability. *Law and Philosophy, 12*(4), 345–383.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99*(1), 3–19.

Engl, F. (2018). A theory of causal responsibility attribution. Available at SSRN 2932769.

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

Feinberg, J. (1968). Collective responsibility. *The Journal of Philosophy, 65*(21), 674–688.

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

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

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a bayesian perspective. *Psychological Review, 90*(3), 239–260.

Fishbein, M., & Ajzen, I. (1973). Attribution of responsibility: A theoretical note. *Journal of Experimental Social Psychology, 9*(2), 148–153.

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

Gerstenberg, T., Goodman, N.D., Lagnado, D.A., & Tenenbaum, J.B. (in press). A counterfactual simulation model of causal judgments for physical events. Psychological Review.

Gerstenberg, T., & Icard, T. F. (2019). 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., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731–1744.

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

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

Giffin, C., & Lombrozo, T. (2015). Mental states are more important in evaluating moral than conventional violations. *Cogsci*.

Golding, N. (2018). greta: Simple and scalable statistical modelling in r. (R package version 0.3.0.9001).

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

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology, 11*, 1069.

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition, 117*(2), 139–150.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods, 48*(3), 829–842.
Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review, 108*(4), 814–834.
Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science, 66*, 413–457.
Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the thirty-second aaai conference on artificial intelligence (aaai-18)* (pp. 1853–1860).
Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887.
Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology, 41*(4), 316–328.
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.
Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology, 21*(1), 107–112.
Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc.
Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition, 190*, 157–164.
Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A logical model of the intuitive analysis of variance. *Journal of Personality and Social Psychology, 53*(4), 663.
Hilton, D. J., McClure, J., & Slugoski, B. (2005). Counterfactuals, conditionals and causality: A social psychological perspective. In D. R. Mandel, D. J. Hilton, & P. Catellani (Eds.), *The psychology of counterfactual thinking* (pp. 44–60). London: Routledge.
Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review, 93*(1), 75–88.
Himmelreich, J. (2019). Responsibility for Killer Robots. *Ethical Theory and Moral Practice*.
Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy, 11*, 587–612.
Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press (OUP).
Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition, 161*, 80–93.
Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition, 40*(8), 1387–1401.
Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(10), 785.
Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology, 77*, 42–76.
Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1–24.
Josephs, M., Kushnir, T., Gräfenhain, M., & Rakoczy, H. (2016). Children protest moral and conventional violations more when they believe actions are freely chosen. *Journal of Experimental Child Psychology, 141*, 247–255.
Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
Kirfel, L., Icard, T. F., & Gerstenberg, T. (2020). Inference from explanation. *PsyArXiv*.
Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J.B. (2015). Inference of intention and permissibility in moral decision making. In D.C. Noelle, R. Dale, A.S. Warlaumont, J. Yoshimi, J. Matlock, T., C.D., & P.P. Maglio (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 1123–1128). Austin, TX: Cognitive Science Society.
Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: intuition and diversity* (vol. 2). The MIT Press.
Kominsky, J. F., Phillips, J., 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., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation, 4*(1), 46–63.
Lagnado, D. A., & Gerstenberg, T. (2015). A difference-making framework for intuitive judgments of responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (vol. 3, pp. 213–241). Oxford University Press.
Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.
Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 47*, 1036–1073.
Lagnado, D. A., & Harvey, N. (2008). The impact of discredited evidence. *Psychonomic Bulletin & Review, 15*(6), 1166–1173.
Latané, B. (1981). The psychology of social impact. *American Psychologist, 36*(4), 343–356.
Lewis, D. (1973). Causation. *The Journal of Philosophy, 70*(17), 556–567.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*(4), 303–332.
Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In *Theories in social psychology* (pp. 72–95). Malden, MA: Wiley-Blackwell.
Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*, 293–318.
Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.
Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology, 79*(3), 309–326.
Malle, B. F., Monroe, A. E., & Guglielmo, S. (2014). Paths to blame and paths to convergence. *Psychological Inquiry, 25*(2), 251–260.
Mao, W., & Gratch, J. (2006). Evaluating a computational model of social causality and responsibility. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems* (pp. 985–992).
Mao, W., & Gratch, J. (2012). Modeling social causality and responsibility judgment in multi-agent interactions. *Journal of Artificial Intelligence Research, 44*, 223–273.
Markman, K. D., & Tetlock, P. E. (2000). 'i couldn't have known': Accountability, foreseeability and counterfactual denials of responsibility. *British Journal of Social Psychology, 39*(3), 313–325.
Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23.
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.
McIntyre, A. (2019). Doctrine of double effect. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2019 ed.). Metaphysics Research Lab, Stanford University.
Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General, 146*(1), 123.
Moore, M. S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford University Press.
Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLoS ONE, 14*(8). e0219704.
Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*(2), 331–355.
Naumov, P., & Tao, J. (2018). Blameworthiness in games with imperfect information. CoRR.
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.
Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology, 100*(1), 30–46.

Philpot, R., Liebst, L. S., Levine, M., Bernasco, W., & Lindegaard, M. R. (2019). Would i be helped? cross-national cctv footage shows that intervention is the norm in public conflicts. *American Psychologist*.
Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science, 14*(3), 267–272.
Quillien, T. (2020). When do we think that x caused y? *Cognition, 205*, 104410.
R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria.
Rabinowitz, G., & MacDonald, S. E. (1986). The power of the states in us presidential elections. *The American Political Science Review*, 65–87.
Rachels, J. (2007). Active and passive euthanasia. In M. A. Sudbury (Ed.), *Bioethics: An Introduction to the History, Methods, and Practice* (pp. 64–69). Jones and Bartlett.
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477–486.
Reeder, G. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological inquiry, 20*(1), 1–18.
Ritov, I., & Baron, J. (1992). Status-quo and omission biases. *Journal of Risk and Uncertainty, 5*(1).
Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of personality and social psychology, 35*(7), 485.
Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition, 156*, 164–176.
Sartorio, C. (2007). Causation and responsibility. Philosophy. *Compass, 2*(5), 749–765.
Schächtele, S., Gerstenberg, T., & Lagnado, D. A. (2011). Beyond outcomes: The influence of intentions and deception. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1860–1865). Austin, TX: Cognitive Science Society.
Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review, 101*(4), 632–652.
Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer-Verlag.
Shoemaker, D. (Ed.). (2015). Oxford studies in agency and responsibility (vol. 3). New York: Oxford University Press.
Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science, 17*(2), 97–108.
Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*(4), 323–348.
Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.
Summers, A. (2018). Common-sense causation in the law. *Oxford Journal of Legal Studies, 38*(4), 793–821.
Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(4), 814–820.
Taylor, A., & Zwicker, W. (1993). Weighted voting, multicameral representation, and power. *Games and Economic Behavior, 5*(1), 170–181.
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.
Tobia, K. P. (2018). How people judge what is reasonable. *Ala. L. Rev., 70*, 293.
Trope, Y. (1974). Inferential processes in the forced compliance situation: A bayesian analysis. *Journal of Experimental Social Psychology, 10*(1), 1–16.
Trope, Y., & Burnstein, E. (1975). Processing the information contained in another's behavior. *Journal of Experimental Social Psychology, 11*(5), 439–458.
Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.
Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition, 116*(1), 87–100.
Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science, 42*(4), 1265–1296.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432.
Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31*(2), 233–256.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In *The oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.
Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language, 26*(1), 21–52.
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: The Guilford Press.
Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology, 15*(1), 1–20.
White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science, 38*(1), 38–75.
Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*(1), 82–111.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.
Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review, 115*(1), 1–50.
Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*(2), 283–301.
Yablo, S. (2002). De facto dependence. *The Journal of Philosophy, 99*(3), 130–148.
Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*(10), 2065–2072.
Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition, 125*(3), 429–440.