Predicting responsibility judgments from dispositional inferences and causal attributions

Antonia F. Langenhoff
University of California, Berkeley

Alex Wiegmann
University of Göttingen

Joseph Y. Halpern
Cornell University

Joshua B. Tenenbaum
Massachusetts Institute of Technology

Tobias Gerstenberg*
Stanford University

Abstract

How do people hold others responsible for their actions? In this paper, we test and extend a computational framework originally introduced by Gerstenberg et al. (2018) that assigns responsibility as a function of 1) a dispositional inference that captures what we learn about a person's character from their action and 2) the causal role that the person's action played in bringing about the outcome. Previously, this framework has been shown to accurately capture responsibility judgments of decision-makers in achievement contexts. Here, we focus on responsibility judgments in voting scenarios, in which political committee members vote on whether or not a policy should be passed. This allowed us to manipulate dispositional inferences and causal attributions in graded ways and show that the predictions of the model hold in group settings that are causally more complex. We provide further support in favor of the framework by showing that its key components can be tested directly by asking how 1) surprising and 2) important a committee member's vote was. Participants' answers to these questions accurately predict the responsibility judgments of another group of participants. Finally, we show that the computational framework also captures participants' judgments in the moral domain, where which component is most relevant shifts from causal attributions to dispositional inferences.

Keywords: responsibility, causality, counterfactuals, pivotality, normality, voting, expectations

*Corresponding author: Tobias Gerstenberg (gerstenberg@stanford.edu), 302 Jordan Hall, Stanford, CA, 94305.

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 – fairly clearly. Imagine for a second, however, that Hilary Clinton had become the next president of the United States. Certainly, Suprun's party colleagues would have held Suprun responsible for contributing to Clinton's victory and Trump's loss in that case. But to what extent? Intuitively, do you think they would have blamed him more than, for example, 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. Do you think Republicans would have blamed Suprun even more in this scenario?

Judgments of responsibility are ubiquitous in our everyday lives. When something goes wrong – for example, when our favored candidate lost an election – we want to know who is to blame. Conversely, when something goes right, we are motivated to find those who deserve credit. How exactly people assign responsibility has puzzled researchers in psychology (Alicke, 2000; Hilton, McClure, & Slugoski, 2005; Lagnado & Harvey, 2008; Shaver, 1985), philosophy (Hart & Honoré, 1959/1985) and the legal sciences (Moore, 2009) for decades. In this paper, we further develop and test a computational framework for responsibility judgments that was originally introduced in Gerstenberg et al. (2018). This framework builds on previous work, but makes quantitative predictions about people's responsibility judgments in a broad range of situations.

The framework predicts that responsibility judgments are influenced by two key processes. Inspired by a rich literature in attribution theory (Ajzen, 1971; Fishbein & Ajzen, 1973; Heider, 1946; Weiner & Kukla, 1970), the first process is a dispositional inference that captures what we learn about a person's character from observing their action. The idea is that, in a given situation, we form an expectation about how another person will act, based on our knowledge about that person and the situation. The more the person's actual behavior diverts from our expectation, the more likely we are to infer that the person's action must have been determined by an unobserved aspect of her disposition. This dispositional inference, in turn, translates into a responsibility judgment: We hold another person responsible to the extent that we see her action as determined by her own dispositions, goals, or desires, rather than by determinants that are out of her control (e.g. Alicke, 2000). 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 framework predicts that Republicans would assign quite a high level of blame to him for contributing to Trump's (hypothetical) loss. Critically, the framework 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 less surprising.

The second process is a causal attribution that determines what role the person's action played in bringing about the outcome. Specifically, the computational framework predicts that a person is held more responsible for an outcome the closer their action was to having made a difference to it, as suggested by Chockler and Halpern (2004). In the version of our hypothetical scenario above, where 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 making a difference to the outcome of the election than in a scenario where 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 framework predicts that Republicans would blame Suprun more in the first, as compared to the second case.

Previously, Gerstenberg et al. (2018) tested the computational framework in a range of different achievement contexts: participants were asked to attribute responsibility to goalkeepers trying to block penalties, game show contestants trying to win money, and gardeners trying to make flowers bloom. Overall, Gerstenberg et al.'s (2018) experiments showed a close match between the computational framework's predictions and participants' actual responsibility judgments. Nevertheless, they left several questions unanswered. Here, we address three of them.

First, does the model pass a more direct test of its two key components: dispositional inferences and causal attributions? In previous tests of the framework, Gerstenberg et al. (2018) manipulated how expected an agent's action was, and whether the action made a difference to the outcome, to see how these factors affected people's responsibility judgments. While participants' responsibility judgments were consistent with the computational framework, Gerstenberg et al. didn't test the components of their model directly. Here, we go one step further: we assess participants' 1) dispositional inferences and 2) causal attributions by asking them to evaluate a) how surprising an agent's actions were, assuming that the more surprising a committee member's vote, the more an observer infers that the vote must have been determined by an unobserved aspect of the committee member's disposition; and b) how important an action was for bringing about the outcome. We then investigate whether these judgments, in turn, predict responsibility judgments as postulated by the computational framework.

Second, do the framework's predictions hold in more complex causal settings? Gerstenberg et al.'s (2018) previous tests of the framework focused on situations in which a single agent brought about an outcome. However, it is often the case that several people together contribute to an outcome, as in our hypothetical example where Clinton won the 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 and Trump's loss. In a different strand of research, Gerstenberg and colleagues have investigated how people distribute responsibility in situations in which the contributions of several individuals combine to yield a group outcome (Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Gerstenberg & Lagnado, 2010; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018; Lagnado, Fenton, & Neil, 2013; Lagnado & Gerstenberg, 2015; Lagnado, Gerstenberg, & Zultan, 2013; Zultan, Gerstenberg, & Lagnado, 2012). Here, we connect this research on responsibility judgments in group settings with Gerstenberg et al.'s (2018) work. In addition to manipulating the causal structure of the situation, we also manipulated action expectations in graded ways. To adequately explain responsibility judgments in these situations, we need a model of causal attribution that infers how important an individual's action was for bringing about the outcome, and a model of dispositional inferences that captures to what extent an action diverted from an expectation.

Third, we test the model further by applying it to the moral domain. So far, the model has been shown to predict how people assign responsibility to differently skilled decision-makers in achievement contexts. Questions of morality naturally elicit judgments of responsibility. Who is to blame for the car crash? Who is responsible for the deaths of the workers who died during the factory fire? We propose that in situations like these, when more than a victory in a soccer game or the growth of a flower is at stake, the weights between dispositional inferences and causal attributions shift. Specifically, 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, Pizarro, & Diermeier, 2015).

In this paper, we report the results of three experiments, each of which was designed to tackle one of these open questions. Experiment 1 provides a direct test of how participants make dispositional inferences (by asking about how surprising a particular action was in a given situation) and causal attributions (by asking about how important an action was for the outcome). Experiment 2 asks participants to make responsibility judgments in a large variety of situations that manipulate action expectations and the causal structure in graded ways. Finally, Experiment 3 applies the framework to the moral domain by manipulating the moral valence of the outcome, as well as what question participants are asked to evaluate. Before describing each experiment in more detail, we provide an overview of our experimental paradigm, followed by a more comprehensive outline of the two key components of the computational model: dispositional inferences and causal attributions. Subsequently, we report our experimental results and relate them to the predictions of our computational framework. We conclude by discussing some challenges that remain to be addressed.

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. Table 1 gives an example of a voting situation 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. At least four votes in favor of the policy were required in order for the policy to be passed. As it turned out, Allie and Christie voted in favor of the policy, while Bridget, Dalia and Emma voted against it. The policy was not passed since only two committee members voted in favor but four votes were required for the policy to pass.

Policy information
Number: # 109383
Supported by: The Democratic party
Votes in favor of policy required: 4

Votes

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

Outcome: The policy was **not passed**. 2 out of 5 committee members voted in favor of the policy and 4 votes were required for the policy to pass.

Figure 1. Exemplary voting scenario.

Experiment 3 did not include information about party affiliation. Instead, we told participants about the content of the policy that was up for vote. 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).

We expected that a committee member's party affiliation in Experiments 1 and 2, and voting for the "morally right" outcome in the morally negative context condition in Experiment 3 (i.e., voting against corporal punishment in schools) would affect participants' expectations about how a committee member would vote. By varying how the committee members actually voted, we manipulated the extent to which the votes of the individual committee members were surprising. We predicted that the more surprising a vote was, the more likely our participants would be to infer that the vote must have been determined by the committee member's character or dispositions (rather than by other factors such as allegiance to the party, or the overall quality of a particular policy).

In all three experiments, the committee members' causal contribution to the outcome was manipulated by varying the patterns of votes and the threshold of votes required for the policy to pass. We predicted that a vote would be seen as more causally important the closer it was to having made a difference to the outcome and the fewer causes had contributed to the outcome. We expand on these predictions below.

Model

We now discuss in more detail how we concretely implemented the two components of the computational framework – dispositional inferences and causal attributions – for the experiments reported here. For each component, we first briefly discuss the broader theoretical background, and then the specific model implementation.

Dispositional inferences

Background. How do we explain other people's behavior? Early attribution theorists suggested Bayesian inference as a normative framework to study this question (Ajzen, 1971; Ajzen & Fishbein, 1975; Fischhoff & Beyth-Marom, 1983; Fishbein & Ajzen, 1973; Morris & Larrick, 1995; Trope, 1974; Trope & Burnstein, 1975). Within the Bayesian framework, we can describe behavioral attributions as arising from a comparison between different hypotheses as explanations for a given action. A hypothesis is chosen as the explanation if it has a high prior probability and if it explains the observed behavior well.

Generally, it has been shown that we consider both internal factors, such as a person's abilities, dispositions, goals, beliefs or desires, and external factors, such as the situation the person was in, as possible behavioral explanations. However, research in attribution theory has also revealed that specifically when we try to make sense of the behavior of others (as compared to our own behavior), we tend to emphasize dispositional or character-based explanations, and neglect the influence of situational and environmental factors (Jones & Harris, 1967; Ross, Amabile, & Steinmetz, 1977).

Moreover, in some domains, we may be generally more likely to make dispositional inferences than in others: Recent work in moral psychology has shown that in the moral domain, we often tend to 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 has been broken (Bartels & Pizarro, 2011; Bayles, 1982; Pizarro, Uhlmann, & Salovey, 2003; Uhlmann et al., 2015; Waldmann, Nagel, & Wiegmann, 2012). Based on these findings, researchers have argued that we are inherently motivated to determine the moral character of others (Uhlmann et al., 2015). We evaluate whether they care for others, whether they are fair and whether they can be trusted. Based on these evaluations, we can determine whether we should cooperate with these others in the future or better avoid them. Thus, when we assign responsibility to an individual in the moral domain, our inferences about that individual's character might play a relatively larger role than considerations about the individual's causal connection to the outcome.

More recently, researchers have modeled behavioral attribution processes computationally (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). These models build on the early work within the Bayesian framework. Importantly, however, they point out that in order to computationally model how an observer attributes another person's action as caused by their abilities, dispositions or mental states, we first need to capture the observer's expectation about how an agent *should* act in a given situation, given these abilities, dispositions, and mental states. From this expectation about how an agent should act, the observer can then work backwards to infer the desires and beliefs that caused the agent's behavior using Bayes' rule (Baker et al., 2009). A general principle that determines how mental states cause actions and that has been much studied at the qualitative level is the *principle of rationality* (Dennett, 1987). It states that we take an "intentional stance" toward others and expect of them that they will act as rationally as possible to achieve their desires and goals, given their beliefs about the world (Baker et al., 2009). Empirical tests of models that represent human action understanding as Bayesian inverse planning based on this principle of rationality have shown a close match between model predictions and human data.

$$\begin{aligned}
\text{Party}_{\text{same}} &\sim \text{beta}(1,\ 3)\\
\text{Party}_{\text{other}} &\sim \text{beta}(3,\ 1)\\
\text{Policy} &\sim \text{beta}(2,\ 2)\\
\text{Vote}_{\text{same}} &= \text{mean}(\text{Party}_{\text{same}}, \text{Policy})\\
\text{Vote}_{\text{other}} &= \text{mean}(\text{Party}_{\text{other}}, \text{Policy})\\
\text{Votes} &= \text{c}(\text{Vote}_{\text{same}}, \text{Vote}_{\text{other}})\\
\text{Yes} &\sim \text{binomial}(\text{size} = \text{length}(\text{Votes}),\ \text{p} = \text{Votes})
\end{aligned}$$

Figure 2. Bayesian model of surprise. *Note*: $\sim$ = distributed as, c() = concatenation of values, p = probability.

Implementation. As described in the previous section, a general idea underlying the computational framework of responsibility judgments is that, in a given situation, people draw inferences about a person's character or disposition from their action, and that these inferences, in turn, affect the extent to which a person is held responsible for an outcome. In this section, we outline how we formalized this idea in our model to generate specific predictions for our experimental paradigm. In our paradigm, the question is how much an observer learns about a committee member from how they voted in a given scenario, and how this inference translates into a responsibility judgment.

We assume that there are three driving forces that affect a committee member's vote: 1) whether or not their party supports the policy, 2) the quality of the policy, and 3) their individual preference. An observer knows 1), and can infer 2) and 3) based on the committee members' votes. We assume that committee members will be held more responsible for the outcome of a vote to the extent that their vote was inferred to be driven by individual preference.

Specifically, we model inference as illustrated in Figure 2. We assume that committee members whose party affiliation is aligned with the party that supports the policy are more likely to vote in favor of the policy ($\text{Party}_{\text{same}}$, where the mean of the prior probability distribution is $p = 0.75$), whereas committee member of the opposing party are more likely to vote against the policy ($\text{Party}_{\text{other}}$, with a mean of $p = 0.25$). In our voting example above, Allie and Bridget should a-priori be more likely to vote in favor of the policy while Christie, Dalia and Emma should a-priori be more likely to vote against it. However, voting behavior is not solely determined by party affiliation. Committee members also take into account the quality of the policy. In our first two experiments, observers do not have any prior information about the policy that might serve as an indicator for its quality. For that reason, we take the mean of the prior probability distribution over the policy to be $p = 0.50$. For our purposes, we assume that there are no general asymmetries in voting behavior between committee members with different party affiliations. For example, a Republican is just as likely to vote for a policy backed by Democrats, as a Democratic is to vote for a policy backed by Republicans. Further, we assume that party affiliation and quality of the policy affect a committee member's vote equally strongly ($\text{Vote}_{\text{same}}$ or $\text{Vote}_{\text{other}}$, depending on the committee member's party affiliation).

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

$$p(\text{Party, Policy}|\text{votes}) \propto p(\text{Votes}|\text{Party, Policy}) \cdot p(\text{Party}) \cdot p(\text{Policy}) \tag{1}$$

Based on these posteriors, the model then forms an expectation about how the committee member of interest should vote. For an example, let's focus on the committee member Bridget in our voting scenario above. Since she is a Democrat and thus from the same party that supports the policy, she is a-priori more likely to vote for rather than against the policy. Now, let's take the evidence into account. There is one additional Democrat, Allie, who voted in favor of the policy. In addition to Allie, one of the Republicans, Christie, also voted for the policy. Thus, two out of the four other committee members voted in favor of the policy. Given Bridget's party affiliation and how the others voted, the model forms an expectation that Bridget is likely to vote in favor of the policy, where we use the mean of the posterior over the committee member's vote as our measure of expectation.

We then 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 ($\text{Vote}_{\text{same}}$ or $\text{Vote}_{\text{other}}$ depending on the committee member's party affiliation). Given that an observer would have expected Bridget to vote in favor of the policy, her actual vote against the policy is somewhat surprising. The inference that Bridget's vote must have been affected by her individual preference (since it's not well-explained by her party affiliation and how the others voted) is then predicted to lead to an increased judgment of responsibility.

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 as shown in Figure 2. `greta` uses Markov-chain Monte Carlo (MCMC) inference to approximate the posterior distribution. The code for the model is available on the project's github repository: `https://github.com/cicl-stanford/voting`

Causal attributions

Background. We now turn to the second key process in the computational framework: A causal attribution about the person's role in bringing about the outcome. One way of capturing whether a person's action is causally connected to an outcome is to run a counterfactual simulation and ask whether the outcome would have been different without the person's action (Lewis, 1973). This test of causation works well in situations that involve only a single agent: Is Martin responsible for the bottle being smashed? Yes, because had Martin not dropped the bottle, the bottle would not have smashed.

However, such a naive use of counterfactuals does not suffice in general as a model of responsibility. Consider again our exemplary voting scenario in Figure 1. At least four of the committee members had to vote in favor for the policy to pass but only two committee members did, so the policy did not pass. When we assign responsibility to the committee members for the policy failing to pass, we probably want to say that Bridget, Dalia and Emma are all responsible for the policy not passing, at least to some extent, because all three of them voted against the policy. However, their individual actions did not make a difference to the outcome because the outcome is causally *overdetermined*: For example, even if Bridget had voted in favor of the policy, it would still have failed since four votes in total were required for the policy to pass.

Halpern and Pearl (2005) introduce a solution to this problem by extending the simple counterfactual model. They define a person's action as a cause of an outcome if the outcome depends on the action *under certain contingencies*. Their definition identifies Bridget, Dalia and Emma as causes even in this case of overdetermination. Building on Halpern and Pearl's (2005) definition of causality, Chockler and Halpern (2004) developed a model of responsibility that makes graded predictions. They define the responsibility of a person for an outcome in terms of the minimal number of changes that have to be made to make the outcome counterfactually dependent on the person's action. Their model predicts that the fewer changes are necessary to move from the actual situation to a situation in which the outcome counterfactually depends on the person's action, the more responsibility is assigned to that person. Gerstenberg and colleagues called this notion the person's *pivotality* for the outcome (for an overview, see Lagnado, Fenton, & Neil, 2013). In a range of experiments (Gerstenberg & Goodman, 2012; Gerstenberg & Lagnado, 2010; Lagnado, Gerstenberg, & Zultan, 2013; Zultan et al., 2012), they showed that a person's pivotality is a significant predictor for how much responsibility people assign to that person for a group outcome in a variety of situations, including sports competitions and strategic gaming.

In our computational model of responsibility, we consider an individual's pivotality as one component of the causal attribution process. However, we believe that when assigning responsibility to individuals in a voting setting, where the causal contribution of each individual that voted in line with the voting outcome is identical, people take an additional factor into account when evaluating the causal contribution of a particular individual to the group outcome; namely, 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 (Darley & Latané, 1968; Latané, 1981; White, 2014). To see that 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 they intervened, the victim would be helped, but nevertheless individuals have a reduced sense of responsibility as the number of people who could help increases. Based on this finding, we predict that in addition to a person's pivotality, people take into account how many causes were involved in bringing about the outcome when evaluating an person's causal contribution to an outcome and, as a consequence of that, their responsibility.

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

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

[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)

where C is the minimal number of changes that are required to make A pivotal for E in S.[2] 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, Bridget's pivotality in our example above is $\frac{1}{2}$ $(\frac{1}{1+1})$, since 1 vote needs to be changed to make Bridget's vote pivotal (either Dalia or Emma would have needed to vote in favor of the policy, rather than against it).

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 number of causes that contributed to the outcome affects how important an individual contribution is perceived. The more causes have contributed to an outcome, the less important is each individual cause perceived.

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

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

where by β_1 and β_2 determine how much weight is put on pivotality and the number of causes when making causal attributions.

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'). For simplicity, we assume that both factors of the model combine additively to affect judgments of responsibility.

$$\textit{Responsibility} = \alpha \cdot \textit{Dispositional inference} + \beta \cdot \textit{Causal attribution} \tag{4}$$

In the remainder of this paper, we report the results of three empirical online studies, designed to answer three outstanding questions: Can the model components be directly assessed by asking people to judge dispositional inferences and causal contributions directly (Experiment 1)? Does the model capture how people assign responsibility in group contexts with more complex causal settings (Experiment 2)? And finally, can the model be applied to the moral domain, where judgments of responsibility play a central role (Experiment 3)?

Experiment 1: Testing dispositional inferences and causal attributions directly

In Experiment 1, the participants' task was to judge to what extent the vote of a politician who voted in a committee on whether a new policy should be passed was 1) surprising and 2) important for the outcome of the vote. With these test questions, we aimed to directly assess the two main components of our computational model; dispositional inferences and causal attributions.

We predicted that participants' judgments of how surprising the vote of an individual committee member was would increase the greater the difference was between the mean

[2] S captures the causal structure of the situation, which is often represented in terms of structural equations. For our voting scenarios, S captures how each committee member voted as well as the number of votes needed for a policy to be passed.

of the posterior distribution of 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. Further, we predicted that judgments of how important an individual vote was would increase 1) the closer the vote was to having been pivotal for the outcome, and 2) the fewer the number of committee members whose votes contributed to the outcome.

Methods

Participants. 40 participants ($M_{age} = 35$, $SD_{age} = 11$, 10 female) 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).

Design. Experiment 1 included 27 voting scenarios. Each scenario featured a different political committee comprised of five members.[3] 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 (pass / did not pass), which political party supported the policy (Democrats / Republicans) and the party affiliation of each committee member. Figure 1 shows an example of one particular situation. 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 surprise and importance model.[4]

Procedure

The experiment was programmed in *Qualtrics*. After receiving instructions, participants answered a set of the 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 1) important and 2) 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 made their ratings on 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 minutes ($SD = 9.29$) to complete the experiment.[5]

[3] Note that unlike the example in Figure 1, we used only male names for the politicians within our actual experiments, in order to control for possible effects of gender.

[4] See Table A1 in the Appendix for a full list of the scenarios.

[5] All materials including data, experiments, and analysis scripts are available here:
`https://github.com/cicl-stanford/voting`

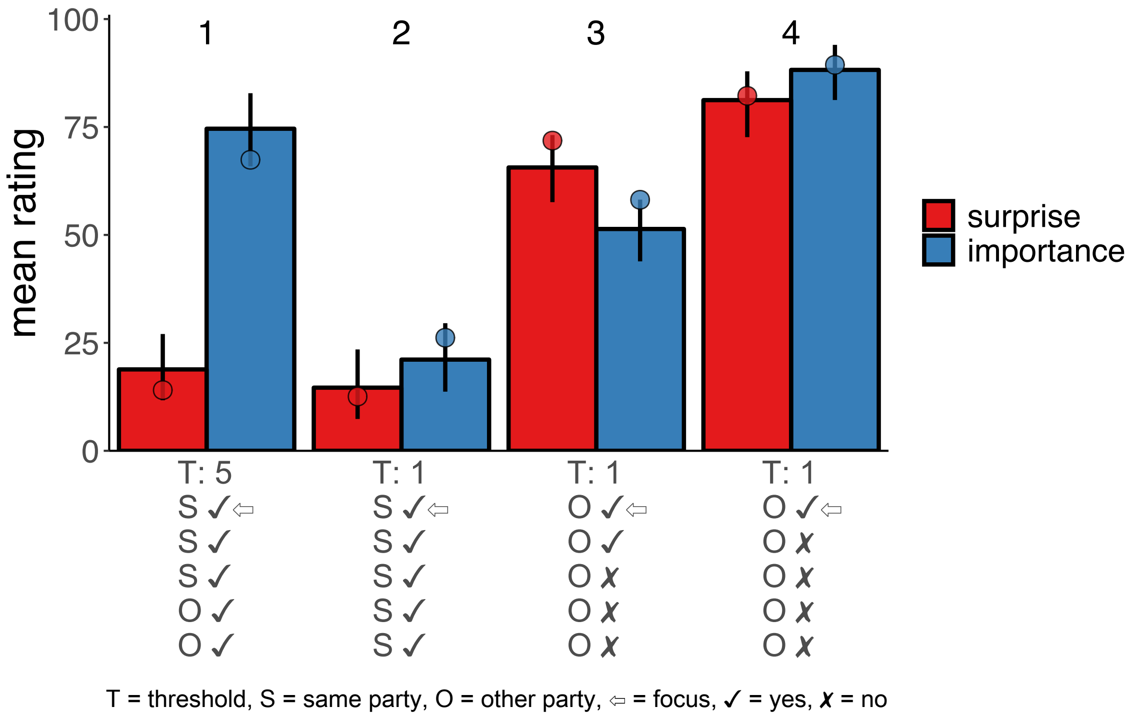


Figure 3. **Experiment 1**: Importance and surprise judgments for trials 1–4. Bars indicate mean judgments, error bars indicate bootstrapped 95% confidence intervals, and points indicate model predictions. The text on the x-axis shows what happened in each situation.

Results

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

Detailed analysis of a selection of scenarios. Figure 3 shows the results of four of the voting scenarios that we used in this experiment. The figure shows participants' mean judgments together with the predictions of the surprise and importance model described above. In all of these four scenarios, the policy was passed because the number of votes in favor met or exceeded the threshold (T).

Let's take a look at surprise judgments first. In all four scenarios, the committee member for whom ratings were assessed (the "focus person", indicated by the arrow in figure 3) voted in favor of the policy. In scenario 1 and 2, the focus person was in the party that supported the policy, whereas in scenario 3 and 4, the focus person was in the other party. We see that in general, participants were more surprised when a person voted in favor of a policy despite being from the opposite party.

However, surprise 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 (10.37 [8.32, 12.37]).[6] The model accurately

[6]For any statistical claim, we report the mean of the posterior distribution together with the 95% continuous highest density interval. Here, for example, the posterior over the difference between scenario 4 and 3 has a mean of 10.37, and the continuous 95% highest density interval ranges from 8.32 to 12.37. All Bayesian

captures this difference. As detailed above, 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. While we cannot observe a policy's quality directly, we can infer the quality by looking at 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. Participants were sensitive to this subtle difference in their surprise judgments, and the model explains how this difference arises.

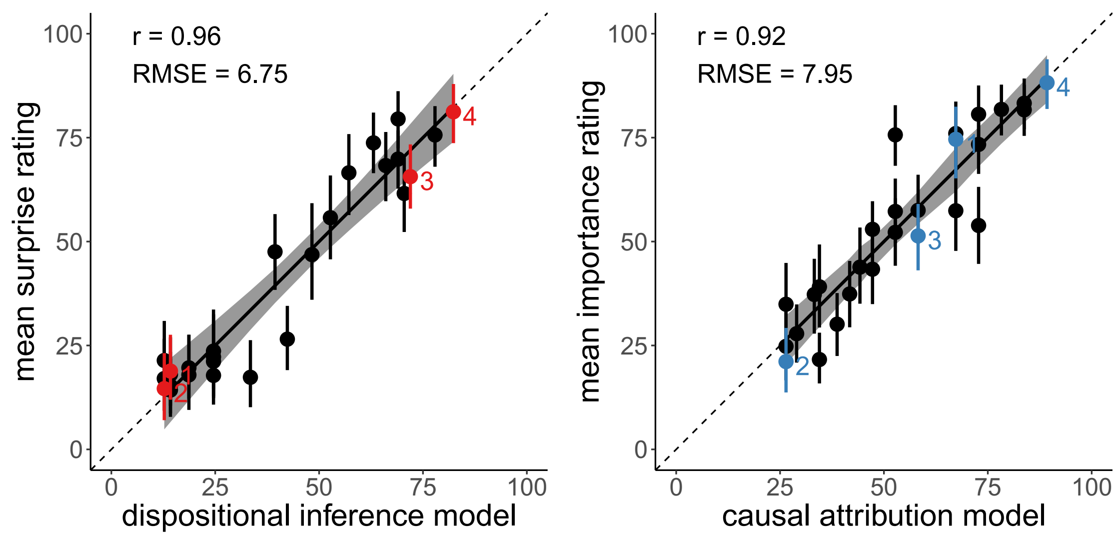
Next, let's take a look at importance judgments. In scenarios 1 and 4, the focus person's vote is pivotal. In both situations, had the focus person voted against the policy, the policy would not have passed. In scenarios 2 and 3, the outcome is overdetermined. However, whereas in scenario 2, all other 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]).

However, 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 compared to 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 the pattern of importance judgments.

Overall results and model comparison. Figure 4 shows scatter plots of the model predictions and participants' surprise and importance ratings for all 27 situations. We fit the model predictions to individual participants' responses specifying linear mixed effects models with random intercepts and slopes. Figure 4 shows the model predictions applied to the mean judgments.

Our *dispositional inference model* captures participants' average surprise judgments very well with r = .96 and RMSE = 6.76 (Figure 4a). A model that considers only whether the committee member voted in line with his party affiliation also correlates well with participants' judgments r = .95 and RMSE = 7.66. We compared the models using using approximate leave-one-out crossvalidation as model selection criterion (PSIS-LOO, cf. Vehtari, Gelman, & Gabry, 2017). According to this criterion, the model that considers only party affiliation performs slightly better than the Bayesian surprise model (difference

models were written in Stan (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019).



(a) Surprise judgments (b) Importance judgments

Figure 4. **Experiment 1**: Surprise and importance judgments. Data points show mean judgments. The colored data points correspond to the four situations shown in Figure 3. The gray ribbon shows the 95% credible interval for the model fit. The error bars indicate 95% boostrapped confidence intervals. *Note*: r = Pearson moment correlation, RMSE = root mean squared error.

in expected log predictive density (elpd) = 26.9, with a standard error of 26.5).[7] However, note that the model that considers only party affiliation just makes two different predictions, and therefore cannot capture the gradedness of participants' judgments.

Figure 4b shows that the *causal attribution model* accounts well for participants' mean importance judgments with $r = .92$ and RMSE = 8.01. Remember that our 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. Our causal attribution model compares favorably with lesioned models that consider only pivotality ($r = .92$ and RMSE = 8.01; elpd = 37.4, standard error = 8.8) or the number of causes that contributed to the outcome ($r = .53$ and RMSE = 17.81; elpd = 233.3, standard error = 23.7).

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 Figure 1). The results show that the extent to which participants found a committee member's vote to be surprising and important for the outcome was systematically affected by this information. To explain participants' surprise judgments, we

[7] 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).

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. While a simple model that only considers whether the committee member of interest voted in line with their party affiliation also explains much of the variance in participants' surprise judgments, it doesn't capture the gradedness of participants' responses.

Participants' judgments about how important a committee member's vote was for the outcome of the vote 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 when it is pivotal (i.e., when the outcome of the overall vote would have been different had the committee member voted differently) and when it was the only cause of the outcome.

Experiment 2: Responsibility judgments in voting scenarios

In Experiment 1, we experimentally manipulated the extent to which a vote was surprising and its importance for the outcome, and assessed how these manipulations affected participants' dispositional inferences and their causal attributions. Since the computational framework predicts that dispositional and causal inferences combine additively to yield responsibility judgments, the extent to which committee members in our voting scenarios are considered responsible for the voting outcome should be influenced by the same experimental manipulations in the exact same way: The more surprising a vote was, the more responsible a committee member should be held for the outcome of the vote. Further, voters should be judged more responsible the closer their vote was to having been pivotal and the fewer committee members voted in line with the voting outcome. To test these predictions, we presented participants in Experiment 2 with voting scenarios like those in Experiment 1 and asked participants to what extent they considered certain committee members to be *responsible* for the outcome of the vote.

Experiment 2 was of much larger scale than Experiment 1. It included 24 of the 27 voting scenarios used in the first experiment, as well as 146 additional ones. We predicted that participants' responsibility judgments for committee members in Experiment 2 would increase to the extent that their votes were surprising and important. Including the 24 voting scenarios from Experiment 1 in the current experiment allowed us to not only look at the extent to which our experimental *manipulations* of surprise and importance affected responsibility judgments, but also to predict the responsibility judgments in Experiment 2 based on participants' surprise and importance judgments in Experiment 1 for a selection of the cases.

Methods

Participants. 208 participants ($M_{age} = 36.24$, $SD_{age} = 13.54$, 86 female) 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%.

Design. In Experiment 2, 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 situations, 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 situations 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 already reduces the number of situations to 340.

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

We split the 170 situations into 10 different conditions with 17 situations each. Each condition included 3 situations with $N_{committee} = 3$, and 14 situations with $N_{committee} = 5$.

Procedure

Participants were randomly assigned to one of 10 conditions. After receiving instructions, each participant made responsibility judgments for a set of 17 situations. 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 reaching from "not at all responsible" (0) to "very much responsible" (100).

Participants were asked only to assign responsibility to committee members whose vote was in line with the outcome. Depending on the situation, 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 at least 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 situations, there were 90 situations in which participants were asked to make a single judgment, and 80 situations in which they made responsibility judgments for two committee members. Thus, we have a total of 250 data points.

In our example situation depicted in Figure 1, three voters voted in line with the voting outcome: Bridget, Dalia and Emma. However, since Dalia and Emma came from the same political party, we only assessed responsibility judgments for one of them in this case. Thus, in this situation, participants made two ratings; one for Bridget (Democrat) and one for Dalia (Republican).

On average, it took participants 6.61 minutes ($SD = 7.03$) to complete the experiment.

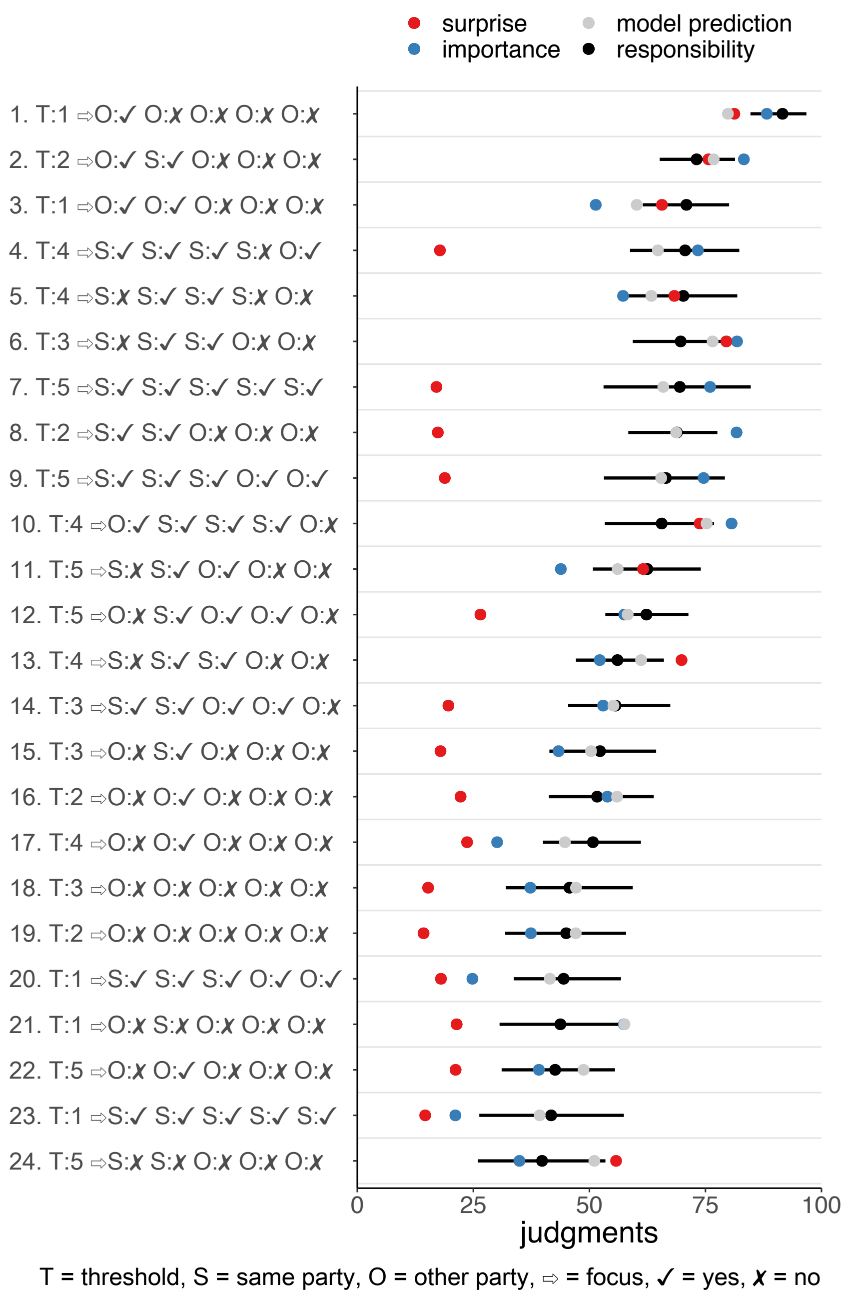


Figure 5. Mean responsibility judgments (black dots) together with the mean surprise (red dots) and importance (blue dots) judgments based on Experiment 1, as well as the model prediction (gray dots) that combines surprise and importance judgments. *Note*: The error bars indicate 95% confidence intervals.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

Results

As in Experiment 1, we first discuss a selection of cases individually before examining the data on a higher level of aggregation to see whether, and to what extent, participants' judgments of responsibility were influenced by dispositional inferences on the one hand and causal contributions on the other hand.

Detailed analysis of a selection of cases. Figure 5 shows participants' judgments for 24 of the 170 situations. These 24 situations are the ones that were 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 situation, the threshold for the policy passing was one (T: 1), all the committee members were from the opposite party (O) than the party which supported the policy. 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 found 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, and the model correctly captures participants' judgment.

In Situation 24, the threshold was 5, but all committee members voted against the policy. Two members were of the same party that supported the policy, and three of the opposite party. Participants in Experiment 1 found it relatively 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 in Situation 1 did (who also voted against the party affiliation). In Situation 1, all other committee members voted against the policy, and the focused member was the only one voting in favor. However, in Situation 24, all of the committee members voted against the policy, thus making the action of the focused member less surprising. Our surprise model captures this because of the effect of the other committee members' votes. In Situation 1, the model infers that the policy overall has little support − hence, the focused member's vote becomes particularly surprising. In Situation 24, the model also infers that the policy wasn't supported, and thus, even though the focused members action is somewhat surprising given his party affiliation, it is less surprising overall because the policy is not supported by anyone.

Participants in Experiment 1 judged that the focus person's action was not particularly important in Situation 24. His vote is far from being pivotal (all of the other four votes would have needed to change), and it's only one among five causes of the outcome. Participants in Experiment 2 judged that the focus person in Situation 24 was not very responsible for the outcome. Again, the model captures this case quite well, although it predicts a slightly higher judgment than what people say.

To derive the model predictions for the 24 situations used in both Experiment 1 and 2, we used participants' mean surprise and importance judgments from Experiment 1 as predictors in a Bayesian mixed effects model of participants' responsibility judgments in Experiment 2, with both random intercepts and slopes. The model accounts very well for the responsibility judgments across the 24 situations, as shown in Figure 5 with r = .86 and RMSE = 5.69. The 95% credible interval 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. Figure 6a shows a scatter plot of the model predictions and participants' responsibility judgments.

Using the surprise models and importance models that were fit to participants' judgments in Experiment 1 as predictors yields a very similar fit to participants' responsibility judgments of $r = .86$ and RMSE = 6.64, with the following posterior estimates for the surprise ($\beta_{\text{surprise}} = 0.20$, 95% HDI [0.08, 0.31]) and importance predictor ($\beta_{\text{importance}} = 0.40$ [0.26, 0.54]). The close correspondence between the predicted responsibility judgments based on participants' empirical surprise and importance judgments and the surprise and importance model was to be expected, given the very high correlations between the models' predictions and participants' judgments in Experiment 1 (see Figure 4).

Overall, we see that participants' responsibility judgments in this selection of 24 situations 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' judgments across the whole range of situations, and also compare the full model to lesioned models that consider only suprise, or consider only importance.

Overall results and model comparison. In order to apply the model to the full set of cases, we took the predictors that are relevant for the dispositional inference component of the model (i.e., the predictions of 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 predicting participants' responsibility judgments based on these predictors including random intercepts and slopes (see Table 1). Figure 6b shows a scatter plot of the model predictions and participants' responsibility judgments for the full set of 170 situations (with 250 judgments). Overall, the model predicts participants' responsibility judgments well with $r = .77$ and RMSE = 8.18. Table 1 shows the estimates of the different predictors. As can be seen, none of the predictor's 95% credible interval overlaps with 0.

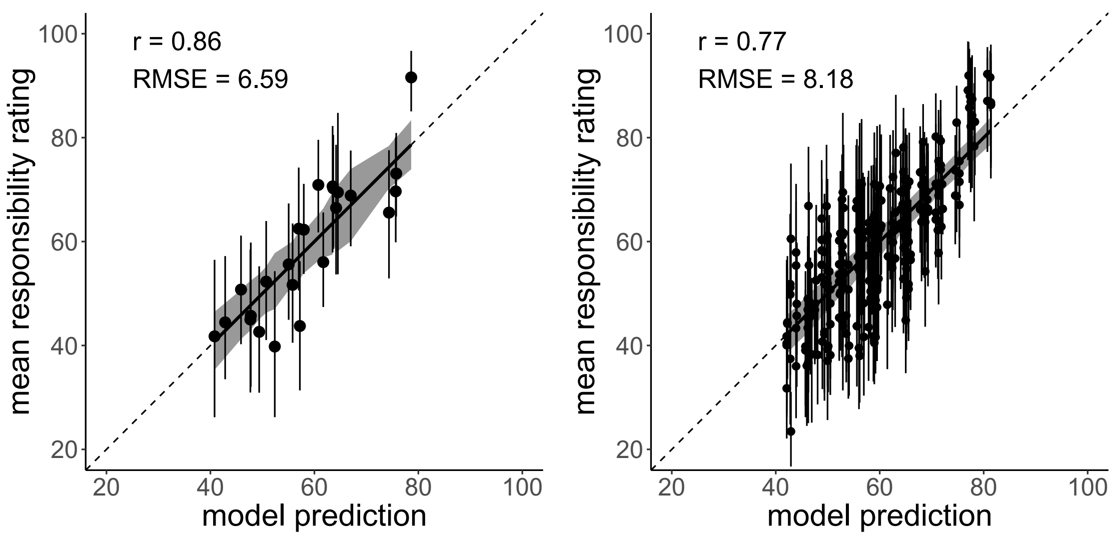
To investigate further whether the different components of the model are needed to adequately capture participants' responsibility judgments, we constructed two lesioned models: one model that considers only the dispositional inference part (i.e., using only `surprise` as a predictor), and one model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

The model that considers only surprise as a predictor performs markedly worse with $r = .37$ and RMSE = 11.98. A model that considers only pivotality and the number of

Table 1
Estimates of the mean, standard error, and 95% credible intervals 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% CI	upper 95% CI
intercept	63.51	2.78	58.82	67.95
pivotality	13.50	1.81	10.53	16.46
n_causes	-5.52	0.48	-6.32	-4.72
surprise	13.36	2.92	8.61	18.12



(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.

Figure 6. **Experiment 2**: Scatter plot between model predictions (x-axis) and mean responsibility judgments (y-axis). The gray ribbon indicates the 95% credible interval for the regression line. The error bars indicate 95% bootstrapped confidence intervals. *Note*: r = Pearson moment correlation, RMSE = root mean squared error.

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 the 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) = 88.5, with a standard error of 16.3). 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 = 7.63 (difference in elpd = 72.9, 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).

Discussion

In this experiment, we asked participants to make responsibility judgments about individual committee members, for a large set of voting scenarios. Our computational framework captured participants' judgments well. While previous work showed that the model accounts well for responsibility judgments about individual decision-makers in achievement contexts (Gerstenberg et al., 2018), the results of this experiment show that the model also does a good job accounting for responsibility judgments about individuals in group settings.

Furthermore, while the responsibility judgments of 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. The surprise and importance judgments of participants in Experiment 1 allow us to predict the responsibility judgments of participants in Experiment 2.

The results further showed that while both components of the model are important, participants' responsibility judgments were most strongly influenced by the causal attribution aspect of our framework which expresses how important a person's action was for bringing about the outcome. 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, the most relevant component may shift from causal attributions to dispositional inferences.

In addition to the factors that our model considers, we also 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). This effect was not predicted by our model, but could in principle be accommodated by it. Right now, our model assumes that, a priori, committee members are just as likely to vote for or against a policy. It is possible, however, that people consider it generally more likely that committee members will vote against a policy rather than in its favor; that is, that the prior distribution over the policy is skewed toward voting "no". This might be because people assume that committee members only vote "yes" if they really agree with the policy, while voting "no" could be a reasonable thing to do in situations where committee members are against the policy, but also in situations where they do not have a strong opinion about the policy (see, e.g. Ritov & Baron, 1992).

Experiment 3: Responsibility judgments in moral contexts

Gerstenberg et al. (2018) previously tested the computational framework 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.

A strength of the computational framework first presented by Gerstenberg et al. (2018) and further developed here is that it is not restricted to specific contexts or a specific domain. When people assign responsibility to an individual for an outcome in the moral domain, the framework predicts that their judgments should be affected by dispositional inferences on the one hand, and causal attributions on the other hand, just like in achievement contexts.

However, 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 character rather than of its consequences on the act (Bartels & Pizarro, 2011; Bayles, 1982; Cushman, 2008; Gerstenberg, Lagnado, & Kareev, 2010; Pizarro et al., 2003; Schächtele, Gerstenberg, & Lagnado, 2011; Uhlmann et al., 2015; Waldmann et al., 2012). Against this background, we hypothesized that when people make responsibility judgments in the moral domain, the relative weights they assign to dispositional inferences versus causal attributions may shift.

We tested this prediction by manipulating the moral valence of the policy that the committees in Experiment 3 voted on (between participants). While in our previous experiments, participants simply read that a certain number of committee members voted for a policy, participants in the current experiment were informed about the content and the consequences of the policy, as well as how the vote went. One group of participants assigned responsibility in a "morally neutral context", where 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 predicted that causal attributions would affect participants' responsibility judgments in both conditions, but that they would play a smaller role in the morally negative condition compared to the morally neutral condition. We made this prediction because we assumed 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.

Importantly, however, in the moral domain, people are not only concerned with determining whether or to what extent a person is responsible for an outcome. They are also motivated to determine whether the person's action was generally right or wrong (Haidt, 2001) and if so, to punish her for her wrongdoing (Cushman, 2008; Darley, 2009). Intuitively, while considerations about the extent to which a person is responsible for an *outcome* should be influenced by both dispositional inferences *and* causal attributions, this does not seem to be the case for judgments about the moral wrongness of a person's *action* Cushman (2008); Teigen and Brun (2011): For example, when evaluating to what extent it was morally wrong that a particular committee member voted in favor of children being spanked and paddled at school, intuitively, we should make that decision independently of how the other committee members voted.

Irrespective of this intuition about how moral wrongness *should* be determined, it is noteworthy that in everyday life, people often try to excuse their behavior by downplaying their causal contribution to the outcome (see, e.g. Falk & Szech, 2013; Glover & Scott-Taggart, 1975; Green, 1991). For example, I might feel that polluting the environment by taking my car to work is less morally wrong because everybody else does it. This opens up the possibility that causal attributions might affect judgments about the moral wrongness of an action *despite* the fact that, intuitively, this should not be the case.

To test how dispositional inferences and causal attributions would affect judgments about the moral wrongness of an action, Experiment 3 included a third condition. In this condition, the policy was also a request to reintroduce corporal punishment in schools. Thus, the moral context of the vote was negative. 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 our experimental manipulations of causal importance.

To sum up, our key predictions in Experiment 3 were that, first, for judgments of responsibility, the importance of causal attribution would decrease in the morally negative compared to the morally neutral context condition. Second, we predicted that for judgments of moral wrongness, the causal attribution component of the model would not matter.

Methods

Participants. 314 participants were recruited via the UK-based internet-platform *Prolific*. Inclusion criteria were English as native language and an approval rate not lower than 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} = 30.83, SD_{age} = 9.3).

Design. As in Experiment 1 and 2, we presented participants in Experiment 3 with scenarios in which a political committee voted on whether or not a motion should be passed. As before, 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). Based on different combinations of these factors, we constructed five different voting scenarios whose structure is illustrated in Figure 7). To keep the scenarios somewhat more simple, the size of the committee (5 members) and the outcome of the vote (policy passed) were held constant in this experiment. In addition, the focus person, for which we assessed responsibility or moral wrongness ratings, was always described as having voted in favor of the policy. Like in Experiments 1 and 2, the focus person's causal contribution to the outcome varied based on how the remaining committee members voted. For example, in situation 2 in Figure 7), the threshold for the policy to pass is 1 and in addition to the focus person, one other committee member ended up voting in favor of the policy. Thus, in this situation, the focus person's pivotality is 0.5 and the voting outcome has two causes.

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. In the current experiment, participants did not receive any information about party affiliation and party support. Instead, we manipulated participants' expectations about how a committee member would vote via information about the moral context of the vote: We specified the moral valence of the policy, such that participants either made their judgments in a "morally neutral" or in a "morally negative" context. In addition, we varied the test question. Instead of asking for responsibility judgments, one group of participants was asked to evaluate the extent to which they considered a particular committee member's action morally wrong. Thus, Experiment 3 had three conditions: *neutral* (morally neutral context, responsibility judgments), *moral* (morally negative context, responsibility judgments) and *wrongness* (morally negative context, moral wrongness judgments).

Procedure

Experiment 3 was programmed in *Unipark*, a German online survey platform. Participants were randomly allocated to one of three conditions. After receiving instructions, they were presented with all five voting scenarios in randomized order. After each scenario, participants made responsibility or moral wrongness 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 that the font of all government documents is changed into Arial", participants in the morally negative context condition read "To what extent is Dallas responsible that corporal punishment is introduced in schools" and participants in the moral wrongness condition read "To what extent is it morally wrong that Dallas voted in favor of introducing corporal punishment in schools?".

Participants made their rating using a slider reaching 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. Participants in the morally neutral context condition were asked "How do you morally judge voting in favor of changing government documents into Arial?" and participants in the morally negative context condition and in the moral wrongness 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".

After having completed all five scenarios, participants received 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 wrongness 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.[8]

On average, it took participants 5.63 minutes ($SD = 2.52$) to complete this experiment.

Results

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

To test our prediction that the different experimental conditions would affect the extent to which the causal attribution component of the model matters, we computed a Bayesian mixed effects model with the two predictors that capture the causal attribution

[8] 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 *wrongness*) in the morally negative context conditions judged introducing corporal punishment in schools as bad. However, there were also a number of participants in the morally neutral context condition who indicated that they considered changing the font of government documents into Arial as bad or good (8 participants), and a number of participants in the morally negative context conditions who indicated that they considered introducing corporal punishment as neutral or good (25 participants in *moral* and 12 participants in *wrongness*). As reported above, these participants were excluded from subsequent analyses.

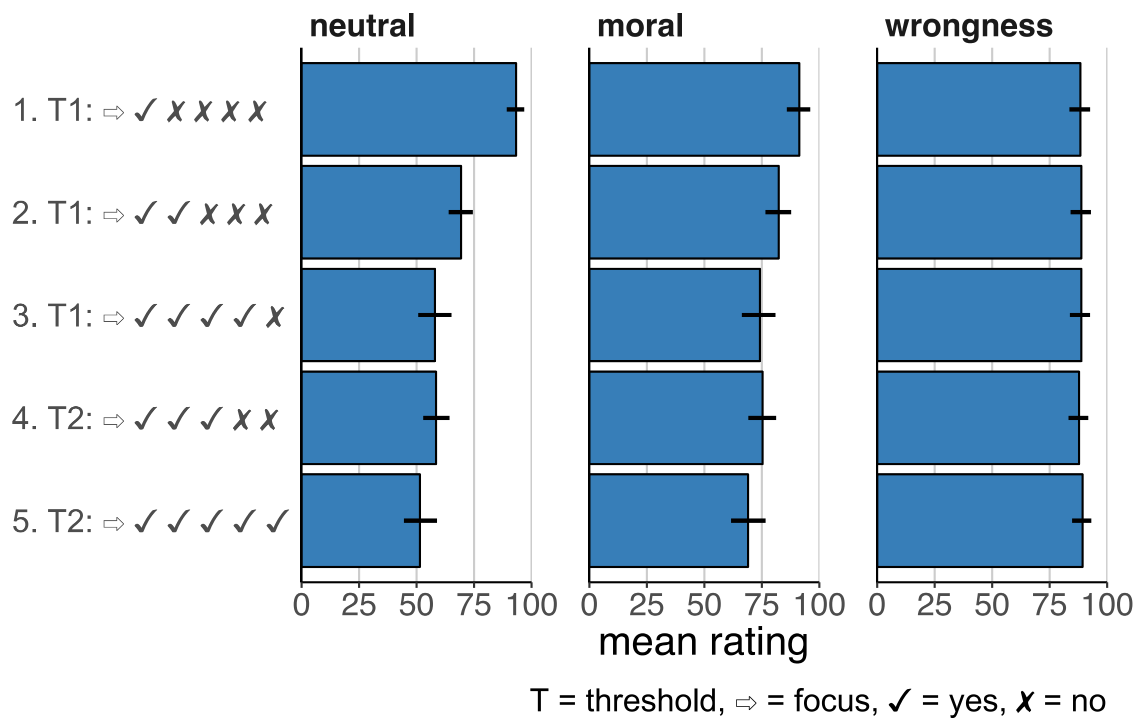


Figure 7. Experiment 3: Participants' mean ratings in the *neutral*, *moral*, and *wrongness* conditions. Note: Error bars indicate 95% confidence intervals. The error bars indicate bootstrapped 95% confidence intervals.

part of the model (`pivotality` and `n_causes`), as well as `condition` and its interactions with the other predictors. Because we only had five observations for each participant, we included only random intercepts but no random slopes.

Table 2 shows a summary of the regression results. As predicted, the extent to which causal attributions affect participants' judgments differed between conditions. In particular, what role pivotality played for participants' judgments differed between conditions as indicated by the pivotality:condition$_{\text{moral}}$ (-29.86 $[-44.92, -14.52]$) and the pivotality:condition$_{\text{wrongness}}$ estimates (-39.06 $[-53.64, -24.35]$). 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 38.45 $[22.63, 54.09]$ and -2.78 $[-5.87, 0.30]$, respectively. In the moral condition, they were 8.90 $[-5.85, 23.12]$ and -3.71 $[-6.44, -1.01]$. Finally, in the wrongness condition, the estimates were -0.10 $[-4.45, 4.16]$ and 0.17 $[-0.69, 1.03]$.

Consistent with our hypothesis, pivotality had a less strong effect on participants' responsibility judgments in the moral condition than the neutral condition. Further, it did not seem to affect participants' wrongness judgments at all in the wrongness condition.

Discussion

In this experiment, we applied the computational framework to the moral domain. We used a similar setup as in Experiments 1 and 2 featuring individuals in a group voting for an outcome. However, instead of manipulating voting expectations via information about the committee members' party affiliation, 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 hypothesized 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 hypothesized that judgments of moral wrongness would not be affected by the causal role of the person at all.

We first consider the two conditions in which we assessed responsibility judgments: *neutral* and *moral*. Our experimental manipulation of pivotality was a good predictor for participants' responsibility judgments in both of these conditions, but the impact of this component was smaller in *moral* than in *neutral*. While we did not assess dispositional inferences about the committee member's directly, this result is in line with our idea, inspired by recent work in moral psychology (e.g., Uhlmann et al., 2015), that in the moral domain, the weights that people put on the two components of the computational model shift so that the influence of causal attributions decreases and dispositional inferences play a larger role. In other words, it seems that participants in *moral* learned something a lot more relevant about the focus committee member as a person than participants in *neutral*; namely, that the focus member is "immoral" or "a bad person".

Now consider on our third condition: *wrongness*. Unlike in *neutral* and *moral*, our experimental manipulation pivotality did not predict the extent to which participants considered the focus committee member morally wrong. This is exactly what we predicted, and conforms to previous work (Cushman, 2008; Darley, 2009; Teigen & Brun, 2011) and

Table 2
Estimates of the mean, standard error, and 95% credible intervals of the different predictors in the Bayesian mixed effects model. Note: `n_causes` *= number of causes.*

`responsibility ~ 1 + (pivotality + n_causes) * condition + (1 | participant)`

term	estimate	std.error	lower 95% CI	upper 95% CI
intercept	55.03	7.14	43.51	66.81
pivotality	38.83	6.33	28.41	49.11
n_causes	-2.71	1.22	-4.70	-0.69
$\text{condition}_{\text{moral}}$	30.29	10.40	12.93	46.96
$\text{condition}_{\text{wrongness}}$	33.57	10.05	16.84	49.96
$\text{pivotality:condition}_{\text{moral}}$	-29.86	9.26	-44.92	-14.52
$\text{pivotality:condition}_{\text{wrongness}}$	-39.06	8.88	-53.64	-24.35
$\text{n\_causes:condition}_{\text{moral}}$	-1.00	1.79	-3.89	1.91
$\text{n\_causes:condition}_{\text{wrongness}}$	2.87	1.72	0.05	5.72

common intuition: When making a judgment about the extent to which a person's action was morally wrong, we should make that decision independently of what other people did in that situation. Thus, while people might often 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.

General Discussion

In this paper, we further developed and extended a computational framework of responsibility judgments, originally introduced by Gerstenberg et al. (2018). This framework predicts that we assign responsibility to an individual for an outcome based on two cognitive processes: dispositional inference and causal attribution. Here, we tested the framework in a voting setting in which multiple members of a political committee voted on whether or not a policy should be passed. This setting allowed us to quantitatively manipulate information relevant to the two key components of the computational model, and systematically investigate how this affects people's responsibility judgments. Specifically, we manipulated the causal structure of the situation (by having committees of different size and different thresholds of how many votes were required in order for a policy to pass), the political affiliation of the committee members (i.e., whether a committee member was affiliated with the party who supported the policy), how each committee member voted, whether the policy was passed, and the moral context of the vote (*neutral* or *moral*).

These factors, in turn, affect the predictions of our computational framework. Specifically, the party affiliations, voting pattern, and the moral context of the vote affect the extent to which a particular committee member's vote is *surprising*. For example, an individual committee member's vote is particularly surprising in a situation in which 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 the threshold to pass was 1 and none of the other committee members voted in favor of the policy.

Within our framework, we map surprise and importance onto the two key cognitive processes: surprise is linked to *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. And 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.

Experiment 1 directly tested the model's key components by assessing participants' judgments of surprise and importance. As predicted, the extent to which participants considered the vote of an individual committee member surprising was affected by the committee member's party affiliation and by how the other committee members voted. 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.

In Experiment 2, we showed that for the subset of voting situations that were used in both Experiment 1 and 2, participants' surprise and importance judgments from Experiment 1 predicted participants' responsibility judgments in Experiment 2. In addition, our responsibility model accounted well for the overall set of patterns that were employed in this large-scale experiment: Participants held committee members more responsible for the outcome when their vote was more surprising, when they were closer to being pivotal, and when fewer causes contributed to the outcome of the vote. It turned out that in addition, the outcome of the vote also affected participants' responsibility judgments. This was not predicted by our account; however, it could in principle be accommodated in our model by assuming that committee members are generally more likely to vote against rather than in favor of a policy (cf. Ritov & Baron, 1992).

In Experiment 3, we applied the computational framework to the moral domain. We showed that, as in previous experiments, participants who made responsibility judgments in a morally negative context considered both what they learned about a committee member as a person, as well as how much the committee member contributed to the voting outcome in their responsibility judgments. However, the impact of the causal attribution component was smaller in the morally negative context condition than in the the morally neutral context condition. We also showed that when participants made judgments about the moral wrongness of a committee member's action, the causal contribution of the committee member's action to the outcome did not play a role. Thus, across the three conditions of Experiment 3 (*neutral* to *moral* to *wrongness*), we showed that the weight people put on the causal contribution factor decreased. These results are consistent with recent proposals in the moral psychology literature that, for questions of moral concern, people predominantly focus on information that is indicative of a person's character or disposition (Bartels & Pizarro, 2011; Bayles, 1982; Cushman, 2008; Gerstenberg et al., 2010; Pizarro et al., 2003; Schächtele et al., 2011; Uhlmann et al., 2015; Waldmann et al., 2012).

In total, the work presented in this paper makes three major contributions toward a comprehensive computational framework of responsibility judgments: First, we developed specific computational implementations for the dispositional inference and causal attribution components of our framework. While prior work had only indirectly tested how these components relate to responsibility judgments (Gerstenberg et al., 2018), we provide a more direct test here. We show that our computational models of dispositional inference and causal attribution accurately predict participants' surprise and importance judgments in Experiment 1, respectively, and that these two components are critical for capturing participants' responsibility judgments in Experiment 2.

Second, this work connects prior research on how expectations and dispositional inferences affect responsibility judgments to individual decision makers, with research that has looked at how responsibility is attributed to individuals in group contexts (Gerstenberg & Lagnado, 2010; Koskuba et al., 2018; Lagnado, Fenton, & Neil, 2013; Lagnado & Gerstenberg, 2015; Zultan et al., 2012). The voting paradigm allowed us to manipulate prior expectations in a natural way, and the results showed that these expectations influenced responsibility judgments. Further, the paradigm featured relatively complex causal settings with expectations manipulated in graded ways, and thus provided a challenging test bed for our computational model.

Third, we show that our framework can be applied to the moral domain as well; our results suggest that people's responsibility judgments in moral contexts are affected by dispositional inferences and causal attributions, just like in other settings. Beyond this more general contribution, the successful application of the framework to the moral domain has important practical implications. For example, a currently much debated topic is how we can design artificial intelligence that behaves responsibly in critical situations (Allen & Wallach, 2009; Friedenberg & Halpern, 2019; Himmelreich, 2019; Mao & Gratch, 2006). How do we want a self-driving car to behave when it has the option to either hit a child on a bike that suddenly turned into its lane or to avoid the child by swerving into an oncoming lane, hitting into another car on this lane and injuring its passengers (Awad et al., 2018; Rahwan et al., 2019)? To answer questions like this one, we need to work toward a prescriptive model for how responsibility should be assigned. Formalizing notions of how people actually assign responsibility in a variety of moral contexts is an important first step toward this goal.

In the remainder of this paper, we discuss avenues for future research on the computational framework. We do so separately for each of the framework's key components; dispositional inferences and causal attributions.

Future directions: dispositional inference

The evidence presented here and in Gerstenberg et al. (2018) clearly speaks in favor of the idea that dispositional inferences are a key component of responsibility judgments. However, ongoing research is needed to develop our understanding of the exact role of dispositional inferences for judgments of responsibility further. In what follows, we discuss three questions that future work should address in more detail: First, what is the role of prior expectations; second, how do action expectations map onto responsibility judgments; and third, how can mental states be incorporated into the framework?

The role of prior expectations. As we have pointed out many times throughout this paper, a central idea within the computational framework of responsibility judgments is that people compare their expectation about how an agent *should* act with the agent's actual behavior.

However, we have not yet specifically discussed where people's initial expectations about how one ought to behave come from. Prior expectations about a person's behavior can stem from at least two different sources. They can be based on what *any reasonable person* would do in that situation or on what the *specific person* under consideration would do (cf. Sytsma, Livengood, & Rose, 2012). In our introductory example, given the angry reactions Suprun's decision to vote for a candidate other than Trump triggered in many Republicans, it seems like they expected him to vote how any "reasonable Republican" would vote in that situation: for Trump. However, imagine that Suprun's friends and family members know him as a strong character, willing to make unconventional decisions if that means he can stay true to his principles. For them, his decision might have been less surprising because they compared Suprun's behavior to what they believed he would do, given his earlier behavior.

In the experiments reported here, participants did not have any specific background information about the committee members. However, in Experiments 1 and 2, participants knew the committee members' party affiliation, and participants in the morally negative context condition in Experiment 3 knew what the "morally right" voting decision was (voting against corporal punishment in schools). Hence, we predicted that this information would affect participants' expectations about how a committee member would vote, so expected participants to use something closer to a "reasonable person" comparison standard. In future work, it would be interesting to compare more specifically between expectations resulting from more general versus person-specific standards of comparisons and investigate how they may influence people's responsibility judgments differently.

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: Our model predicts that the more surprising a committee member's vote, the stronger the dispositional inference about that committee member and thus the higher the level of responsibility that is assigned to the committee member for the outcome of the vote.

It is important to note, however, that in other settings, this direct mapping may not be possible. In achievement contexts, for example, actors are sometimes given more responsibility for unexpected actions (Brewer, 1977; Fincham & Jaspars, 1983; Malle, Monroe, & Guglielmo, 2014; Petrocelli, Percy, Sherman, & Tormala, 2011), and sometimes for expected actions (Johnson & Rips, 2015). In their initial paper on the computational framework for responsibility judgments, Gerstenberg et al. (2018) were able to explain these puzzling findings by demonstrating that violating expectations in itself does not result in more (or less) responsibility, but that the dispositional inference *mediates* the relationship between action expectations and responsibility judgments. In other words, 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 framework predicts that in this context, unexpected actions will yield more responsibility or credit. 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 framework predicts that unexpected actions will be assigned less credit.

Thus, while in our voting paradigm we can map action expectations directly onto responsibility judgments, the computational framework is flexible enough to explain how, in other settings, action expectations can affect responsibility judgments differently. In future work, it would be interesting to test the framework in even more diverse settings and investigate how exactly action expectations might map onto responsibility judgments in each of them.

Modeling mental states. Responsibility is a rich and multifaceted concept (cf. Hart, 2008). Research has shown that there are many different factors that can affect the way in which people assign responsibility. One group of "inputs" to responsibility judgments that has received particular attention are the mental states of the agent whose responsibility is assessed (see Young & Tsoi, 2013, for a review). For example, it has been shown that 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; Gerstenberg et al., 2010; Guglielmo & Malle, 2010; Pizarro et al., 2003; Schächtele et al., 2011), and whether they were able to foresee the consequences of their actions (Lagnado & Channon, 2008; Markman & Tetlock, 2000; Young & Saxe, 2009) when they assign responsibility to those agents.

Several models of how exactly mental states affect responsibility judgments have been put forward. For example, Malle, Guglielmo, and Monroe's (2014) Path Model of Blame predicts that once an observer has established whether an agent has caused an outcome, considerations about the agent's intention lead the observer to two different information-processing paths to blame: If the agent is found to have caused the outcome intentionally, the observer considers his reasons for acting. In contrast, if the outcome was unintentional, the observer considers the agent's obligation and capacity to prevent the harmful outcome from occurring (see also Monroe & Malle, 2017). While empirical tests of the Path Model have yielded convincing results, the model's disadvantage is that it is purely qualitative; as most other theories of responsibility judgments, it leaves unanswered the question of how *much* responsibility people assign to an individual for an outcome in a given situation.

An alternative *computational* model for responsibility judgments that can make quantitative predictions about responsibility and that explicitly incorporates mental states is due to Mao and Gratch 2003; 2005; 2006. The central idea underlying their model is that responsibility judgments are formed based on an agent's causal knowledge on and on conversation interactions from which the observer infers the agent's mental states. The model has not been extensively tested so far. However, the few tests that have been conducted show some limitations (Mao & Gratch, 2006). Specifically, the model diverges from people's actual responsibility judgments in situations that involve several people who contributed to an outcome.

So far, mental states have not been explicitly incorporated within the computational framework for responsibility judgments presented here. For example, in our voting experiments, it is taken for granted that the committee members cast their votes intentionally, with full awareness of the consequences of their actions. Explicitly modeling mental states within the computational framework will be an important next step for future research. We suggest that an agent's mental state should affect the framework's dispositional inference component. Specifically, an agent's mental state should affect the expectations an observer develops about how the agent will act, and thus the extent to which the observer draws an inference about the agent's character. For example, when someone knowingly votes in favor of a policy that has morally negative consequences, this most likely leads us to a stronger inference about that person's character than when someone votes in favor of such a policy without actually knowing what these consequences are. As mentioned earlier, researchers have recently begun to model how people infer mental states from actions computationally (Baker et al., 2017, 2009; Jara-Ettinger et al., 2016). This work could be a good starting point for the endeavour of modeling mental states within the computational framework for responsibility judgments.

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, as well as by incorporating the number of causes that contributed to the outcome as an additional sub-component of the causal attribution process.

However, the model of causal attributions employed here is still a fairly simple one. We know from other work in causal cognition that causal attributions can be far more nuanced than this (e.g., Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Allen et al., 2015; Einhorn & Hogarth, 1986; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Lombrozo, 2010; White, 2014; Wolff, 2007). In what follows, we briefly discuss two factors that have been shown to affect causal attributions and that future work on the computational framework for responsibility attributions could take into account: The extent to which the cause is spatiotemporally connected to the outcome, and the causal function that determines how the individual actions are integrated to yield the group outcome.

Physical processes. As established earlier, the computational framework for responsibility judgments is based on counterfactual theories of causation (Lewis, 1973), which capture causation by determining whether the candidate cause made a difference to the outcome. However, in philosophy, there is a second major theoretical framework for thinking about causation besides counterfactual theories: so-called *process theories of causation*. Process theories establish causal relationships by analyzing whether there was a physical connection that linked the candidate cause and the effect. Empirical work has shown that in some situations, participants' causal judgments are predominantly influenced by information about physical connections (Dowe, 2000; Lombrozo, 2010; Walsh & Sloman, 2011; Wolff, 2007), whereas in others, participants use counterfactual analyses to infer causation (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014).

So far, the computational framework's focus on counterfactual analyses has proven successful at capturing people's causal attributions. It should be noted, however, that we have not yet explicitly manipulated information about physical connections. In our voting scenarios, a committee member's causal contribution to the voting outcome depends only on how the other committee members voted. The physical connection (if one can call it that at all) between each committee member that voted in line with the voting outcome and the outcome is identical: the committee member's vote. In real life, however, we are often confronted with situations in which the individual physical connections of different group members to an outcome differ. In the European Parliament, for example, seats are allocated to each country according to population; thus, larger countries get more votes. Based on previous research (Dowe, 2000; Lombrozo, 2010; Walsh & Sloman, 2011; Wolff, 2007), it seems plausible that people would integrate such information into their responsibility judgments. Hence, future research on the computational framework should look at situations in which the extent to which the individual causes are physically connected to a joint outcome differ quantitatively.

Causal integration functions. Another aspect that has been shown to affect judgments of causation and of responsibility is the way in which individual contributions combine to determine a group outcome (Gerstenberg & Lagnado, 2010; Waldmann, 2007). For example, are the individual contributions added together, so each cause contributes something to the overall outcome (addition)? Do all causes need to surpass a certain threshold (conjunction)? Or is one cause sufficient for bringing about the outcome (disjunction)?

In our experiments, we varied the way in which the individual votes combined to yield the voting outcome to some extent by varying the threshold of votes required for the policy to pass. For example, a voting situation in which the threshold was five represents a conjunctive situation, while a situation in which only one committee member had to vote in favor of the policy for the policy to pass represents a disjunctive situation. As reported above, the way in which our participants assigned responsibility indicates that they were sensitive to these variations. In future work, it would be interesting to more specifically manipulate different causal functions. For example, one could create situations in which the individual committee members differ in how much power they have to influence the voting outcome and look at how this affects responsibility judgments.

Conclusion

Deciding whether and to what extent someone is responsible for an outcome is something we do on a day-to-day basis. In this paper, we have further tested and extended a computational framework Gerstenberg et al. (2018) that postulates two key processes in responsibility judgments: dispositional inferences and causal attributions. We have shown that the framework's predictions hold when its two key components are assessed directly, when the framework is employed in more complex causal settings, and when it is tested in the moral domain. In doing so, we have provided further evidence that the computational framework for responsibility attribution is applicable as a unified tool for obtaining quantitative predictions about how people assign responsibility to others in a variety of contexts.

Acknowledgments

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 by 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
Scenarios presented in Experiment 1

Table A1
List of 27 trials presented in Experiment 1.

		party					vote						
trial	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 = same party as the party who 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 B
Scenarios presented in Experiment 2

Table B1
List of 30 situations with three committee members in Experiment 2. Note that if there was a member from each party who caused the outcome, participants were asked to assign responsibility to each member on separate sliders.

	party			vote				
trial	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 = same party as the party who 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 B2

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

	party					vote						
trial	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

	party					vote						
trial	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
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
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

	party					vote						
trial	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
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

	party					vote						
trial	p1	p2	p3	p4	p5	v1	v2	v3	v4	v5	threshold	outcome
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
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 = same party as the party who 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.

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, C., & Wallach, W. (2009). Moral machines.

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 et al. (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., … Rahwan, I. (2018, October). The Moral Machine experiment. *Nature*. doi: 10.1038/s41586-018-0637-6

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, mar). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. Retrieved from `https://doi.org/10.1038%2Fs41562-017-0064` doi: 10.1038/s41562-017-0064

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

Bartels, D. M., & Pizarro, D. A. (2011, oct). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161. Retrieved from `http://dx.doi.org/10.1016/j.cognition.2011.05.010` doi: 10.1016/j.cognition.2011.05.010

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. doi: 10.18637/jss.v080.i01

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

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, jul). Moral appraisals affect doing/allowing judgments. *Cognition*, *108*(1), 281–289. Retrieved from http://dx.doi.org/10.1016/j.cognition.2008.02.005 doi: 10.1016/j.cognition.2008.02.005

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.

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

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

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

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

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

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. *arXiv preprint arXiv:1903.04102*.

Gerstenberg, T., & Goodman, N. D. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.

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., & 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, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from https://doi.org/10.1177%2F0956797617713053 doi: 10.1177/0956797617713053

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

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

Golding, N. (2018). greta: Simple and scalable statistical modelling in r [Computer software manual].

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

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., ... 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., & 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.

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.

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.

Himmelreich, J. (2019, June). Responsibility for Killer Robots. *Ethical Theory and Moral Practice.*

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. Retrieved from `https://doi.org/10.1016%2Fj.tics.2016.08.007` doi: 10.1016/j.tics.2016.08.007

Johnson, S. G., & Rips, L. J. (2015, mar). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76. Retrieved from `http://dx.doi.org/10.1016/j.cogpsych.2015.01.003` doi: 10.1016/j.cogpsych.2015.01.003

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of experimental social psychology*, *3*(1), 1–24.

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., & 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., Guglielmo, S., & Monroe, A. E. (2014, Apr). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186. Retrieved from http://dx.doi.org/10.1080/1047840x.2014.877340 doi: 10.1080/1047840x.2014.877340

Malle, B. F., Monroe, A. E., & Guglielmo, S. (2014, apr). Paths to blame and paths to convergence. *Psychological Inquiry*, *25*(2), 251–260. Retrieved from http://dx.doi.org/10.1080/1047840x.2014.913379 doi: 10.1080/1047840x.2014.913379

Mao, W., & Gratch, J. (2003). The social credit assignment problem. In *International workshop on intelligent virtual agents* (pp. 39–47).

Mao, W., & Gratch, J. (2005). Social causality and responsibility: Modeling and evaluation. In *International workshop on intelligent virtual agents* (pp. 191–204).

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).

Markman, K. D., & Tetlock, P. E. (2000, sep). 'i couldn't have known': Accountability, foreseeability and counterfactual denials of responsibility. *British Journal of Social Psychology*, *39*(3), 313–325. Retrieved from http://dx.doi.org/10.1348/014466600164499 doi: 10.1348/014466600164499

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

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, 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.

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), 267–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12741752

R Core Team. (2019). R: A language and environment for statistical computing [Computer

software manual]. Vienna, Austria.
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... Wellman, M. (2019, April). Machine behaviour. *Nature*, *568*(7753), 477-486. doi: 10.1038/s41586-019-1138-y
Ritov, I., & Baron, J. (1992, Feb). Status-quo and omission biases. *Journal of Risk and Uncertainty*, *5*(1). Retrieved from `http://dx.doi.org/10.1007/BF00208786` doi: 10.1007/BF00208786
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.
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.
Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* Springer-Verlag, New York.
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.
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.
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.
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.
Vehtari, A., Gelman, A., & Gabry, J. (2017, September). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432. doi: 10.1007/s11222-016-9696-4
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.
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., & 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. Retrieved from `http://dx.doi.org/10.1111/cogs`

.12075 doi: 10.1111/cogs.12075

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.

Young, L., & Tsoi, L. (2013, August). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass*, *7*(8), 585–604. doi: 10.1111/spc3.12044

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