10
A Difference-Making Framework for Intuitive Judgments of Responsibility

David Lagnado and Tobias Gerstenberg

INTRODUCTION

'Would-be politician Robert "Bobby" McDonald has learnt the hard way that every vote can count. McDonald, 27, tied with Olivia Ballou for a seat on Walton City Council in Kentucky on 669 votes ... When his wife Katie rang him with ten minutes to go before the polls closed to say she hadn't had time to vote, he told her not to bother ... McDonald, whose fate is likely to be decided by a toss of a coin if a recount does not split them, did not blame his wife, who works nights as a patient care assistant at a hospital and is finishing training as a nurse ... "She feels bad enough," he said. "She worked extra hours, goes to school and we have three kids, so I don't blame her. She woke up about ten minutes before the polls closed and asked if she should run up, but I told her I didn't think one vote would matter."'[1] (*Herald Sun*,[2] November 9, 2012).

Assigning responsibility is a complex matter. Not only are there multiple criteria by which we evaluate people's actions, but there are often distinct senses of responsibility in play. Take Bobby McDonald's wife, Katie, in the

[1] We thank Ro'i Zultan for this example, and for his substantial contribution to the previous work on which this chapter is based. We also thank two anonymous reviewers for insightful comments.

[2] Based on a report in the *Kentucky Enquirer*, available at:

http://www.news.com.au/world-news/robert-bobby-mcdonald-tied-with-olivia-ballou-on-walton-city-council-after-he-told-wife-not-to-vote/story-fndir2ev-1226513833424 voting example. To what extent was she responsible for the tied outcome? Bobby didn't blame her: he had told her not to vote, and she had good reasons for missing the poll. But her failure to vote turned out to be crucial. Had she voted (assuming she would have voted for her husband!), Bobby would have won outright. There seem to be conflicting intuitions here. On the one hand she seems responsible for the outcome—if she had voted, her husband would have won. But on the other hand, it seems wrong to hold her responsible. After all, no one expected the outcome of the election to be this close, and her failure to vote is excusable. Not only was she working hard, but even her husband told her not to bother.

We believe this tension reflects the interplay of distinct principles that operate when people assign responsibility, and the fact that the term "responsibility", depending on context, can refer to one or other of these principles (Cane 2002; Hart 1968/2008; Schlenker et al 1994; Vincent 2011). Based on our previous work on responsibility attribution in groups (Gerstenberg and Lagnado 2010, 2012; Zultan, Gerstenberg, and Lagnado 2012), we outline a novel psychological account of responsibility attribution that identifies two separate but interrelated components. In particular, we argue that common judgments of responsibility are a function of both retrospective and prospective factors. We show that our account maps neatly onto a recently proposed structural taxonomy of responsibility concepts (Vincent 2011). Our account takes difference-making as the core mechanism for attributions, and thus is closely tied to a counterfactual analysis (Lewis 1979; Halpern and Pearl 2005). However, the counterfactual analysis is extended in several ways, avoiding problems of overdetermination, and allowing for graded measures of the key components that constitute our account.

In developing our account, we will focus on social settings such as voting, team competitions, and public goods games in which the contributions of multiple individuals combine to determine the outcome. However, our account also applies to situations in which the outcome is the result of complex mechanical devices. Group contexts are commonplace, and present a challenge to responsibility theories, especially those based on counterfactual analyses, due to problems of overdetermination. Moreover, providing a graded measure of responsibility proves particularly important when assigning responsibility to multiple agents (cf. Braham and van Hees 2009, 2013).

10.1 CONCEPTS OF RESPONSIBILITY

Legal and moral philosophers distinguish various concepts of responsibility (Cane 2002; Hart 1968/2008). In contrast, psychological research on responsibility attribution often treats responsibility as a single concept, subject to diverse influences and biasing factors (Alicke 2000; Knobe 2010; Knobe and Hitchcock 2009; Lagnado and Channon 2008). We maintain that a more fine-grained analysis of the concept of responsibility is appropriate, and that various examples of supposed "irrationality" in people's responsibility attributions might be due to the theorists' failure to allow that people operate with distinct notions. This is not to argue that laypeople's conceptions of responsibility can be fractionated as neatly as those of a philosopher or legal theorist. But we do propose that the underlying cognitive principles of responsibility attribution are more complex and multifaceted than has previously been acknowledged in the literature. This can be illustrated by looking at a recent taxonomy of responsibility concepts advanced in philosophy (Vincent 2011) and seeing how it maps onto psychological theorizing about people's responsibility judgments.

Vincent (2011) proposes a structured taxonomy of responsibility concepts: she delineates six separate concepts and shows how they are structurally related. Her taxonomy is represented in Figure 10.1. The relations between the concepts stand for justificatory relations which we will recast as functional relations in our psychological account below. What Vincent terms "outcome responsibility" corresponds to the main concept of responsibility used in philosophy and psychology. It is a retrospective concept: given an outcome of interest, it involves looking backwards to establish who is responsible for that outcome, as a precursor to assigning blame or praise. However, there are two distinct concepts that determine (or justify) an assignment of outcome responsibility: (i) *causal responsibility* and (ii) *role* (or task) *responsibility*. The former is also a retrospective concept, and corresponds to the causal contribution that the agent made to the outcome. The latter is prospective and often has normative implications (Schlenker et al 1994). It corresponds to an agent's tasks or duties, not just in an institutional sense, but also in terms of what they are expected to do in the context in question. Applied to the voting example, causal responsibility involves the extent to which Katie's action (not voting) caused the

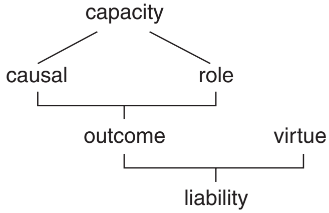

Figure 10.1 Vincent's (2011) structured taxonomy of responsibility concepts.

outcome (poll tied). Role responsibility in this context corresponds to her duties as a potential voter, and as we shall argue below, expectations about what role her contribution could play in the final outcome. Vincent argues that these two concepts combine to determine (or justify) outcome responsibility. Thus, Katie's outcome responsibility for the tied poll is determined both by her causal contribution to the actual result as well as her prospective role in the voting process.

At the top of the structure is the notion of *capacity* responsibility—which concerns the mental and physical capabilities of the agent. It can be seen as furnishing preconditions for causality and role responsibility, or as modulating them in some way. Thus, a suitable mental capacity is required for someone to be assigned role responsibility for a task or duty—they usually need to understand and agree to the role assigned to them, and therefore be capable of carrying it out. In the voting context this might correspond to being eligible to vote: we would not consider Bobby's young children, nor an adult who is not allowed to vote in the election, as having the relevant capacity responsibility. Capacity exerts a similar influence on causal responsibility—presumably the extent to which an agent's action causes the outcome depends on their capabilities to bring about the outcome. This might be cashed out in terms of their levels of skill or competence (Alicke 2000; Guglielmo and Malle 2010; Malle 1999).

At the bottom of the structure is the concept of *liability responsibility*—which corresponds to the final attribution of blame (or praise) once general mitigating or aggravating characteristics of the agent have been taken into account (cf. Shaver 1985). It has a clear instantiation in criminal law, when punishment is to be meted out, but more generally corresponds to the level of redress people are held to for what they are outcome responsible for. *Virtue responsibility* is a blanket notion to cover these general factors (e.g., good or bad character in criminal law). The key point is that *liability responsibility* takes into account more general knowledge about the person's character that might not have stood in any direct relationship to the actual outcome for which the person is held responsible. To illustrate, consider Katie in the voting example. We are reluctant to blame her too much (liability responsibility) because she seems of very good character—working extra hours, attending school, bringing up three kids. These factors give her a good reason to miss the vote, and also accord her good character more generally (virtue responsibility). However, suppose instead that she was unreliable and duplicitous, and was dallying with her lover rather than going to vote. In this case, her outcome responsibility would not change, but we would blame her more for her actions.

Vincent's claim that liability responsibility is a function of both outcome and virtue responsibility resonates with recent research in psychology showing that people's causal and responsibility attributions are influenced by the moral valence of the agents involved (Alicke 1992, 2000; Knobe 2010). However, the psychological research often operates with a "thick" concept of responsibility that runs together outcome and liability responsibility, and it is an open empirical question how sensitive laypeople's concepts are to the finer distinctions given in Vincent's taxonomy. The distinction between outcome and liability responsibility will not be explored in this paper.

Vincent's structural taxonomy is a useful roadmap for the rest of our paper. We will use her structure to motivate some key distinctions we have independently made with regard to psychological components of responsibility attribution (Lagnado et al 2013). Cast in her terms, our prime focus in this paper is on judgments of outcome responsibility, and how they are determined by retrospective assessments of causal responsibility and prospective assessments of role or task responsibility. We replace these placeholders with specific models that generate predictions about how the separate concepts combine to yield graded judgments of outcome responsibility.

10.2 A DIFFERENCE-MAKING MODEL OF OUTCOME RESPONSIBILITY

We will first outline our psychological model of outcome responsibility (for fuller details see Lagnado et al 2013), and then present new empirical data that support this model. The model is based on recent formal work on actual causation (Halpern and Pearl 2005; Pearl 2000) and causal responsibility (Chockler and Halpern 2004). Central to our model is the idea that both prospective and retrospective factors contribute to judgments of outcome responsibility. In particular, when people assign responsibility to an agent for an outcome, they take into account both how critical an agent's contribution is for bringing about a positive outcome (which we term "criticality"), and the agent's causal contribution to what actually happened (which we term "pivotality"). Criticality is assessed ex ante, before the outcome is known, while pivotality is determined ex post, after the outcome is known. As noted above, this maps onto Vincent's taxonomy, with *pivotality* serving as a model for causal responsibility, and *criticality* as a model for role or task responsibility.

At the heart of our account lies the notion of causation as difference-making, which ties causal judgments closely to counterfactuals.[3] However,

[3] Like Halpern and Pearl (2005) we do not endorse a reductive counterfactual analysis of causation but allow for actual causation to be modeled in terms of more general causal

we make use of recent innovations in computer science and philosophy (Chockler and Halpern 2004; Halpern and Pearl 2005; Woodward 2003) which extend the notion of counterfactual dependence in several crucial ways.

The standard approach here is to consider whether an agent made a difference to the outcome, by comparing what actually happened with what *would have* happened had the agent done something different (Lewis 1986; Schaffer 2005, 2010). Applied to the voting example, one compares what actually happened—Katie did not vote and the poll was tied—with what would have happened had she voted. The key question is whether the outcome would have been different. Essentially we are asking whether there is a counterfactual dependence between Katie's action and the outcome. We will use the term "pivotality" for this relation, and follow Halpern and Pearl (2005) by extending the concept to apply not just in the actual situation (e.g., where the votes were tied) but also in counterfactual situations (e.g., where the votes could have been cast differently). Note that the claim that an agent's action is pivotal for an outcome involves two counterfactuals that correspond to the necessity and sufficiency of the action in the given circumstances (cf. Woodward 2006): (i) if the agent had not acted, then the outcome would not have occurred (counterfactual necessity); (ii) if the agent had acted, then the outcome would have occurred (counterfactual sufficiency). What is novel about this analysis is that these counterfactuals are assessed not just in the situation that actually occurred, but also under alternative counterfactual situations. This allows us to deal with problems of overdetermination, and yields a graded notion of responsibility.

10.2.1 Closeness to Pivotality

Pivotality is all-or-none: either the outcome was counterfactually dependent on the action in question, or it was not. Either Katie's vote was pivotal for the election outcome or it wasn't. And in actual fact it turned out to be pivotal. But the notion of pivotality is naturally extended to yield a graded measure of how close an action (or event) was to being pivotal (Chockler and Halpern 2004). Even though Katie's vote turned out to be pivotal in the actual circumstances (e.g., the votes were tied) there were many possible situations in which her vote would not have been pivotal. The notion of "closeness" to pivotality can be illustrated if we imagine how the results of the poll could have turned out differently.

concepts such as the notion of an ideal causal intervention (see Woodward 2003 for a detailed interventionist account of causation).

First, suppose that Bobby's competitor, Olivia Ballou, had won by one vote. In this situation Katie's failure to vote would still have been pivotal for the outcome—if she had voted the outcome would have been a tie. Now imagine that Olivia Ballou won by *two* votes. In this case Katie's vote would no longer have been pivotal for the outcome—even if she had voted for Bobby, he would still have lost. But she was "close" to being pivotal, in the sense that it would have only taken a small change to the actual situation—for example, removing one of Ballou's voters—to make Katie pivotal again. Generally, as the margin between Olivia and Bobby increases, Katie's causal responsibility diminishes, because it would take increasingly more changes to the actual situation to render her pivotal. Taken to the extreme, if there had been a landslide in favor of Olivia, Katie would bear almost no causal responsibility for the outcome. She would have been very far from being pivotal.

These intuitions are cashed out in a formal model of causal responsibility proposed by Chockler and Halpern (2004), and we will adopt their measure for our psychological model. They propose that the degree of causal responsibility assigned to an agent is a function of the number of changes to the actual situation required to make the agent pivotal for the outcome. In particular, using our terminology (where *pivotality*[4] maps onto *causal responsibility* in Vincent's framework):

$$\textit{pivotality}\,(A,O,S) = \frac{1}{(N+1)}$$

where *pivotality* (*A,O,S*) is agent *A*'s causal responsibility for outcome *O* in a setting *S*, and *N* is the minimal number of changes required to make *A* pivotal for *O*. Applied to the voting example, *O* is the outcome "Bobby does not win due to a tied vote," *S* describes the causal structure of the situation, which in this case can be summarized as a simple majority rule, and *N* is the number of voters one would need to change in order to make Katie's action pivotal for *O*. Thus, in the actual situation when the votes are tied, Katie is pivotal for the outcome ($N = 0$)—her *pivotality* is 1. Similarly, in a slightly different situation where Olivia wins by one vote, Katie is still pivotal, because if she had voted the outcome would have been different (the poll would have tied). Thus in this situation her pivotality would still have been 1 (although here she would have been pivotal for the loss rather

[4] Note that the measure of pivotality given by this equation yields a number between 1 and 0, and this corresponds to the *closeness* of the action to being pivotal for the outcome, where being pivotal is still either true or false.

than the tie). However, in the situation where Olivia wins by two votes, Katie is no longer pivotal. A minimum of one vote would have needed to be removed (*N* = 1) to make her pivotal, so her pivotality in this situation is ½. More generally, an agent's pivotality for a particular outcome decreases as *N* increases.

As well as yielding graded responsibility judgments, this extension of the counterfactual analysis allows us to deal with problems of overdetermination. This is crucial if we are to develop an account of responsibility attributions in groups, because overdetermination is rife in such contexts (Gerstenberg and Lagnado 2010, 2012; Zultan et al 2012). In the current voting example we have focused on Katie, and to what extent she was causally responsible for the outcome. But another key question is how to assign responsibility to the voters in the poll more generally—how much causal responsibility does each voter bear for the final outcome? Moving to an analysis in terms of "counterfactual" pivotality provides a solution to this problem: each voter is causally responsible for the outcome according to the number of votes that would need to be changed to render them pivotal. For example, in the actual election every voter was pivotal—but in an election where the gap between the leading candidates is *N* votes, then *N*-1 changes would have been needed to make each voter pivotal.[5]

10.2.2 Criticality

As well as having a retrospective component, judgments of outcome responsibility also have a prospective component. Legal theorists have emphasized this forward-looking dimension (Cane 2002; Hart 1968/2008), and Vincent (2011) captures it in terms of role responsibility. We use the term "criticality" (Lagnado et al 2013) to capture one key prospective component to responsibility judgments.

A person's criticality captures the degree to which a future positive outcome is expected to be dependent on their action. Returning to the voting example, we might ask how critical Katie's vote is for the outcome, before the votes are cast. Neither Bobby nor Katie expected her vote to be critical, which seems a reasonable supposition given the low chance of a tie in

[5] Chockler and Halpern (2004) define the notion of a "change" in terms of an intervention on a variable in the underlying causal model that represents the situation. This means that an agent's pivotality is sensitive to the causal representation of the situation that dictates what changes are licensed. For simplicity, our example rests on the assumption that voters either vote for a specific candidate or refrain from voting. Hence, we rule out the possibility that a voter could have switched from one candidate to the other (see Livengood 2011 for a critical discussion of actual cause theories in the context of voting scenarios). In the empirical studies reported below, the notion of a change is unambiguous.

the election (which decreases as the number of voters increases). We can, however, imagine situations in which Katie's vote would be judged more critical. For example, if there was a small number of voters, or if the vote was expected to be extremely close.

Criticality is a graded notion. As noted above, Katie's criticality will depend on factors such as the number of voters, the expected distribution of votes, as well as her a priori voting power in situations in which voters cast a different number of votes (Felsenthal and Machover 2004; Kerr 1989; Rapoport 1987). Differences in this prospective factor can influence people's responsibility judgments even when events unfold in exactly the same way (e.g., the agent's actions and the outcome stay the same). For example, if we hold constant what actually happened in the voting case (i.e., Katie did not vote and the poll was tied), but vary Katie's perceived criticality, this will affect the degree to which she is judged responsible for the outcome. If the vote is expected to be incredibly tight, then her vote has a higher degree of criticality than if it is expected to be a landslide for Bobby.

There are different ways of defining criticality more formally. A plausible approach would be to define it in terms of an expectation that the agent will make a difference to the outcome; in other words, in terms of the agent's expected pivotality.[6] Applied to the case of an election, this means that a voter's criticality is determined by the probability that the vote they cast will be pivotal (cf. Rapoport 1987).

Defining a person's criticality in terms of their expected pivotality captures our intuitions for the described situation in which the outcome is determined via a majority rule. However, as we shall see, there are situations in which expected pivotality does not correspond to our intuitive sense of a person's criticality. To illustrate the problem, consider a simple voting example with only two voters, Sarah and Jim, each of whom must vote whether or not to pass a bill. Now compare two different voting situations: (i) conjunctive—both agents must vote in favor of the bill for it to be passed; (ii) disjunctive—the bill is passed so long as at least one agent votes in favor. Intuitively it seems that each voter is more critical for the bill being passed in case (i) than in case (ii). But suppose that we focus on Sarah's expected pivotality, and assume that she has no idea whether or not Jim will vote in favor of the bill (i.e., she assigns a probability of $p = 0.5$ that Jim will vote for the bill). In this situation, Sarah's expected pivotality is the same in the conjunctive and disjunctive case. In (i) she will be pivotal to the bill passing when Jim votes in favor; in (ii) she will be pivotal to the bill

[6] Chockler and Halpern (2004) distinguish between the notions of responsibility and blame. Whereas responsibility, in their account, corresponds to our notion of pivotality, blame is defined as expected pivotality.

passing when Jim votes against. Given the assumption that Jim is as likely to vote for the bill as against it, Sarah has the same expected pivotality in both cases. However, intuitively Sarah's vote is more critical in the conjunctive compared to the disjunctive situation.

Indeed, Lagnado et al (2013) have shown that in simple two-agent cases like this voting example, most people judge agents to be more critical in the conjunctive than the disjunctive set-up. Note that expected pivotality corresponds to the probability that an agent's action will be both necessary and sufficient for a particular outcome. However, the asymmetry between conjunctive and disjunctive situations described above suggests that people's perception of criticality is more closely tied to necessity than sufficiency.

Lagnado et al (2013) demonstrated that participants' criticality judgments are closely tied to the expectation that a person's contribution is necessary for the positive outcome (see Pearl 1999). Thus, the criticality of agent A_{i}'s contribution in situation *S*, is defined as the probability that their contribution x_i is necessary for the positive group outcome *y*:

$$criticality(A_i, S) = \frac{p(y \mid x_i) - p(y \mid \neg x_i)}{p(y \mid x_i)}$$

Applied to the two-person voting example: in the conjunctive set-up (i) each voter is fully critical, since the probability of the bill being passed without their vote is zero (e.g., $p(y \mid \neg x_i) = 0$. In the disjunctive set-up (ii), an agent's criticality depends on the probability that the other voter will vote for the bill. Assuming that voter A_{i} is uncertain about whether or not voter A_{j} will vote, the model predicts a criticality of 0.5—given that in a disjunctive set-up, $p(y \mid \neg x_i) = 1$ (the bill will definitely pass if A_i votes for it) and $p(y \mid \neg x_i) = 0.5$ (whether the bill will pass if A_i does not vote for it depends on the probability that A_j votes for it). More generally, the model predicts that each agent's criticality decreases with the number of agents in the group in disjunctive situations (assuming agent has at least some probability of succeeding). In conjunctive situations, an agent's criticality is maximal and not affected by the size of the group. Lagnado et al (2013) found that this model predicted participants' criticality judgments for a variety of causal structures that varied the number of group members and the way in which the individual contributions mapped onto the group outcome.

In line with Vincent (2011), we claim that outcome responsibility is a function of both pivotality and criticality. For example, when judging Katie's failure to vote, we incorporate both her pivotality for the outcome (if she had voted, Bobby would have won) and her perceived criticality (no one expected a situation where her vote would make the difference). Exactly how these two components are combined is a further question that we will discuss below. Note that our notion of criticality is just one factor amongst several that underpin the concept of role responsibility. In particular, it does not entirely capture the notion of a duty—because there might be a difference between an agent with a duty to do *X* and whether or not we actually expect the agent to do *X*. Statistical norms and moral norms can sometimes come apart (cf. Knobe and Hitchcock 2009). For example, parents have a duty to look after their children, but if we know that a particular parent is incompetent then we might expect them to fail in this duty. Indeed the notion of duty, independent of what we *actually* expect, is probably driving some of the tension in the voting example. Katie has a duty towards her husband, irrespective of how much we expected her vote to make a difference, because she is Bobby's wife. [7] In future research, we will tease apart to what extent judgments of criticality are driven by our actual expectations versus what we think the person ought to do given their duty.

10.3 EMPIRICAL TESTS OF THE DIFFERENCE-MAKING MODEL OF OUTCOME RESPONSIBILITY

An important benefit of specifying concrete models of "causal responsibility" in terms of *pivotality* and "role responsibility" in terms of *criticality* is that we can test the quantitative predictions that this framework makes about an individual's degree of "outcome responsibility". Because we restrict ourselves to these three different aspects of responsibility, we will from now on simply use the term *responsibility* when referring to "outcome responsibility". To reiterate, our difference-making model of responsibility predicts that responsibility judgments are a function of both pivotality and criticality. Generally, our account predicts that when two agents are equally critical, the agent whose contribution was closer to being pivotal will be held more responsible. Similarly, when two agents are equally pivotal, the agent who was perceived ex ante to have a more critical role will be held more responsible.

More concretely, we model an agent *A*'s responsibility for an outcome *O* in a given type of situation *S* as

$$\textit{responsibility}(A,O,S) = f(\textit{criticality}(A,S),\ \textit{pivotality}(A,O,S))$$

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

where *criticality* (*A,S*) denotes *A*'s criticality in setting *S* and *pivotality* (*A,O,S*) *A*'s pivotality with respect to the outcome *O* in setting *S*. For present purposes we assume that the function which translates criticality and pivotality into responsibility is a weighted linear combination of both factors.

In our experiments, we vary the setting *S* by manipulating the causal structure of the group task. Different settings differ in how individual performances are mapped onto the group outcome. We consider a space of settings that encompasses different group sizes as well as different causal structures that dictate the way in which the group members affect the joint outcome. Individual contributions either combine disjunctively (i.e., at least one of the agents needs to succeed in order for the group to be successful), conjunctively (i.e., all of the agents needs to succeed) or as mixtures of disjunction and conjunction (e.g., *A* needs to succeed and at least one out of *B, C*, and *D*). We vary outcomes *O* by creating different profiles of individual actions in a given setting *S*.

Figure 10.2 shows two different settings (a disjunctive task and a conjunctive task with two agents each) and four different outcome patterns. As outlined above, our model of criticality predicts that *A* is perceived to be more critical in the conjunctive structure than in the disjunctive structure. Whereas in conjunctive situations *A*'s contribution is necessary for the group to be successful ($C = 1$), in disjunctive structures it is not ($C = 1/2$). *A*'s pivotality varies with the outcome. *A* is fully pivotal for the loss in disjunctive situation (Figure 10.2a) and the win in the conjunctive situation (Figure 10.2d) where the group outcome is not overdetermined ($P = 1$). *A*'s pivotality is reduced for the win in the disjunctive structure (Figure 10.2b) and the loss in the conjunctive structure (Figure 10.2c) where the group outcome *is* overdetermined ($P = 1/2$). In each of overdetermined cases, the performance of B would need to be changed to make *A* pivotal (from a success to failure in Figure 10.2b and from failure to success in Figure 10.2c).

In one of our experiments (see Lagnado et al 2013), the participants' task was to evaluate the performance of contestants in a hypothetical game show. In the game show, each contestant played a game in which they had to click on a dot on the screen. Each time the dot was clicked it appeared on a new random location on the screen. In order to succeed in the game, a player had to reach a certain number of dot clicks within a given time period. After a period of practice trials, contestants were randomly assigned to group challenges that differed in terms of the number of people in the group and the underlying causal structure. Participants in the experiment were presented with the results of different team challenges and they were asked to judge to what extent player *A* was responsible for the group's outcome. Participants always saw the results of four different team challenges

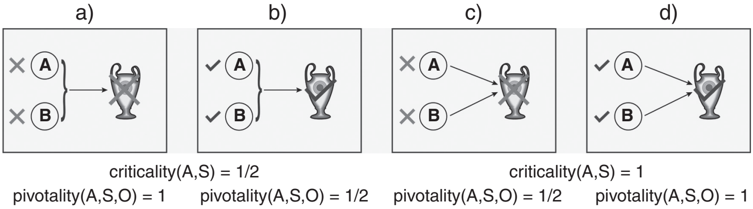

Figure 10.2 Two different settings (*disjunctive*: a) and b) vs. *conjunctive*: c) and d)) and two different outcomes (a win and loss for each situation) with predictions about agent *A*'s criticality in each of the two settings and *A*'s pivotality for each outcome. *Note*: ✕ indicates failure and ✓ indicates success. Curly braces indicate that individual contributions combine disjunctively.

simultaneously on the same screen. They indicated their responsibility judgments via sliders that were located underneath each of the four challenges.[8] Figure 10.3 shows participants' responsibility judgments for the four different situations discussed above.[9] Responsibility judgments increased with pivotality for the same level of criticality. Player *A* was held more responsible in situation 1 than in situation 2 and in situation 4 compared to situation 3. Similarly, *A* was held more responsible for situations in which pivotality was held constant but criticality increased as can be seen by comparing situation 1 with situation 4 and situation 2 with situation 3. Indeed, since participants weighed criticality more heavily than pivotality in these situations, *A* was held more responsible for the group's loss in the conjunctive structure (situation 3) than in the disjunctive structure (situation 1) despite the fact that *A* could have made the group win in situation 1 whereas the loss was overdetermined in situation 3. The pattern of responses shows that neither pivotality nor criticality itself can fully account for participants' responsibility judgments. Only a model that combines both notions can adequately capture participants' judgments. The model predictions shown in Figure 10.3 are based on a model that only uses one free parameter to determine the degree to which responsibility judgments are influenced by pivotality and criticality. This renders the model's predictions sensitive to the full set of situations that participants experienced over the course of the experiment. Overall, the

[8] You can access demos of the experiments here:

http://www.ucl.ac.uk/lagnado-lab/experiments/demos/causal_stucture_demos.html

[9] In the experiment, participants viewed situations 1 and 3 and situations 2 and 4 on separate screens but we combined the results here for convenience of comparison.

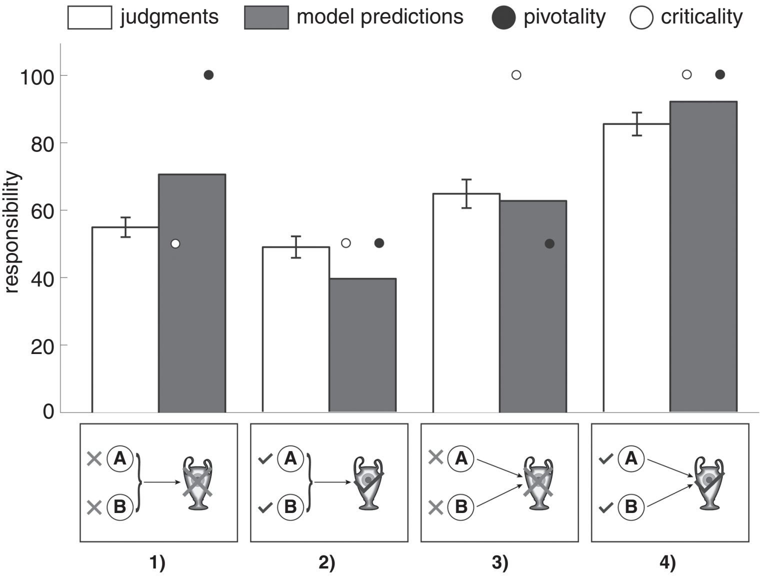

Figure 10.3 Mean responsibility judgments and model predictions for four different challenges that vary the causal structure (disjunctive vs. conjunctive) and the group outcome (loss vs. win). Error bars indicate ±1 standard error of the mean.

model assigns slightly more weight to pivotality than to criticality and hence the predicted responsibility for situation 1 is slightly higher than for situation 3. However, a model that relaxes the strong assumption that how much participants weigh criticality and pivotality is context-insensitive, correctly accounts for this pattern of results (see Lagnado et al 2013 for details).

One might be concerned that some of the comparisons above involved comparing *A*'s responsibility for losses versus wins. Maybe these differences can be explained in terms of an asymmetry between how people assign responsibility for negative and positive outcomes. To demonstrate that this is not the case, Figure 10.4 shows participants' judgments for two different sets of team challenges for which the group outcome was held constant. In Figure 10.4a, *A*'s criticality was held constant across the four challenges. In order for the group to succeed in this type of challenge, *A* and *B* had to succeed and at least one out of *C* and *D*. *A*'s pivotality was varied via manipulating the performance of the other players in the team. In situation 1, all players failed in their individual tasks. In this situation, a minimum of two changes is required to make *A* pivotal, namely changing *B* and either *C* or *D*.

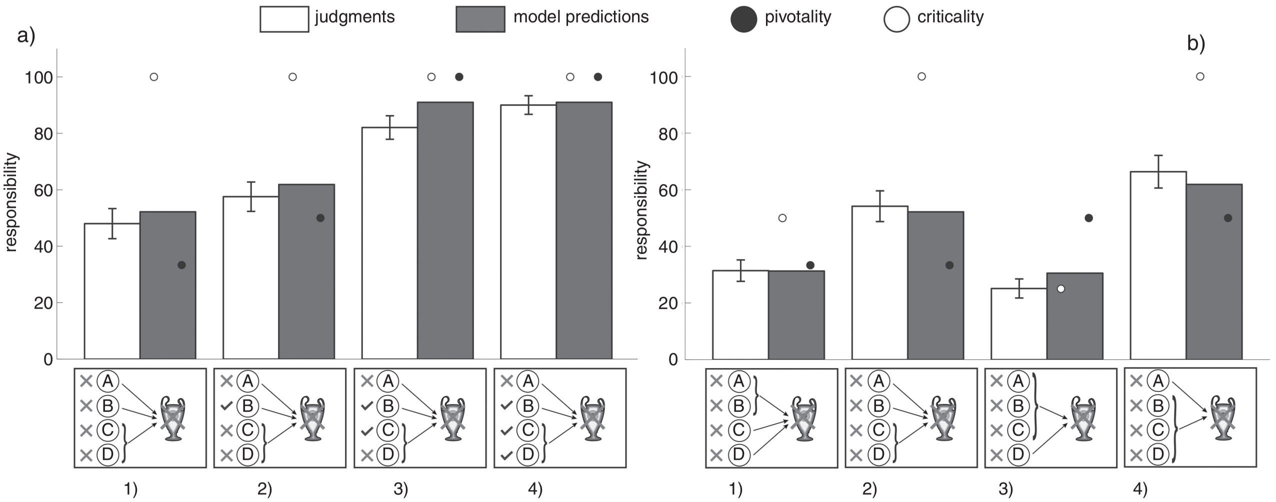

Figure 10.4 Responsibility judgments and model predictions for two different sets of situations. a) Responsibility increases with pivotality when criticality is held constant. b) Responsibility increases with criticality when pivotality is held constant (1 vs. 2 and 3 vs. 4). Error bars indicate ±1 standard error of the mean.

In situation 2, only one change is required to make *A* pivotal because *B* has already succeeded. In situations 3 and 4, *A* is pivotal. As can be seen, participants' responsibility judgments are clearly sensitive to how close *A* was to being pivotal. Responsibility judgments to *A* increased the closer he was to being pivotal and were particularly high when he was pivotal in the actual situation. This pattern of judgments demonstrates that participants are sensitive to the underlying causal structure of the task. Since the notion of pivotality is defined in terms of the minimal number of causal interventions that would be required to render a person pivotal, it can capture people's sensitivity to the structural dependencies (Chockler and Halpern 2004).

However, as outlined in Vincent's (2011) taxonomy, people's responsibility judgments are not only a function of a person's causal contribution to the outcome in the actual situation as captured by the pivotality model. In addition, participants are sensitive to a player's role in the group, that is, the extent to which a player's performance is critical for the team's success in a given challenge. Figure 10.4b shows a set of situations in which the structure of the team challenges was varied but the performance of all the players was held constant. Between situations 1 and 2, *A*'s pivotality was held constant. In both situations, a minimum of two changes would have been required to make *A* pivotal. However, *A*'s criticality is higher in situation 2 than in situation 1. In situation 2 the group has no chance of winning the challenge without *A*'s being successful. In contrast, in situation 1 *A*'s being successful is not necessary for the group to succeed. As predicted by our responsibility model, participants' judgments increased with criticality. *A* was held more responsible for the team's loss when he was perceived to have played a more critical role. The same pattern holds for situations 3 and 4. Again, while *A*'s pivotality was held constant, responsibility judgments increased with criticality.

In the experiment, participants saw nine sets of situations with four challenges each and thus made thirty-six judgments in total. With a single free parameter that was fit to determine how much participants' weighted criticality and pivotality when making their judgments, our responsibility model accounted for 81 percent of the variance in the data (see Lagnado et al 2013 for details). In fact, there was not a single case for which the model made the wrong qualitative prediction. That is, holding criticality constant, judgments always increased with pivotality. Holding pivotality constant, responsibility judgments increased with criticality. While neither criticality nor pivotality is individually sufficient, a weighted combination of both aspects captures participants' judgments very well. Moreover, we did not need to assume any asymmetry between responsibility attributions for positive and negative outcomes.

To sum up, participants' responsibility attributions to an individual in a group are not solely determined by the individual's performance but are systematically influenced by i) how critical the individual's contribution was perceived to be for the team's success ex ante and ii) how close the individual's contribution was to making a difference to the outcome ex post.

10.4 EXPLORING THE GENERALITY OF THE RESPONSIBILITY MODEL

The experiments reported in Lagnado et al (2013) explored people's responsibility judgments in an achievement context in which the team's outcome was a function of the team members' performances. We have also explored our responsibility model in different domains ranging from decisions about contributions in a public goods game to the functioning of different parts in mechanistic devices (see Figure 10.5). In these experiments, we exposed participants to the same causal structures and outcome patterns as described above but varied the framing of the context. In the public good framing (see Figure 10.5b), each of the group members decided whether or not to contribute their personal endowment to a public good (see Rapoport 1987). The causal structure of the task determined how many and which of the group members needed to contribute their endowment in order for the public good to be provided. If the public good is provided, each of the players in the team receives a bonus no matter whether or not they themselves contributed their endowment. If the public good is not provided, each contributed endowment is lost. This creates a well-known social dilemma situation (Hardin 1968; for a review, see Dawes 1980): each player has the incentive to maximize their individual payoff by deciding not to contribute their own endowment in the hope that the others will contribute theirs. This often leads to a suboptimal outcome on the group level: the public good may not be realized due to players deciding not to contribute. On the individual level, an individual who didn't contribute is always better off compared to other group members who contributed their endowment. In case the public good is provided, the non-contributing players keep their endowment *and* receive their equal share of the public good. In case the public good is not provided, non-contributing players at least kept their endowment and thus end up with more than contributing players whose efforts were wasted on a lost cause.

In the "public bad" framing, the game is structurally equivalent to the public good game (cf. Sonnemans, Schram, and Offerman 1998). However, rather than facing a decision of whether or not to contribute, individual players can either take money or refrain from taking money. Whether

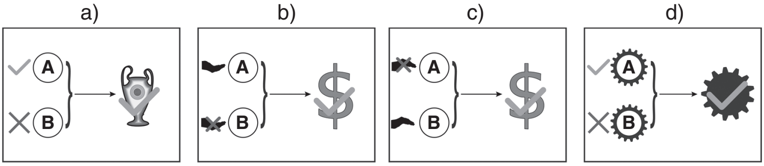

Figure 10.5 Four different framings of the task set-up. a) Team game in which the group outcome is a function of each individual's performance. b) Public good game in which the group outcome depends on whether players give money. c) Public bad game in which the group outcome depends on whether or not players take money. d) Non-social set-up in which a machine's functioning depends on its components.

the additional public good for each player is provided depends on which and how many players decide to take money individually. For example, in Figure 10.5c, the public good is provided as long as both players decide not to take the money.

In the non-social framing (see Figure 10.5d), we replaced the agents in the group with components of abstract mechanistic devices. For example, for disjunctive devices, only at least one of the components needed to operate in order for the device to work whereas for conjunctive structures, all of the components were required to operate. Participants' task in this experiment was to judge to what extent different components of the machines caused the machine to pass or fail a test trial.

Together these experiments constitute a comprehensive test of the generality of our responsibility model. Are considerations of criticality and pivotality limited to achievement domains or do they apply to a much broader class of situations in which causal factors combine to bring about a joint outcome? In particular, are criticality considerations restricted to social settings in which agents are aware of their role in the group or do they extend to the non-social domain? One might argue that the concept of "role responsibility" as endorsed in Vincent's (2011) taxonomy only makes sense for intentional agents who have the capacity to assume the normative forces that certain roles carry with them. For example, an individual whose contribution is necessary for the provision of the public good knows that he plays a more central role for the group outcome than an individual whose contribution combines disjunctively with other group members. We might thus expect that, by way of being aware of that role, more critical individuals are more likely to contribute their endowments than less critical group members. Similarly, in an achievement context, more critical players can be expected to try harder and exert more effort in order to ensure a good performance (Kerr and Bruun 1983).

If normative influences are restricted to the social domain, then we should expect no effects of criticality on judgments about the mechanistic devices. However, Hitchcock and Knobe (2009) have recently argued that norms do indeed extend to the non-social domain. Generally, they maintain that causal judgments are strongly influenced by normative considerations. Endorsing a difference-making conception of causation, Hitchcock and Knobe argue that norms influence causal judgments via affecting what counterfactuals people are likely to consider (cf. Kahneman and Miller 1986). This account provides an explanation for why people have a tendency to select abnormal events rather than normal events as causes for outcomes (Hilton and Slugoski 1986). While abnormal events trigger the consideration of more normal counterfactuals, the reverse doesn't hold: normal events tend not to make us consider what would have happened under abnormal counterfactual circumstances (Hart and Honoré 1959/1985; see Halpern and Hitchcock 2014, for a formal account of how normality influences causal judgments).

Hitchcock and Knobe acknowledge that there are multiple aspects that influence our conception of normality. They distinguish between different types of norms such as moral norms, statistical norms, and norms of proper functioning. Moral norms determine what a person ought to do in a given situation and only apply to social agents who can appreciate the normative force. Statistical norms apply to both the social and non-social domain such as the ratio of female to male students at MIT or the ratio of Mac to PC users in the academic community. Norms of proper functioning also apply to both the social and non-social domain. In the social domain, many fast food restaurants impose the norm of disposing one's garbage after having eaten. In the non-social domain, there is a norm that mechanical devices such as toasters or dishwashers are supposed to work. Note that these different norms can be in conflict with each other. While it might be the case that statistically most people just leave their garbage on the table, the norm of proper social functioning dictates that they shouldn't do so.

Hitchcock and Knobe demonstrated that the norm of proper functioning influences people's causal judgments in a non-social domain. Participants read a vignette about a machine with two wires whereby the black wire was designed to touch a battery and the red wire was supposed to remain in some other part of the machine. When a short circuit was caused due to both wires touching the battery, participants tended to agree with the statement that the red wire caused the machine to short circuit and to disagree that it was the black wire's fault.

In the non-social condition of our experiment, we did not directly manipulate the norm of proper functioning. However, it is plausible that participants' causal judgments were nevertheless sensitive to normative influences. Even for abstract mechanical devices, components whose functioning is necessary for the machine to operate are perceived more critical than components for which there exists a certain degree of redundancy by means of a disjunctive combination function (cf. Chockler and Halpern 2004). If the influence of criticality on responsibility does indeed extend to the non-social domain, we would expect very similar patterns of responsibility judgments for our four different experimental conditions. While criticality captures the moral norms prevalent in the context of individuals contributing to a joint outcome, criticality considerations might reflect the norm of proper functioning in the non-social domain.

Indeed, the patterns of results of the different experimental conditions, which were run between participants, were extremely closely aligned. The correlation between participants' judgments in the achievement condition and the public good condition, public bad condition, and mechanistic devices condition was $r = .92$, $r = .89$, and $r = .95$, respectively.

Figure 10.6 shows participants' judgments in all four conditions for two sets of situations. The pattern of judgments in the different conditions is virtually identical. Figure 10.6a shows that on average, participants' judgments were more strongly influenced by criticality than pivotality in situations in which these two considerations pointed in opposite directions. For example, in all four experimental conditions, *A* was judged to be more responsible when the causal structure was conjunctive (situations 2 and 4) than when it was disjunctive (situations 1 and 3). Despite the fact that *A* is pivotal in situation 3 and three changes would have been required to render *A* pivotal in situation 4, *A* was judged to be *more* responsible in 4 than in 3.

Similarly, in Figure 10.6b, the pattern of judgments between the four experimental conditions was almost identical. This set of judgments again highlights the importance of pivotality considerations in people's attributions (see also Zultan et al 2012). A simple diffusion of responsibility account (Darley and Latané 1968) predicts that *A*'s responsibility for the negative outcome increases the fewer other players (or components) also failed. In contrast to that, *A*'s responsibility actually *decreased* when *B* succeeded. For causal structures in which the contributions of *A* and *B* combine disjunctively, the success of *B* moves *A* further away from being pivotal. Hence, the pivotality model correctly predicts the reduction in responsibility from situation 1 to situation 2. In contrast, the success of *C* (and *D*) move *A* closer to pivotality. Accordingly, responsibility attributions in situations 3 and 4 increased compared to situation 1. Hence, despite the fact that participants' responsibility judgments were sometimes more strongly influenced by criticality than pivotality (cf. Figure 10.6a), pivotality considerations played an important role as well (as shown in Figure 10.6b). Indeed, trying to explain

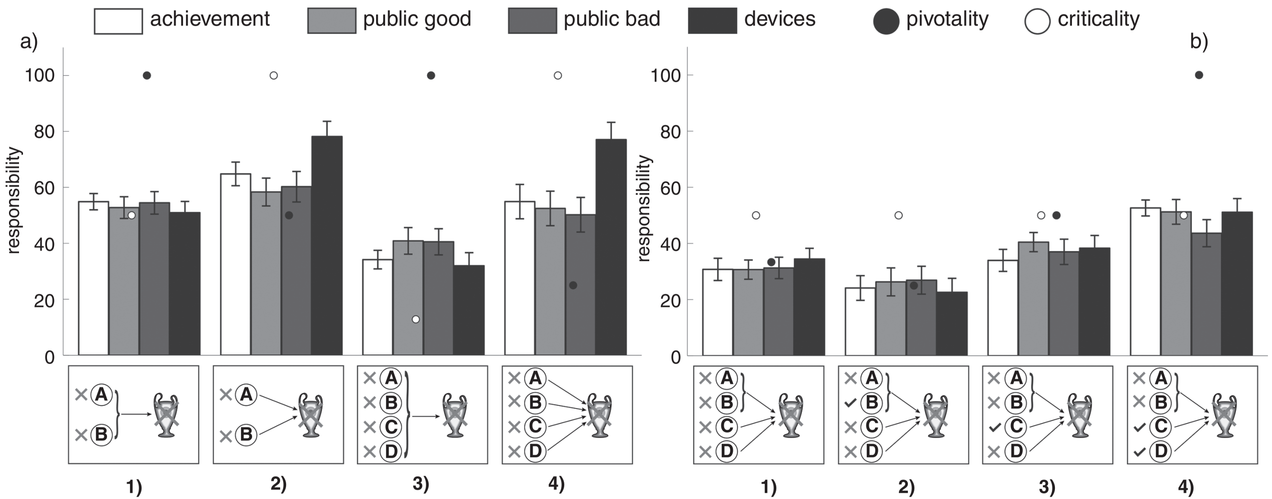

Figure 10.6 Mean responsibility judgments in achievement condition (*N* = 40 participants), public good condition (*N* = 30), and public bad condition (*N* = 30) and mean causality judgments in mechanistic devices condition (*N* = 38) for two different sets of outcome patterns. Error bars indicate ±1 standard error of the mean.

participants' responsibility judgments merely in terms of criticality or pivotality led to significantly lower correlations (see Lagnado et al 2013).

The very close similarity between participants' responsibility ratings in the public good and public bad condition demonstrates that the notion of criticality needs to be defined in terms of the action being critical for bringing about a *positive* or *good* outcome. Participants assigned more responsibility to players taking money in the public bad condition when it was sufficient that one player took money for the bonus to be cancelled. In this situation, it is necessary for each player *not to take* the money in order for the bonus to be provided. The same situation can be characterized as either disjunctive for a negative (or bad) outcome or conjunctive for a positive (or good) outcome. Our model of criticality is defined with respect to positive outcomes. Thus, we predict that an individual in a group is held more responsible for bringing about a bad outcome when each individual's action was sufficient for causing the bad outcome. In this situation, each person was maximally critical for bringing about the good outcome—only if all of individuals do the right thing will the bad outcome be avoided.

This prediction of our model resonates with findings by Spellman and Kincannon (2001) who compared participants' causal ratings for situations in which a negative outcome was brought about by multiple sufficient or multiple necessary causes. They found that participants rated two actors in a murder story who simultaneously shot a third person more causal when each of their individual shots was sufficient for bringing about the death compared to a situation in which both shots were necessary. They replicated their finding for a non-social scenario in which two fires either overdetermined a negative outcome or were both necessary for bringing it about.[10]

To sum up, the very close similarity between participants' responsibility judgments in the four experimental conditions demonstrates that criticality and pivotality are general principles that influence responsibility judgments across different domains. Despite the striking differences between the four experimental conditions on the surface level, participants' judgments were determined by the deeper underlying causal structure common to all four conditions. The results showed that criticality considerations are not only important in the social domain but have a significant influence on people's judgments in the non-social domain as well. Interpreted in Hitchcock and

[10] Note that in Spellman and Kincannon (2001), there is a potential confound between how sufficiency and necessity was manipulated and the inferences one can draw about the quality or strength of the candidate causes. For example, in the murder scenario, a shot that was individually sufficient to bring about a person's death was presumably a "better" shot compared to a shot that was merely necessary and only jointly sufficient with another shot. In our experiments, we avoid these confounds as each person's individual action is the same no matter whether they are in a conjunctive or disjunctive task.

Knobe's (2009) terms, the norm of proper functioning in the non-social domain exerted as strong an influence on people's judgments as the moral or performance norms in the social domain.

10.4.1 A Note on Individual Differences

So far, we have just presented participants' judgments in terms of average responses. However, there were quite substantial differences in how much participants weighted criticality and pivotality to determine their responsibility judgments. Rather than fitting a single weight parameter based on participants' mean judgments, we can also fit the weight parameter for each individual participant to see how strongly their judgments reflected considerations of criticality and pivotality. We use the following model to fit participants' responsibility judgments to player A:

$$\text{responsibility}(\text{A}) = w * \text{criticality}(\text{A}) + (1 - w) * \text{pivotality } (\text{A}),$$

where w is free to vary between 0 and 1.[11] Hence, a w of 0 indicates that a participant's judgments were best explained by a model that only uses pivotality as a predictor. A w of 1 indicates that a participant solely cared about criticality. Values in between indicate that participants' judgments reflected a consideration of both aspects. The average weights w for the achievement (w_a), public good (w_{pg}), public bad (w_{pb}), and mechanistic devices (w_m) conditions were w_a = 0.42 (SD = 0.20), w_{pg} = 0.34 (SD = 0.16), w_{pb} = 0.43 (SD = 0.19), and w_m = 0.55 (SD = 0.21). Hence, on average, participants were sensitive to both criticality and pivotality. Interestingly, criticality weighted most strongly in the mechanistic devices condition. However, across all four conditions, we observed the full range of weights with some participants only caring about pivotality (i.e., w = 0) and others only caring about criticality (i.e., w = 1).

What drives these inter-individual differences needs to be explored more closely in future research. Here, we will just note that it is relatively easy to shift one's intuitions about whether criticality or pivotality should matter more. On the one hand, it seems reasonable to say that only criticality should matter as a determinant for responsibility attribution. For example, in the achievement condition, a person in a conjunctive task knows that in order for the team to win, she has to succeed in her individual task. By

[11] In Lagnado et al (2013) we also fit a global intercept α. However, since we will evaluate model performance here merely in terms of correlation, we don't need to worry about the intercept here.

taking into account pivotality, however, our responsibility judgments to an individual for a group outcome are influenced by factors that were beyond her control. Compare two different ways in which the team could lose a conjunctive challenge with four players: (i) all players fail vs. (ii) all but *A* succeed. *A* is pivotal in (ii) but far from being pivotal in (i) (three changes would be required to render *A* pivotal). However, whether or not the other players succeeded in their task was clearly beyond *A*'s control. Assuming that we don't learn anything about the difficulty of the task through the player's performances, it seems unfair to give *A* more responsibility in (ii) than in (i). If we did hold *A* more responsible in (ii) than in (i), this would appear to be a clear instance of resultant moral luck (cf. Nagel 1979; Williams 1981; Sartorio 2012).

On the other hand, basing one's responsibility judgments solely on criticality doesn't take into consideration what actually happened. But part of what it means to be responsible for a particular outcome is that one's contribution played a significant causal role in bringing it about. The notion of pivotality captures this degree of causal involvement. In that sense, there is a close connection between holding someone responsible and providing an explanation for why a certain outcome came about (Lombrozo 2010). Thus, holding *A* responsible is a better explanation for why the team failed in (ii) than in (i).

The inter-individual differences found in our experiments could thus reflect differences in people's beliefs about what purpose responsibility judgments ought to serve. Participants who believe that assignments of responsibility ought to be free of factors that go beyond the agent's control might base their judgments mostly on criticality. Participants who believe that assignments of responsibility are supposed to reflect what actually happened and provide a good explanation for what causal factors were responsible for the outcome will take pivotality into consideration.

10.5 GENERAL DISCUSSION

We have argued that when people attribute responsibility to members of a group, their judgments depend on: (i) Criticality—how much they expect each member's action to be necessary for a positive outcome; and (ii) Pivotality—how close each member's action actually was to making a difference to the outcome.[12] These principles are based on core notions of

[12] We allow that responsibility judgments might depend on other factors too, so strictly speaking our claim is that responsibility judgments depend on *at least* these two components.

causality and counterfactual dependence (Pearl 2000), extended to deal with issues of overdetermination and degrees of responsibility (Chockler and Halpern 2004; Halpern and Pearl 2005). We have shown that criticality and pivotality are general principles that underlie people's responsibility judgments in both social and physical domains. Thus, people use both prospective and retrospective factors whether they are assigning responsibility to individual players in group games or to components in abstract mechanical devices. Finally, we have shown that people differ in how much weight they assign to criticality and pivotality when judging responsibility.

10.5.1 Capacity Responsibility

Our account fits neatly with a philosophical taxonomy of responsibility concepts proposed by Vincent (2011). In this paper we just focus on three key components in this taxonomy—mapping criticality onto (one component of) *role responsibility*, and pivotality onto *causal responsibility*. Together these components combine to yield judgments of *outcome responsibility*. However, we think that our psychological framework can be extended to include the other concepts in Vincent's taxonomy. For instance, in Vincent's taxonomy both role and causal responsibility are dependent on capacity responsibility (see Figure 10.1). Someone is responsible for an outcome only if they have the capacity to fulfill their role, and they actually exerted sufficient control over the outcome. One way to incorporate the notion of capacity responsibility into our formal framework is in terms of a generic causal model of someone's actions, including a probability distribution over their capabilities and actions (cf. Jara-Ettinger, Tenenbaum, and Schulz 2013). This prior model could express a person's skill in the achievement context, a person's generosity in a public goods game, or the reliability of a particular component in a machine.

Moreover, our framework makes straightforward predictions about how manipulating priors in this causal model will affect responsibility attributions. Recall that our notion of criticality is closely related to whether or not a person's contribution is necessary for the positive group outcome. Thus, our model predicts that in a conjunctive situation with two group members, manipulating the prior of one member will not influence the perceived criticality of the other. In a conjunctive situation, each person's contribution is necessary no matter how likely the others members are to succeed in their tasks. In a disjunctive situation, however, we do predict an influence of one person's prior chance of success on the other person's perceived criticality. The more likely one group member is going to succeed, the lower the chance that the other group member's contribution will be necessary. We are currently testing these predictions with new empirical studies, and have found support for our model.

Our framework also predicts that manipulating information about the prior capacities of team members will affect how close someone is to being pivotal. Contrast the following two situations. In situation 1, John plays with a highly skilled team member Mary. In situation 2, John plays with Brian, who has very low skill. In both situations, the team played a conjunctive team task and both players failed in their individual tasks. Given that both players failed in their task the group's loss was overdetermined. However, in situation 1, John was arguably closer to being pivotal than in situation 2. It is easier to imagine that Katy, who has high skill, would have succeeded in her task than it is to imagine a positive outcome for the low-skilled Brian. In future work, we will explore the predicted effects of manipulating capacity responsibility in terms of priors on responsibility attributions.

10.5.2 Counterfactual Replacements

At the core of our account of responsibility attribution is the notion of counterfactual difference-making, which we cashed out in terms of criticality and pivotality. Our experiments were designed so that the relevant counterfactuals were clear. For example, when someone failed to contribute to the public good, the obvious counterfactual is what would have happened if they had contributed. Expressed in terms of the formal accounts on which our model rests, counterfactuals are implemented via setting the value of a variable through an idealized intervention (Halpern and Pearl 2005; Pearl 2000; Woodward 2003). However, there are situations in which this operation does not fit our understanding of the causal structure of the situation. In many contexts, counterfactuals are better thought of in terms of replacements: what would someone else have done in the given situation (Falk and Szech 2013; Fincham and Jaspars 1983)? For example, consider the counterfactual "If Michael Jordan had not played for the Bulls then they would not have won the NBA championship". When we evaluate this counterfactual we do not just remove Jordan from the team and have them play four against five. Rather, we consider what other player would have likely taken Jordan's position and simulate whether the Bulls would have won with that replacement player.[13] Similarly, in legal cases of negligence, counterfactually replacing a defendant with the reasonable man often serves as an evaluative

[13] In baseball the wins above replacement statistic (WAR) captures the number of wins that a particular player contributes over and above a hypothetical replacement player.

yardstick (cf. Schaffer 2005, 2010). A person is only considered guilty of negligence if it can be argued that a reasonable man would have acted in such a way that the negative outcome would have been avoided. In future work, we will extend our account to such situations in which counterfactuals are better conceived of in terms of replacements rather than mere presences or absences of actions (see Gerstenberg et al 2014).

Finally, let us return to Bobby, Katie, and the Kentucky council elections. We are now in a better position to explain the conflicting intuitions we have about Katie's responsibility for the tied poll. In the actual situation her vote was pivotal for the outcome; but she was not expected to be critical, and she had good mitigating reasons for missing the vote. But our analysis misses a final component. We don't know whether Bobby won or lost in the coin toss that decided the election. Presumably this should not matter to Katie's responsibility, but we bet it does![14]

References

Alicke, M. D. (2000). "Culpable Control and the Psychology of Blame." *Psychological Bulletin* 126(4): 556–74.

Braham, M. and van Hees, M. (2009). "Degrees of Causation." *Erkenntnis* 71(3): 323–44.

Braham, M. and van Hees, M. (2013). "An Anatomy of Moral Responsibility." *Mind* 121(483): 601–34.

Cane, P. (2002). *Responsibility in Law and Morality* (Oxford: Hart Publishing).

Chockler, H. and Halpern, J. Y. (2004). "Responsibility and Blame: A Structural-model Approach." *Journal of Artificial Intelligence Research* 22(1): 93–115.

Darley, J. M. & Latane, B. (1968). Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology*, 8(4): 377-383.

Dawes, R. M. (1980). "Social Dilemmas." *Annual Review of Psychology* 31(1): 169–93.

Falk, A. and Szech, N. (2013). "Morals and Markets." *Science* 340(6133): 707–11.

Felsenthal, D. S. and Machover, M. (2004). "A Priori Voting Power: What is it All About?" *Political Studies Review* 2(1): 1–23.

Fincham, F. D. and Jaspars, J. M. (1983). "A Subjective Probability Approach to Responsibility Attribution." *British Journal of Social Psychology* 22(2): 145–61.

Gerstenberg, T. and Lagnado, D. A. (2010). "Spreading the Blame: The Allocation of Responsibility Amongst Multiple Agents." *Cognition* 115(1): 166–71.

[14] In fact Bobby lost the coin toss. However, in a further twist to the story, his competitor Olivia decided not to take the seat. We still don't know whether Bobby was given the seat, or had to fight another election.

Gerstenberg, T. and Lagnado, D. A. (2012). "When Contributions Make a Difference: Explaining Order Effects in Responsibility Attributions." *Psychonomic Bulletin & Review* 19(4): 729–36.

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., and Tenenbaum, J. B. (2014). "Wins Above Replacement: Responsibility Attributions as Counterfactual Replacements." In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, (ed. M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth), 2263–68 (Austin, TX, 2014. Cognitive Science Society).

Guglielmo, S. and Malle, B. F. (2010). "Enough Skill to Kill: Intentionality Judgments and the Moral Valence of Action." *Cognition* 117(2): 139–50.

Halpern, J. Y. and Hitchcock, C. (2014). "Graded Causation and Defaults." *The British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axt050

Halpern, J. Y. and Pearl, J. (2005). "Causes and Explanations: A Structural-Model Approach. Part I: Causes." *The British Journal for the Philosophy of Science* 56(4): 843–87.

Hardin, G. (1968). "The Tragedy of the Commons." *Science* 126: 1243–8.

Hart, H. L. A. (1968/2008). *Punishment and Responsibility* (Oxford: Oxford University Press).

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

Hilton, D. J. and Slugoski, B. R. (1986). "Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model." *Psychological Review* 93(1): 75–88.

Hitchcock, C. and Knobe, J. (2009). "Cause and Norm." *Journal of Philosophy* 11: 587–612.

Jara-Ettinger, J., Tenenbaum, J. B., and Schulz, L. E. (2013). "Not So Innocent: Reasoning About Costs, Competence, and Culpability in Very Early Childhood." In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, (eds. M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth), 663–68 (Austin, TX: Cognitive Science Society).

Kahneman, D. and Miller, D. T. (1986). "Norm Theory: Comparing Reality to its Alternatives." *Psychological Review* 93(2): 136–53.

Kerr, N. L. (1989). "Illusions of Efficacy: The Effects of Group Size on Perceived Efficacy in Social Dilemmas." *Journal of Experimental Social Psychology* 25(4): 287–313.

Kerr, N. L. and Bruun, S. E. (1983). "Dispensability of Member Effort and Group Motivation Losses: Free-rider Effects." *Journal of Personality and Social Psychology* 44(1): 78–94.

Knobe, J. (2010). "Person as Scientist, Person as Moralist." *Behavioral and Brain Sciences* 33(4): 315–65.

Hitchcock, C. and Knobe, J. (2009). "Cause and Norm." *Journal of Philosophy* 11: 587–612.

Lagnado, D. A. and Channon, S. (2008). "Judgments of Cause and Blame: The Effects of Intentionality and Foreseeability." *Cognition* 108(3): 754–70.

Lagnado, D. A., Gerstenberg, T., and Zultan, R. (2013). "Causal Responsibility and Counterfactuals." *Cognitive Science* 47: 1036–73.

Lewis, D. (1979). “Counterfactual Dependence and Time’s Arrow.” *Noûs* 13(4): 455–76.

Lewis, D. (1986). “Causal Explanation.” *Philosophical Papers* 2: 214–40.

Livengood, J. (2011). “Actual Causation and Simple Voting Scenarios.” *Noûs,* 47(2): 1–33.

Lombrozo, T. (2010). “Causal-explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions.” *Cognitive Psychology* 61(4): 303–32.

Malle, B. F. (1999). “How People Explain Behavior: A New Theoretical Framework.” *Personality and Social Psychology Review* 3(1): 23–48.

Nagel, T. (1979). *Mortal Questions* (Cambridge: Cambridge University Press).

Pearl, J. (1999). “Probabilities of Causation: Three Counterfactual Interpretations and Their Identification.” *Synthese* 121(1–2): 93–149.

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

Rapoport, A. (1987). “Research Paradigms and Expected Utility Models for the Provision of Step-level Public Goods.” *Psychological Review* 94(1): 74–83.

Sartorio, C. (2012). “Two Wrongs Do Not Make a Right: Responsibility and Overdetermination.” *Legal Theory* 18(4): 473–90.

Schaffer, J. (2005). “Contrastive Causation.” *The Philosophical Review* 114(3): 327–58.

Schaffer, J. (2010). “Contrastive Causation in the Law.” *Legal Theory* 16(04): 259–97.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., and Doherty, K. (1994). “The Triangle Model of Responsibility.” *Psychological Review* 101(4): 632–52.

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness* (New York: Springer-Verlag).

Sonnemans, J., Schram, A., and Offerman, T. (1998). “Public Good Provision and Public Bad Prevention: The Effect of Framing.” *Journal of Economic Behavior & Organization* 34(1): 143–61.

Spellman, B. A. & Kincannon, A. (2001). The Relation Between Counterfactual (“But For”) and Causal Reasoning: Experimental Findings and Implications for Jurors' Decisions. *Law and Contemporary Problems*, 64(4): 241-264.

Vincent, N. A. (2011). “A Structured Taxonomy of Responsibility Concepts.” In *Moral Responsibility: Beyond Free Will and Determinism*, ed. N. A. Vincent, I. van de Poel, and J. van den Hoven, 15–35 (Dordrecht: Springer).

Williams, B. (1981). *Moral Luck: Philosophical Papers, 1973–1980* (Cambridge: Cambridge University Press).

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation* (Oxford: Oxford University Press).

Woodward, J. (2006). “Sensitive and Insensitive Causation.” *The Philosophical Review* 115(1): 1–50.

Zultan, R., Gerstenberg, T., and Lagnado, D. A. (2012). “Finding Fault: Counterfactuals and Causality in Group Attributions.” *Cognition* 125(3): 429–40.