Cognition 177 (2018) 122–141

Contents lists available at ScienceDirect

Cognition

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

Original Articles

# Lucky or clever? From expectations to responsibility judgments

Tobias Gerstenberg[a,*], Tomer D. Ullman[a], Jonas Nagel[b], Max Kleiman-Weiner[a], David A. Lagnado[c], Joshua B. Tenenbaum[a]

[a] Massachusetts Institute of Technology, United States
[b] Göttingen University, Germany
[c] University College London, United Kingdom

ARTICLE INFO

*Keywords:*
Responsibility
Causality
Expectations
Counterfactuals
Pivotality

ABSTRACT

How do people hold others responsible for the consequences of their actions? We propose a computational model that attributes responsibility as a function of what the observed action reveals about the person, and the causal role that the person's action played in bringing about the outcome. The model first infers what type of person someone is from having observed their action. It then compares a prior expectation of how a person would behave with a posterior expectation after having observed the person's action. The model predicts that a person is blamed for negative outcomes to the extent that the posterior expectation is lower than the prior, and credited for positive outcomes if the posterior is greater than the prior. We model the causal role of a person's action by using a counterfactual model that considers how close the action was to having been pivotal for the outcome. The model captures participants' responsibility judgments to a high degree of quantitative accuracy across three experiments that cover a range of different situations. It also solves an existing puzzle in the literature on the relationship between action expectations and responsibility judgments. Whether an unexpected action yields more or less credit depends on whether the action was diagnostic for good or bad future performance.

## 1. Introduction

In the quarter final of the 2006 FIFA World Cup, the Germany versus Argentina match came down to penalty shots. Unbeknownst to the Argentinian team, the German goalkeeper, Jens Lehmann, was handed a piece of paper that indicated where each of the Argentinian players was likely to shoot. Lehmann ended up saving two penalties, and the German team won the game. Clearly, Lehmann deserves credit for the team's win. But how much, and on what grounds?

Let us suppose that the following took place: Lehman was told that the first shooter often aims the ball at the left corner. Lehmann jumped to this corner and saved the ball. For the second shooter, Lehmann was told again to expect a shot in the left corner. However, this time Lehmann jumped in the opposite corner, and again saved the shot, even though his opponent kicked the ball in the unexpected direction. Would you give Lehmann more credit for the first, or the second save? And suppose Lehmann had failed to save both shots. Would you have blamed him more for failing to save the shot that went in the expected direction, or the unexpected one?

In this paper, we investigate how people hold others responsible for their actions. Most existing accounts predict that unexpected actions elicit greater attributions of responsibility than expected actions (Brewer, 1977; Fincham & Jaspars, 1983; Malle, Guglielmo, & Monroe, 2014; Petrocelli, Percy, Sherman, & Tormala, 2011), and, more generally, that unexpected events are more likely to be cited as the cause of an outcome (Halpern & Hitchcock, 2015; Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). However, recently Johnson and Rips (2015) reported a series of experiments in which participants held agents *more* responsible when positive outcomes resulted from *expected* actions. In their experiments, an agent faced a choice between multiple options that differed in their probability of bringing about a positive outcome. They found that participants held the agent more responsible for a positive outcome when the agent chose an option that was better than any of the alternatives, and less responsible when the agent chose an inferior option.

Together, these findings present a puzzle: When do we assign more responsibility for unexpected actions (as most theories predict), and when do we assign less responsibility? We present a computational model that solves this puzzle. The model relies on two processes: the first process is a *dispositional inference* that captures what an action reveals about a person. Specifically, we propose that a person will be credited (or blamed) to the degree that their action reveals they are the sort of person who will get things right (or wrong) in the future. To go

* Corresponding author at: MIT Building 46-4053, 77 Massachusetts Avenue, Cambridge, MA 02139, United States.
*E-mail address:* tger@mit.edu (T. Gerstenberg).

https://doi.org/10.1016/j.cognition.2018.03.019
Received 4 December 2017; Received in revised form 21 March 2018; Accepted 23 March 2018

back to our opening example, Lehmann will be credited more for saving the unexpected shot because we infer by that action that Lehmann is a skilled goalie. However, if Lehmann chose an unexpected action in a pure game of chance, this would be diagnostic of poor future performance and so our model predicts little credit in this case.

The second process is a *causal attribution* of the role that a person's action played in bringing about the outcome. People are held more responsible to the extent that their action was pivotal in bringing about the outcome.

Our formal framework for explaining responsibility judgments draws on a rich literature in attribution theory, as well as recent work on modeling causal judgments. We briefly review each of these strands of research, focusing on the aspects that are most relevant for our framework. We then present our computational model in detail, and subsequently test the fine-grained predictions of our model in three experiments that vary action expectations, and the extent to which a person's action made a difference to the outcome. We discuss how our model relates to previous work, and how different comparison standards may affect judgments of responsibility. We conclude by highlighting future avenues of research motivated by the model and results presented here.

### 1.1. Dispositional inference: from actions to persons

Early attribution theorists proposed Bayesian inference as a normative framework for making diagnostic inferences about a person from observing their actions (Ajzen, 1971; Ajzen & Fishbein, 1975, 1978; Fischhoff & Beyth-Marom, 1983; Fischhoff & Lichtenstein, 1978; Morris & Larrick, 1995; Trope, 1974; Trope & Burnstein, 1975). For the Bayesian framework to support inferences from observed variables (behavior) to latent variables (mental states), it requires a model that captures how the latent and observed variables relate. Essentially, in order to assign responsibility to others, we need a model of decision-making that expresses how we believe people make choices based on their mental states. A key assumption for making sense of other people's behavior in this way is the *principle of rational action* (Dennett, 1987). It states that a person chooses an action that is expected to achieve a desired goal in the most efficient way, subject to the person's beliefs and abilities (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Gilbert, 1998; Goodman et al., 2006; Heider, 1958; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Malle & Knobe, 1997; Pantelis et al., 2014; Wellman & Bartsch, 1988).

To the extent that a person acts in line with our expectations, we do not learn much beyond what we already know, and need not update our beliefs. However, when a person's action violates our expectation then we need to make sense of their behavior, either finding situational factors that influenced their actions, or updating our beliefs about who they really are (Duff, 1993; Frieze & Weiner, 1971; Koster-Hale & Saxe, 2013; Uhlmann, Pizarro, & Diermeier, 2015; Weiner, 1985; Weiner, Heckhausen, Meyer, & Cook, 1972). Did the agent have some special skill and behave optimally in light of having this ability, or did the agent lack the relevant skill, and the positive outcome was the lucky result of poor decision-making (cf. Morse, 2003; Rachlinski, 2002–2003; Sinnott-Armstrong & Levy, 2011; van Inwagen, 1978)?

Our model predicts that attributions of responsibility are closely linked to our expectations. We credit a person if their action indicates that they are better than a comparison standard. Conversely, we blame a person if their action reveals that they are worse than we expected.

### 1.2. Causal attribution: from actions to outcomes

Research on causal attribution has identified a host of factors that influence people's causal judgments (Einhorn & Hogarth, 1986; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014, 2015; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Sloman, 2005; Sloman & Lagnado, 2015; White, 2014; Wolff, 2007). In order to be held responsible for an outcome, a person's action must be causally connected to the outcome. We predict that the extent to which a person is blamed or credited for an outcome depends on the perceived causal influence that their action had on the outcome. To determine what role an action played in bringing about the outcome, we need a causal model of the situation that captures how the action of interest and other candidate causes affected the outcome. Here, we take inspiration from work in philosophy (Woodward, 2003; Yablo, 2002) and computer science (Halpern & Pearl, 2005; Pearl, 2000) that models causal relationships in terms of counterfactual contrasts over a causal model of the situation.

Within this framework, a variable qualifies as a cause of an outcome if the outcome would have been different had the variable taken on a different value (Lewis, 1973). However, this test of counterfactual dependence runs into problems when outcomes are overdetermined by multiple, individually sufficient causes. For example, in elections, the outcome would often not have been any different if a single voter had changed her mind. However, we still want to say that each voter has some degree of responsibility for the outcome. Halpern and Pearl (2005) proposed a structural model of causal attribution that handles this and other problems by replacing the simple counterfactual test of causation with a test of counterfactual dependence under contingency. A variable can qualify as a cause even when it did not make a difference in the actual situation, as long as there was a possible situation that could have arisen, in which the event would have made a difference.[1] Chockler and Halpern (2004) have proposed that the closer a person's action was to having been pivotal, the greater their causal responsibility for the outcome. Prior research has shown that pivotality is an important factor in how people attribute responsibility (Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Lagnado, 2010, 2012, 2014; Lagnado & Gerstenberg, 2015; Lagnado, Gerstenberg, & Zultan, 2013; Wells & Gavanski, 1989; Zultan, Gerstenberg, & Lagnado, 2012).

In this paper, we will look at relatively simple settings in which a decision-maker chooses between two options. In some of the situations, their actions turn out to be pivotal – the outcome would have been different if they had acted differently – whereas in other situations, their actions aren't pivotal – the outcome would have been the same even if they had chosen the other option. We predict that a person is viewed as more responsible for an outcome when her action was pivotal.

## 2. Computational model

Our model assigns a degree of responsibility to people making decisions under uncertainty that result in positive or negative outcomes (cf. Botti & McGill, 2006; Leonhardt, Keller, & Pechmann, 2011; Nordbye & Teigen, 2014; Parkinson & Byrne, 2017). Our model has two components: (i) a dispositional inference from the person's action to their character, which affects the model's expectation about the person's future behavior, and (ii) a causal inference about the relationship between the person's action and the outcome. We will discuss each component in turn.

### 2.1. Dispositional inference and expectation change

The first component of our model formalizes how we update our expectations about a person's future performance after having observed the person's action, and the outcome that resulted. This inference involves two steps. The first step is to update our belief about the type of person the decision maker is. The second step is to transform this new belief into an updated expectation about how well the person will do in the future. We will discuss each step in turn.

[1] Much of the work goes into specifying which contingencies are allowed when checking for whether a counterfactual dependence holds between the candidate cause and effect (cf. Livengood, 2011). In this paper, we will focus on settings in which these difficulties do not arise.