Cognition 115 (2010) 166–171

Contents lists available at ScienceDirect

Cognition

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

Spreading the blame: The allocation of responsibility amongst multiple agents

Tobias Gerstenberg, David A. Lagnado *

Department of Cognitive, Perceptual and Brain Sciences, University College London, UK

ARTICLE INFO

Article history:
Received 3 March 2009
Revised 27 November 2009
Accepted 15 December 2009

Keywords:
Collective responsibility
Attribution
Counterfactual reasoning
Causal models
Experimental games

ABSTRACT

How do people assign responsibility to individuals in a group context? Participants played a repeated trial experimental game with three computer players, in which they counted triangles presented in complex diagrams. Three between-subject conditions differed in how the group outcome was computed from the individual players' answers. After each round, participants assigned responsibility for the outcome to each player. The results showed that participants' assignments varied between conditions, and were sensitive to the function that translated individual contributions into the group outcome. The predictions of different cognitive models of attribution were tested, and the Structural Model (Chockler & Halpern, 2004) predicted the data best.

1. Introduction

Who do you blame when your soccer team loses in the final minutes of the game? Your goalkeeper for conceding a soft goal? Your strikers for missing several good opportunities? The whole team for playing below par? Attributing responsibility is a commonplace activity that has attracted widespread research in psychology (Alicke, 2000; Hilton, McClure, & Slugoski, 2005; Lagnado & Channon, 2008; Shaver, 1985), philosophy (Hart & Honoré, 1959/1985) and law (Moore, 2009). The typical focus is on how people attribute blame to individual agents; however, in many situations it is a group of individuals that collectively determines the outcome, and responsibility must be distributed amongst the group. Team sports provide a paradigm example, but issues of group responsibility arise in many areas, including business, medicine, and law. The allocation of credit or blame in such contexts can be problematic, because it is often hard to isolate the separate contributions that each individual made. Consider the traditional sport tug-of-war. Individual power, stamina and technique, as well as coordination within the team, are important determinants of success. How much responsibility should each player bear for the team's win or loss? Should players be held responsible according to their individual contribution? Or perhaps according to whether their contribution made a critical difference to the team's result?

This difficulty in allocating responsibility is compounded by the fact that causes can combine in various different ways to bring about an outcome. Thus, there are several different functions that can translate the actions of each individual member into the group outcome (Steiner, 1972). The nature of this combination function can depend on the rules of the game, the relevant physical or social laws, or practical aspects of the situation (Waldmann, 2007). Three common functions are *addition*, *conjunction* or *disjunction*. In the additive case, each cause contributes something to the final outcome. Tug-of-war is a prototypical example, where each member contributes to the team's overall success. In the conjunctive case, all causes need to surpass a certain threshold. The final outcome is determined by the weakest member of the team.

* Corresponding author. Address: Department of Cognitive, Perceptual and Brain Sciences, University College London, Gower Street, London WC1E 6BT, UK.
E-mail address: d.lagnado@ucl.ac.uk (D.A. Lagnado).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.12.011

For example, a climbing team is only as fast as its slowest member. In the disjunctive case, it only takes one cause to bring about the outcome. The team is as good as its best member. One example is a team quiz, where a correct answer from just one member is sufficient for the team to win the point. How sensitive are people's responsibility judgments to these different combination functions?

Despite the importance of these questions for attribution research, they have received little attention in the psychological literature. This paper introduces a novel experimental set-up to examine how people distribute credit or blame amongst team members, and whether they are sensitive to the different ways that members can combine to produce an outcome. It also evaluates how well these judgments are captured by three competing models of responsibility attribution.

2. Models of attribution

People's judgments of credit or blame are presumed to be based on prior causal attributions, but modulated by various factors such as intention, foresight, mitigating circumstances or potential justifications (Lagnado & Channon, 2008; Shaver, 1985). This paper focuses on the causal attribution stage and investigates to what extent the allocation of responsibility is influenced by people's knowledge of the causal function that translates individual actions into a group outcome. We test three models of responsibility allocation. All models involve two steps: (1) determine which of the agents in the group are causes of the collective outcome and (2) distribute responsibility amongst the identified causes.

2.1. The Matching Model

The Matching Model sees each agent within the group as a cause of the collective outcome. It predicts that people assign responsibility in direct proportion to the individual contribution of each agent. Applied to the tug-of-war example, each player's pulling power might serve as a proxy for responsibility allocation. However, this strategy becomes problematic when the individual contributions are hard to estimate. Furthermore, there is a strong intuition that a player should only be held responsible if his action had the potential to make a difference to the team's result. If the team would have won irrespective of what the player did, we are hesitant to attribute any responsibility to him. Despite these limitations, the Matching Model serves as a useful benchmark against which to compare other models.

2.2. The Counterfactual Model

The Counterfactual Model incorporates the intuition that the potential of making a difference is a precondition for being held responsible. In the first step, it employs the counterfactual theory of causality (Lewis, 1973) to decide which of the agents caused the collective outcome. On this theory two conditions must be met to qualify an event A as the cause of another event B: A and B must both have occurred, and if A had not occurred then B would not have occurred. In the second step, the model assigns full responsibility to each agent identified as a cause. Several problems with counterfactual theories have been pointed out, the major one being that of causal overdetermination (Collins, Hall, & Paul, 2004). Consider a variation of the tug-of-war example where team A, consisting of four players, beats team B, consisting of only three players. Suppose that team A would have won even if only three of their players had engaged in the game. In this situation the Counterfactual Model would assign a responsibility of 0 for the win to each player in team A. None of the players would be identified as a cause because each player's individual action did not make a critical difference to the team's outcome.

2.3. The Structural Model

Chockler and Halpern (2004) have developed a model of responsibility attribution that accommodates cases of overdetermination. Their model is cast in the framework of causal models to capture the counterfactual dependencies between sets of events (Halpern & Pearl, 2005). In the first step, their theory offers a relaxed criterion of counterfactual dependence. A is a cause of B if and only if there is a *possible situation* under which B counterfactually depends on A. In the second step, the degree of responsibility of an individual cause a_1 (from a set of causes a_i) for an effect b is determined by the equation: $\mathrm{Resp}(a_1) = 1/(N+1)$. N denotes the *minimal number of changes* that must be made to the original situation to obtain a modified situation where b counterfactually depends on a_1. Applied to the tug-of-war example, this means that each of the four players in team A receives a responsibility of 1/2 for their win. Only if one player had dropped out (i.e. one change from the actual situation) would each remaining player's action have been critical for the outcome of the contest.

This paper aims to test these three models of responsibility attribution. Although the relation between causal and counterfactual judgments has been extensively investigated (Kahneman & Miller, 1986; Roese, 1997), the Structural Model has not yet been subjected to empirical test. We aim to discover how participants attribute responsibility to individual persons for outcomes they have brought about collectively and whether differences in the underlying causal structure influence participants' responsibility ratings.

3. Experiment

To investigate how people attribute responsibility in group contexts we developed the Triangle Game. This is an interactive computer game. The participant's task was to count triangles presented in complex diagrams for a brief period of time. Participants were instructed that they were not playing the game individually but in a group together with three computer players. Whether a particular round in the game was won or lost depended on the accuracy of each player in the group.

Each round of the game consisted of two consecutive steps. In the first step, participants were shown the