Cognition 167 (2017) 201–211

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Inferences about moral character moderate the impact of consequences on blame and praise

Jenifer Z. Siegel [a,b,1], Molly J. Crockett [a,b,c,*,1], Raymond J. Dolan [b,d]

[a] Department of Experimental Psychology, University of Oxford, United Kingdom
[b] Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom
[c] Department of Psychology, Yale University, USA
[d] Max Planck Centre for Computational Psychiatry and Ageing, University College London, United Kingdom

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 15 April 2017
Accepted 1 May 2017
Available online 17 May 2017

Keywords:
Morality
Moral judgment
Omission bias
Blame
Praise
Character

ABSTRACT

Moral psychology research has highlighted several factors critical for evaluating the morality of another's choice, including the detection of norm-violating outcomes, the extent to which an agent caused an outcome, and the extent to which the agent intended good or bad consequences, as inferred from observing their decisions. However, person-centered accounts of moral judgment suggest that a motivation to infer the moral character of others can itself impact on an evaluation of their choices. Building on this person-centered account, we examine whether inferences about agents' moral character shape the sensitivity of moral judgments to the *consequences* of agents' choices, and agents' role in the *causation* of those consequences. Participants observed and judged sequences of decisions made by agents who were either bad or good, where each decision entailed a trade-off between personal profit and pain for an anonymous victim. Across trials we manipulated the magnitude of profit and pain resulting from the agent's decision (consequences), and whether the outcome was caused via action or inaction (causation). Consistent with previous findings, we found that moral judgments were sensitive to consequences and causation. Furthermore, we show that the inferred character of an agent moderated the extent to which people were sensitive to consequences in their moral judgments. Specifically, participants were more sensitive to the magnitude of consequences in judgments of bad agents' choices relative to good agents' choices. We discuss and interpret these findings within a theoretical framework that views moral judgment as a dynamic process at the intersection of attention and social cognition.

1. Introduction

A longstanding question in moral psychology is a concern with the criteria people use when assigning blame to others' actions. Theories of blame highlight several critical factors in determining an agent's blameworthiness for a bad outcome (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Heider, 1958; Malle, Guglielmo, & Monroe, 2014; Shaver, 1985; Weiner, 1995). The first step is detecting some bad outcome that violates a social norm. Next comes an evaluation of whether the agent caused the outcome, followed by an assessment of whether the agent intended the outcome. People are considered more blameworthy for harmful actions than equally harmful omissions (Baron, 1994; Baron & Ritov, 1994; Cushman, Murray, Gordon-McKeon, Wharton, & Greene, 2012; Spranca, Minsk, & Baron, 1991) because the former are viewed as more causal than the latter (Cushman & Young, 2011). Moreover, people are blamed more for intentional compared to unintentional (i.e., accidental) harms (Karlovac & Darley, 1988; Shultz & Wright, 1985; Shultz, Wright, & Schleifer, 1986). Causation and malintent are each alone sufficient to ascribe judgments of blame for bad outcomes. In the case of accidental harms, people blame agents for bad outcomes that they caused but did not intend (Ahram et al., 2015; Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Martin & Cushman, 2015; Oswald, Orth, Aeberhard, & Schneider, 2005). There is also evidence that people blame agents for bad outcomes that they intend or desire but do not cause (Cushman, 2008; Inbar, Pizarro, & Cushman, 2012).

Other work has highlighted how inferences about moral character impact the assignment of blame and praise. For example, judges and juries frequently condemn repeat offenders to harsher penalties than first-time offenders for equivalent crimes (Roberts, 1997), and conviction rates are correlated with jurors' knowledge

* Corresponding author at: Department of Psychology, Yale University, USA.
E-mail address: mj.crockett@yale.edu (M.J. Crockett).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.cognition.2017.05.004
0010-0277/

of a defendant's previous crimes (T. Eisenberg & Hans, 2009), particularly when past crimes are similar to a current offence (Alicke et al., 2015; Wissler & Saks, 1985). In the laboratory, people assign more blame to dislikable agents than likable agents (Alicke & Zell, 2009; Kliemann, Young, Scholz, & Saxe, 2008; Nadler, 2012). These observations are consistent with a *person-centered* approach to moral judgment, which posits that evaluations of a person's moral character bleed into evaluations of that person's actions (Uhlmann, Pizarro, & Diermeier, 2015). In other words, despite being instructed to assess whether an *act* is blameworthy, people may instead evaluate whether the *person* is blameworthy.

In line with this view, there is evidence that evaluations of causation and intent are themselves sensitive to inferences about an agent's character (Alicke, 1992; Alicke, 2000; Knobe, 2010; Knobe & Fraser, 2008; Mazzocco, Alicke, & Davis, 2004). That is, people tend to conflate moral evaluations of agents with their perceptions of agents' intentions and causation. For example, in the culpable control model of blame, a desire to assign blame to disliked agents influences perceptions of their control over an accident (Alicke, 2000; but see Malle et al., 2014). In an early demonstration of this phenomenon, participants were told that a man speeding home got into a car accident, leaving another person severely injured (Alicke, 1992). The man was described as rushing home to hide either an anniversary present or a vial of cocaine from his parents. Participants judged the delinquent cocaine-hiding individual as having more control by comparison to the virtuous present-hiding man. Similar effects are seen when participants are given more general information about the agent's character (Alicke & Zell, 2009; Nadler, 2012). People also judge an agent who breaks a rule as being more causally responsible for an outcome that breaks a rule than an agent who takes the same action but does not break a rule, suggesting negative moral evaluations increase causal attributions (Hitchcock & Knobe, 2009).

Moral judgments of agents also affect evaluations of intent. For instance, harmful foreseen side-effects are seen as more intentional than helpful foreseen side effects, suggesting that negative moral evaluations lower the threshold for inferring intentionality (Alicke, 2008; Knobe, 2010; Knobe & Fraser, 2008; Nadelhoffer, 2004; Ngo et al., 2015; Uhlmann et al., 2015). In a study where participants played an economic game with agents who were either trustworthy or untrustworthy, and then evaluated the extent to which the agents intended various positive and negative outcomes, the untrustworthy agent was more likely to be evaluated as intending negative outcomes than the trustworthy agent (Kliemann et al., 2008). Greater activation was seen in the right temporoparietal junction, a region implicated in evaluating intent, when assigning blame to an untrustworthy relative to a trustworthy agent (Kliemann et al., 2008). Thus there is a substantial literature supporting a 'person-as-moralist' view of blame attribution (Alicke et al., 2015; Knobe, 2010; Tetlock, 2002), which posits that people are fundamentally motivated to assess the goodness and badness of others, and perceive others' intent and causation in a way that is consistent with their moral evaluations.

To assign blame and praise it is necessary to infer an agent's mental state based on their actions, by considering the likely end consequences of their action (Malle, 2011). Recent work has shown that from an early age people readily infer people's intentions by observing their decisions, deploying a "naïve utility calculus" that assumes people's choices are aimed at maximizing desirable consequences and minimizing undesirable consequences, where desirability is evaluated with respect to the agent's preferences (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). This means that in situations where agents make deterministic choices, their intentions can be inferred from the consequences of their choices. Evaluations of moral character are intimately linked to inferences about intentions, where accumulated evidence of bad intent leads to a judgment of bad character (Leifer, 1971; Leslie, Knobe, & Cohen, 2006; Uhlmann et al., 2015). What remains unknown is whether, and how, the formation of character beliefs impacts on moral judgments of individual actions. In other words, when people repeatedly observe an agent bring about either harmful or helpful consequences, do learnt inferences about the agent's character influence how people make judgments regarding the agent's individual acts?

Our research addresses several open questions. First, although studies have shown that perceptions of character influence separate assessments of consequences, causation, and blameworthiness, it remains unknown how precisely character evaluations affect the *degree to which* consequences and causation shape blame attributions (Fig. 1). Second, the bulk of research in this area has focused on judgments of blameworthiness for harmful actions with less attention to how people judge praiseworthiness for helpful actions (Cushman et al., 2009; Pizarro, Uhlmann, & Salovey, 2003; Weiner, 1995). Furthermore, those studies that have investigated praiseworthy actions have generally used scenarios that differ from those used in studies of blame not only in terms of their moral status but also in terms of their typicality. For example, studies of blame typically assess violent and/or criminal acts such as assault, theft, and murder, while studies of praise typically assess good deeds such as donating to charity, giving away possessions or helping others with daily tasks (Eisenberg, Zhou, & Koller, 2001; Pizarro et al., 2003). Thus, our understanding of how consequences and causation impact judgments of blame versus praise, and their potential moderation by character assessments, is limited by the fact that previous studies of blame and praise are not easily comparable.

In the current study we used a novel task to explore how inferences about moral character influence the impact of consequences and causation on judgments of blame and praise for harmful and helpful actions. Participants evaluated the blameworthiness or praiseworthiness of several agents' harmful or helpful actions. These varied across trials, in terms of their *consequences* and also in terms of the degree to which the actions *caused* a better or worse outcome for a victim. In Study 1, participants evaluated a total of four agents: two with *good character,* and two with *bad character.* In Study 2 we replicate the effects of Study 1 in a truncated task where participants evaluated one agent with good character and one agent with bad character. We used linear mixed models to assess the extent to which blame and praise judgments were sensitive to the agents' consequences, the agents' causation of the outcomes, the agents' character, and the interactions among these factors. The advantage of this approach is that it allows us to capture the influence of consequences, causation, and character on integrated moral judgments, without requiring participants to directly report their explicit (i.e., self-reported) evaluations of these cognitive subcomponents (Crockett, 2016) (Fig. 1). For example, we can measure whether the effects of perceived causation on blame differs for good and bad agents, without asking participants directly about the perceived causation of good vs. bad agents. With this approach we can more closely approximate the way assessments of consequences and causation influence blame judgments in everyday life, where people might assign blame using implicit, rather than explicit, evaluations of causation and consequences.

We manipulated the agents' consequences by having the agents choose, on each trial, between a *harmful option* that yields a higher monetary reward at the expense of delivering a larger number of painful electric shocks to an anonymous victim, and a *helpful option* that yields a lower monetary reward but results in fewer painful shocks delivered to the victim (Fig. 2A). Across trials we varied the amount of profit and pain that result from the harmful relative to the helpful option. Thus, an agent in choosing the harmful option might inflict a small or large amount of pain on the victim