Journal of Experimental Social Psychology 79 (2018) 149–163

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Optimality bias in moral judgment

Julian De Freitas[a,*,1], Samuel G.B. Johnson[b,1]

[a] Department of Psychology, Harvard University, United States of America
[b] Division of Marketing, Business, & Society, University of Bath School of Management, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Handling editor: Robbie Sutton

Keywords:
Moral judgment
Lay decision theory
Theory of mind
Decision-making
Causal attribution

ABSTRACT

We often make decisions with incomplete knowledge of their consequences. Might people nonetheless expect others to make optimal choices, despite this ignorance? Here, we show that people are sensitive to *moral optimality*: that people hold moral agents accountable depending on whether they make optimal choices, even when there is no way that the agent could know *which* choice was optimal. This result held up whether the outcome was positive, negative, inevitable, or unknown, and across within-subjects and between-subjects designs. Participants consistently distinguished between optimal and suboptimal choices, but not between suboptimal choices of varying quality — a signature pattern of the Efficiency Principle found in other areas of cognition. A mediation analysis revealed that the optimality effect occurs because people find suboptimal choices more difficult to explain and assign harsher blame accordingly, while moderation analyses found that the effect does not depend on tacit inferences about the agent's knowledge or negligence. We argue that this moral optimality bias operates largely out of awareness, reflects broader tendencies in how humans understand one another's behavior, and has real-world implications.

1. Introduction

We hold others accountable for their actions based on what they were thinking. If a student cheats on an exam, a scientist fabricates a result, or a company mistreats a customer, our judgment depends on their motives and beliefs. Empirical studies confirm this intuition (e.g., Cushman, Sheketoff, Wharton, & Carey, 2013; Gray & Wegner, 2008), and all theories of blame must account for it (e.g., Cushman, 2008; Malle, Guglielmo, & Monroe, 2014; Shaver, 1985; Uhlmann, Pizarro, & Diermeier, 2015). Yet, people often seem to blame others for things they could not possibly have known about. In 2009, a group of seismologists issued a statement indicating that an earthquake in L'Aquila, Italy was unlikely; when an earthquake struck and killed 308 people, they were charged with manslaughter. Despite the defense's insistence that it is simply beyond the powers of science to predict earthquakes, the scientists were sentenced to prison. Although the convictions were ultimately overturned, incidents like this highlight ways in which our moral judgments can sometimes directly contradict inferences about agents' intentions. It is perfectly clear that the scientists did not — and *could not* — know that the earthquake would hit, yet many people blamed the scientists all the same. What psychological principles could explain such paradoxical judgments?

One likely factor is the outcome bias (e.g., Baron & Hershey, 1988) wherein people blame agents for negative *consequences* despite positive intentions. For example, the scientists might have been blamed so long as the earthquake occurred, even if the scientists took pains to avoid making the incorrect prediction that the earthquake would not occur. In real world cases, however, multiple factors are often at play. Not only did the scientists' choice result in a bad *outcome*, but, unknown to the scientists, it was also *suboptimal*. That is, even before the earthquake itself, an omniscient scientist could have known that the earthquake was likely to occur. Thus, the optimal choice would objectively have been to recommend evacuation. Given that scientists and other humans are not omniscient, the scientists did not and could not have *known* that their choice was suboptimal. Yet might people nonetheless blame agents for making suboptimal choices, even when agents have no way of knowing that their choices are suboptimal?

1.1. The Efficiency Principle

We propose that moral judgments are influenced by a principle people use for understanding others' behavior, which can override inferences about mental states: People expect agents to behave *optimally* or *efficiently*, relative to the agent's goals and the constraints of the

* Corresponding author.
E-mail address: defreitas@g.harvard.edu (J. De Freitas).
[1] Equal contributions.

https://doi.org/10.1016/j.jesp.2018.07.011
Received 12 November 2017; Received in revised form 23 July 2018; Accepted 23 July 2018

situation (Dennett, 1987; Gergely & Csibra, 2003). To use an analogy outside of moral judgment, if the car to our right changes lanes, we could understand that decision in terms of the assumed beliefs and desires of the car's driver; but in most cases, we probably use the simpler strategy of understanding the car's behavior in terms of more general features of the world, such as common goals (avoiding collisions) and broad situational constraints (intuitive physics and geometry), and assuming optimal decision-making relative to those constraints. This *Efficiency Principle* runs psychologically deep. It develops before a representational theory of mind (Csibra, Gergely, Bíró, Koós, & Brockbank, 1999; Gergely, Bekkering, & Király, 2002) and may scaffold later-emerging mental-state inferences. It also plays important roles — often outside of awareness — in other domains of cognition, including visual perception (Gao & Scholl, 2011) and language understanding (Davidson, 1967; Grice, 1989).

Most of the time, efficiency-based thinking leads to the same conclusions as mentalizing — after all, people behave in a reasonably rational manner much of the time. For example, imagine that Jill is deciding which of three shampoos to buy, wanting to make her hair smell like apples. Suppose that the three brands have different likelihoods of achieving this goal — one has a 70% efficacy (call this "Best"), one a 50% efficacy ("Middle"), and one a 30% efficacy ("Worst") — and that Jill knows these probabilities. If we think about Jill's mental states, we realize she is most likely to choose Best (since Jill believes, correctly, that this choice is optimal), but we can also reach this conclusion by merely considering what is optimal *in the world*, since Jill's mental states track the world. That is, when an agent's beliefs match the world, efficiency-based thinking is a useful shortcut for predicting behavior. This is why most game theory models assume optimal decision-making from one's opponents (e.g., Morgenstern & von Neumann, 1947; Nash, 1951).

However, there are some situations where normative prediction requires us to override the Efficiency Principle — cases in which the agent is *ignorant* of key information. For example, imagine that Jill is in the same situation as before, but falsely believes that all three shampoos are equally likely to achieve her goal. In this case, our representational theory-of-mind tells us that Jill is equally likely to choose each of the three brands, since she has no reason to choose one over the others. Yet, the Efficiency Principle says that Jill would behave optimally relative to the *true* situational constraints, not relative to her *representation* of those constraints — she would be likely to choose the 70% option, and unlikely to choose the other two options.

Surprisingly, even adults are susceptible to such efficiency-based thinking, which can override theory-of-mind. People believe that Jill, even when ignorant about the relevant probabilities, is most likely to choose the optimal (70%) option, and less likely to choose the suboptimal (50% or 30%) options (Johnson & Rips, 2014). Critically, people also believe that Jill is *equally likely* to choose each of the suboptimal (50% and 30%) options; hence, their predictions track optimality as such, rather than the objective probability of success. This stands in contrast both to normative mental-state inferences (i.e., Jill is equally likely to choose each option) and to the predictions people make for agents who do know the probabilities (i.e., she is more likely to choose Best than Middle, but also more likely to choose Middle than Worst). Thus, this *stepwise pattern* of responses — higher predictions for optimal choices, but roughly equal predictions among different suboptimal choices — is a unique signature of efficiency-based reasoning about ignorant agents. This pattern has been found in both predictions of behavior as well as explanations: People believe that suboptimal choices are more in need of explanation than optimal choices because such choices violate our expectations about optimal behavior, eluding the efficiency-based schema we can typically apply (Johnson & Rips, 2014).

1.2. Optimality and morality

These findings led us to predict that suboptimal actions would also lead to (non-normatively) harsher *moral* judgments, in light of people's belief that suboptimal choices are more in need of explanation (Johnson & Rips, 2014). This hypothesis follows from several streams of research.

First, people feel muted affect toward events that are explained (Wilson & Gilbert, 2008) — that is, if they can (intrapersonally) assign meaning to that event. In one study, students studying in the library unexpectedly received a dollar coin attached to an index card. The students maintained a positive mood for a shorter duration when the index card contained text explaining why they had received the gift, compared to when the text on the card eluded explanation (Wilson, Centerbar, Kermer, & Gilbert, 2005). This logic applies to negative events too. Participants encouraged to focus on "why" rather than "what" when recalling an angering experience were less likely to experience negative affect (Kross, Ayduk, & Mischel, 2005). For this reason, people faced with bereavement can cope better with their loss if they are able to find meaning in the death of their loved one (e.g., Bonanno et al., 2002).

Second, affective evaluations are closely linked with moral judgments (Haidt, 2001; Moll, de Oliveira-Souza, Bramati, & Grafman, 2002). This leads to the prediction that merely understanding a behavior (thereby muting affect) can make that behavior seem more consistent with moral norms and less blameworthy — as documented in several studies. For example, when mental disorder symptoms are ordered in a coherent causal chain, people rate individuals with those symptoms as less abnormal (Ahn, Novick, & Kim, 2003; Meehl, 1973). Likewise, jurors are less likely to convict defendants when the defense can tell a coherent story using a given set of facts (Pennington & Hastie, 1992), and people are more likely to be seen as lying when they engage in unusual behaviors — even if the behaviors are irrelevant to deception (Bond et al., 1992). These findings all point to the same underlying phenomenon — when behaviors can be readily explained and meaning can be easily assigned, these behaviors are seen as more typical, more normative, and less blameworthy; conversely, when there is no available explanation for behaviors, they are seen as less normative and more blameworthy.

Third, we can ask what are the key antecedents to the feeling that an explanation is needed (e.g., Bruckmüller, Hegarty, Teigen, Böhm, & Luminet, 2017; Legare, 2012). In addition to lack of a causal chain (Ahn et al., 2003) or coherent order (Pennington & Hastie, 1992), we add perhaps the most critical antecedent of all — violation of expectations. Humans constantly predict the future and modify those predictions in light of actual events (Bar, 2007; Rescorla & Wagner, 1972). For this reason, people are strongly motivated to explain divergences from predicted behavior (Legare, Gelman, & Wellman, 2010; Wong & Yudell, 2015). As we noted earlier, people expect others to behave optimally, even when ignorant of critical information, which in turn leads people to find suboptimal behavior less readily explained than optimal behavior (Johnson & Rips, 2014).

Now we can put these ideas together. When an agent behaves suboptimally, people find that behavior difficult to explain because it violates their expectations — it does not conform to the optimal choice schema and resists attempts to make meaning of it. This feeling leads to more pronounced affective reactions to suboptimal choices, corresponding to more severe moral judgments. We thus predicted an *optimality bias* in evaluations of moral decisions, which would be mediated by the presence or absence of a coherent explanatory schema. Following previous work (Ahn et al., 2003; Johnson & Rips, 2014), we measure this explanatory gap by asking participants to indicate the extent to which they feel that an explanation is needed for the agent's behavior: If the agent behaved optimally, then participants should not feel that an explanation is needed; if the agent behaved suboptimally, then they should. These explanatory judgments should mediate the relationship between optimality and blame (as we test in Study 3).

Although this hypothesis has theoretical support, it has not been tested. The most closely related studies are the many demonstrations of the *outcome bias* (e.g., Baron & Hershey, 1988; see also Martin &