Evolution and Human Behavior 38 (2017) 770–778

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Killing for the greater good: Action aversion and the emotional inhibition of harm in moral dilemmas☆

Melissa M. McDonald [a,*], Andrew M. Defever [b], Carlos David Navarrete [b]

[a] Department of Psychology, Oakland University, United States
[b] Department of Psychology, Michigan State University, United States

ARTICLE INFO

Article history:
Initial receipt 9 March 2017
Final revision received 5 June 2017

Keywords:
Trolley problem
Emotional arousal
Action aversion
Virtual reality
Moral judgment
Moral decision making

ABSTRACT

Moral judgment is influenced by both automatic and deliberative processing systems, and moral conflict arises when these systems produce competing intuitions. We investigated the role of emotional arousal in inhibiting harmful action in a behavioral study of utilitarian tradeoffs in a 3D digital simulation of two classic "trolley" scenarios in which participants decided whether to harm one person in order to avert harm to five others. Physiological arousal was measured via skin conductance response in real time. Results showed that physiological arousal is increased in situations in which using personal harm is necessary to achieve a utilitarian outcome relative to when the same outcome can be achieved with impersonal harm, and is linked to a decreased likelihood of engaging in harmful action, though a test of mediation was not statistically significant. In addition, when the use of personal harm was required to save lives, arousal was higher *pre*-action relative to *post*-action. Overall, our findings suggest that physiological arousal may be part of an affective system that functions to inhibit harmful action against others.

1. Introduction

There is a long-standing tension between utilitarian and deontological philosophical perspectives on morality. The deontological tradition emphasizes the rights and duties of individuals as intrinsic axioms, regardless of the consequences (Broad, 1930). In contrast, utilitarianism is necessarily *consequential* in its judgment of action, seeking the maximum welfare for the greatest number (Mill, 1863). Consider a dilemma implemented by Greene, Sommerville, Nystrom, Darley, and Cohen (2001). You are with a group of people hiding from an aggressive militant group attempting to seek you out. If they find you, they will kill the entire group. You are holding a baby that begins to cry loudly, which will attract the attention of your pursuers. You can attempt to smother the baby to silence its cries, but in doing so you will kill it. For deontologists, the act of smothering the baby is immoral, not because of the consequences of doing it, but because we have a moral duty to avoid actions that cause harm. But for utilitarians, smothering a baby is not intrinsically wrong, and may be permissible if it saves a large group of people hiding from soldiers who would have otherwise heard its cries.

The *dual process* approach to moral psychology describes the historical tension between utilitarian and deontological philosophy as the competition between two separate psychological systems rooted in our species' neurocognitive architecture. Moral dilemmas arise when we contemplate pitting consequential considerations for a greater good against our non-consequentialist intuition to avoid harm, each generated by a distinct psychological system within the mind (Greene, 2013). In dilemmas such as the crying baby scenario above, one system judges the consequences of actions in utilitarian terms (e.g., "one harmed is better than many harmed"), and relies on processes that are controlled, deliberative, and logical. The other system is informed by affective feedback about one's action (e.g., "I feel terrible about this"), and relies on processes that are automatic, intuitive, and emotional.

The dual process perspective posits that utilitarian outcomes requiring harmful action can only occur when signals from the affective system impeding harmful action are quelled by the "cooler" deliberative

☆ This material is based on work supported by the National Science Foundation under Grant BCS-0847237. This sponsor did not have involvement in study design; data collection, analysis, or interpretation; in the writing of the report; or in the decision to submit the article for publication. We thank the research assistants in the Morality and Intergroup Relations Lab for their help conducting this study. We also thank Todd Shackelford, Dan Ilgen, and Rebecca Gore for their feedback.

* Corresponding author at: Oakland University, Department of Psychology, 654 Pioneer Dr., Rochester, MI 48309-4482, United States.
E-mail address: mmmcdonald@oakland.edu (M.M. McDonald).

http://dx.doi.org/10.1016/j.evolhumbehav.2017.06.001
1090-5138/

processing areas of the mind involved in weighing the costs and benefits of a particular decision.[1] Conversely, our "hot" emotional aversion to harm may override consequentialist reasoning, stymieing the process of deliberate utilitarian action in favor of deontological judgments, particularly for intentionally harmful actions requiring the use of personal force (Greene et al., 2009). For example, in the classic trolley scenarios, most people judge it morally permissible to save the lives of five railway workers in the path of a runaway trolley by impersonally pulling a switch to divert the trolley to a sidetrack where it kills only one worker. However, when saving the five requires forcefully pushing a large man off a footbridge into the path of the oncoming trolley, most judge it to be morally wrong (e.g., Cushman, Young, & Hauser, 2006).

1.1. Action aversion

Past research has provided some support for the notion that the consideration of the consequences of one's actions, such as empathy for a victim in distress, makes the performance of harmful acts aversive (e.g., Batson, 1983; Batson & Ahmad, 2009). In contrast to such "outcome aversion" approaches to harm avoidance, other researchers posit that the aversion to harmful action may be triggered by the mere anticipation of the motor properties of a harmful action, without the consideration of its consequences (e.g., Blair, 1995). Congruent with this line of reasoning, Cushman, Gray, Gaffey, and Mendes (2012) found evidence for an automatic aversion to committing physical harm, evoked by the intrinsic properties of the action alone. Across two studies, research participants' autonomic nervous system activity was recorded while committing simulated harmful actions such as "stabbing" an experimenter with a rubber knife or "shooting" them in the face with a gun replica. They found that committing simulated harmful actions was associated with heightened autonomic activity relative to when the same actions were observed being committed by a third party.

The authors posit that humans may be endowed with a neurophysiological mechanism for harm avoidance, articulated as the *action aversion hypothesis*, which states that "physiological aversion can be triggered by only the motor or perceptual properties of harmful action" (Cushman et al., 2012, p. 3). When a person prepares for the performance of an act that would typically result in harm, the same physiological processes are activated as when harm would actually occur even if it is a mere simulated action. Additionally, that the autonomic activity occurred before and during the harmful acts, and decreased after the action was completed suggests that the physiological response has an inhibitory function (Cushman et al., 2012). According to Cushman et al., this mechanism operates as a function of the anticipation of the actions themselves, and does not necessarily depend on the "real world" consequences of the actions.

1.2. Evolutionary roots of an action aversion system

A critical component of the original dual-process approach as applied to action aversion is the notion that action aversion to first-person harm should be operative primarily when the anticipated harmful act requires the use of interpersonal force in close contact, and less so when the actions are conducted via impersonal action where causal chains are distally linked (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2009). There are at least two reasons to expect this distinction between personal and impersonal action for harm aversion, which we discuss below.

1.2.1. A phylogenetic by-product

The "hot" moral judgment system is likely to be most strongly activated when mentally simulating or engaging in violent actions with motor patterns that have deep phylogenetic roots that would be familiar to a range of primates due to our common ancestry. This system is markedly less active when considering harmful actions engaging the more recently evolved systems, such as those underlying the unique human ability for "cooler" abstract reasoning processes about impersonal harm requiring complex reasoning about both animate and inanimate mechanical causation. For example, pulling a lever that changes the direction of a trolley to avert a tragedy requires certain kinds of reasoning abilities not widely shared among mammals, and likely has more shallow phylogenetic roots than the cognitive abilities to process harmful action via personal contact.

1.2.2. A reputational adaptation

A functional approach to the moral psychology of harm suggests that harmful actions are aversive to the actor because of the potential costs involved. For example, harmful action, even if implemented for a net benefit to others, can lead to aggressive resistance or retaliation by the victim or third parties, and may have negative long-term reputational consequences. Such an automatic negative reaction to harm involving personal force may function to prevent actions for which the potential for plausible deniability of culpability is limited (DeScioli, Bruening, & Kurzban, 2011; Greene, 2013; Pinker, 2007). Consider, for example, the certainty with which we know Jack Ruby killed Lee Harvey Oswald, relative to the confidence we have that Oswald killed President John F. Kennedy. The details involved in each case are such that counterfactuals involving culpability are more readily generated for the assassination of Kennedy relative to the killing of Oswald.

In sum, an action aversion mechanism that produces an automatic reaction to personal harm subjectively experienced as negative arousal (e.g., fear, disgust) should be activated most strongly when actions require the use of personal force (pushing a bystander to their death to stop a trolley), and less so in impersonal interaction (flipping a switch to divert the trolley). The underlying reason for this could be a result of the phylogenetic age of mechanisms for performing complex vs simple motoric action, activation of reputation preservation mechanisms, or other factors not explored here. These functional explanations are not mutually exclusive, and do not contradict Cushman et al.'s (2012) or Greene et al.'s (2001) accounts of the proximate neurophysiological responses. The current research speaks to, but does not directly test, these ultimate explanations. Rather we examine the implications of the dual process model within the action aversion paradigm.

1.3. The present research

The notion of action aversion is particularly relevant in moral dilemmas contrasting impersonal harm at a distance versus harm that is "up-close-and-personal." The variants of the trolley dilemma described above illustrate the relevance of the action aversion hypothesis, as research shows that utilitarian judgments in surveys are more likely when the hypothetical harm is imagined to be committed impersonally compared to personally (e.g., Cushman et al., 2006).

Greene et al. (Greene et al., 2004; Greene, 2007) provided preliminary evidence for the presence of a dual-process system. Utilizing neuroimaging techniques and examining neural activation, they found that when participants imagined a dilemma requiring personal force (e.g., pushing a person) to kill one person to save five others, "a prepotent, negative emotional response" was activated (Greene, 2007, p. 322). The "hot" response served to inhibit the initial harmful act in some, whereas the "cooler" deliberative system allowed others to overcome the response and engage in the action.

[1] Classifying the outcome of a decision as utilitarian is often presumed to imply that an individual reached the decision through the use of conscious utilitarian reasoning, in which an individual decided whether to take action on the basis of which option would maximize the welfare of those involved in the dilemma. This assumption may not hold, as the true reasons behind a moral judgment may be inaccessible or simply post-hoc rationalizations (Haidt, 2001). Thus, it should be noted that throughout this article the use of the terms utilitarian and deontological denote a classification of an outcome that could be perceived as utilitarian or deontological, but not that a particular person is reasoning with such specific philosophical premises in mind.