ELSEVIER

Evolution and Human Behavior 33 (2012) 323 – 333

Evolution and Human Behavior

Original Article

Hamilton vs. Kant: pitting adaptations for altruism against adaptations for moral judgment

Robert Kurzban[a,*], Peter DeScioli[b], Daniel Fein[a]

[a]Department of Psychology, University of Pennsylvania, Philadelphia PA 19104, USA
[b]Departments of Psychology and Economics, Brandeis University, Waltham, MA 02453, USA

Initial receipt 22 February 2010; final revision received 3 November 2011

Abstract

Prominent evolutionary theories of morality maintain that the adaptations that underlie moral judgment and behavior function, at least in part, to deliver benefits (or prevent harm) to others. These explanations are based on the theories of kin selection and reciprocal altruism, and they predict that moral systems are designed to maximize Hamiltonian inclusive fitness. In sharp contrast, however, moral judgment often appears Kantian and rule-based. To reconcile this apparent discrepancy, some theorists have claimed that Kantian moral rules result from mechanisms that implement simple heuristics for maximizing welfare. To test this idea, we conducted a set of studies in which subjects (N=1290) decided whether they would kill one person to save five others, varying the relationship of the subject with the others involved (strangers, friends, brothers). Are participants more likely to observe the Kantian rule against killing in decisions about brothers and friends, rather than strangers? We found the reverse. Subjects reported greater willingness to kill a brother or friend than a stranger (in order to save five others of the same type). These results suggest that the rule-based structure of moral cognition is not explained by kin selection, reciprocity, or other altruism theories.

Keywords: Morality; Altruism; Kin selection; Inclusive fitness; Judgment & decision making; Reasoning

1. Introduction

1.1. Human moral cognition is nonconsequentialist

Consider an organism faced with a dilemma. It can either kill one of its offspring, which will allow five others to live, or it can do nothing, in which case five of its offspring will die. Hamilton's (1964) theory of kin selection predicts that evolution will favor designs for killing one relative to save five others. Indeed, this behavior has been observed in many species, such as the dramatic case of the burying beetle, which kills some offspring in order to feed the bodies to other offspring (Mock, 2004). The burying beetle's decisions are, just as kin selection predicts, *consequentialist*, based exclusively on outcomes. More precisely, the mechanism that causes the burying beetle's infanticide-and-regurgitation behavior was selected by virtue of the inclusive fitness consequences of its choices.

Immanuel Kant would argue, however, that when humans face this dilemma, they should *not* kill one to save five because there is an inviolable moral rule against killing that cannot be broken regardless of the consequences. Kant's view—*nonconsequentialism*—is reflected in a characteristic feature of human moral cognition: Moral judgment is rule-based and focuses on *behavior per se*, the means used to accomplish outcomes, rather than on the *outcomes*, or ends (DeScioli & Kurzban, 2009a). For instance, in the footbridge version of the Trolley Problem (Foot, 1967; see Methods, below), 90% of people judge that it is impermissible to kill one person to save five people (Hauser, Young, & Cushman, 2008).

Why is the burying beetle's behavior consequentialist while human moral judgment is nonconsequentialist? Instead of using simple rules such as "never kill," "never steal," or "never eat pork," humans could make moral decisions based on only the costs and benefits of their options. The phenomenon of nonconsequentialism in moral

* Corresponding author. Department of Psychology, University of Pennsylvania, Philadelphia PA 19104, USA. Tel.: +1 215 898 4977; fax: +1 215 898 7301.
E-mail address: kurzban@psych.upenn.edu (R. Kurzban).

1090-5138/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.evolhumbehav.2011.11.002

judgment is easily overlooked as a puzzle because it is so familiar and intuitive (Cosmides & Tooby, 1994). But this feature poses a problem: Why do humans focus moral decisions on behavior rather than considering only the consequences?

1.2. Describing the problem with choice theory

Basic choice theory clarifies the distinction between consequentialism and nonconsequentialism. In choice theory, there is a decision maker who selects an action, *a,* from a set of possible actions, A, and each action is associated with possible outcomes. Here, an outcome consists of payoffs to the organism and other relevant organisms, i.e., a vector $\mathbf{y}$ of payoffs to the self and others. Finally, the decision maker has a standard utility function for ranking outcomes depending on the resulting payoffs, $u(\mathbf{y})$.

A consequentialist decision procedure would choose an action, a^*, to maximize utility:

$$\max_{a \in A} u(\mathbf{y}), \tag{1}$$

where $u(\mathbf{y})$ depends only on the vector of payoffs, $\mathbf{y}$. This encompasses a range of utility functions including any weighted sum of payoffs to the self and others, whether characterized by extreme selfishness, universal altruism, or altruism skewed toward family and friends.

In contrast, the Kantian decision procedure would choose an action not only based on the payoffs $\mathbf{y}$, but also based on whether the action is labeled morally wrong. Morally wrong actions are excluded regardless of the payoffs they generate. The Kantian approach can be expressed as maximization subject to constraints on the actions:

$$\max_{a \in A} u(\mathbf{y}), \text{subject to the constraint}, a \notin W, \tag{2}$$

where W refers to a set of actions labeled morally wrong. The Kantian decision rule excludes actions in W and then maximizes utility subject to this constraint.

Moral dilemmas occur when maximization based on payoffs (1) conflicts with moral constraints (2). Specifically, a dilemma arises when the action that maximizes utility for the consequentialist, a^*, is in the set W of moral wrongs. In these situations, decision processes (1) and (2) lead to different choices. Empirical observations show that people's choices are sometimes most consistent with (1), as in the switch version of the Trolley Problem, and sometimes with (2), as in the footbridge version of the Trolley Problem (Hauser, 2006). This pattern of results suggests that both of these conflicting decision processes are used to some extent, which is presumably why humans perceive these problems as "dilemmas" rather than having clear-cut solutions.

Consequentialist mechanisms pose no theoretical difficulty because evolution favors adaptations based on the payoffs they produce. Kin selection, for example, favors mechanisms that maximize a weighted sum of individuals' payoffs based on relatedness (Hamilton, 1964). That is, Hamilton's theory is consistent with decision procedure (1), and observations in species such as the burying beetle support the theory. Similarly, reciprocity (Trivers, 1971), mutualism (Sachs, Mueller, Wilcox, & Bull, 2004), and costly signaling (Zahavi, 1975) can also favor consequentialist mechanisms with positive weights on others' payoffs.

In contrast, Kant's moral philosophy is described by decision procedure (2), and current theories do not straightforwardly explain what selection pressures give rise to it, leaving a gap in our understanding of human moral judgment. When people obey moral constraints, choosing actions other than a^*, as in (2), an explanation is required. What is the function of the constraining mechanism?

1.3. Does prohibiting beneficial acts generate benefits?

One common proposal is that simple moral rules of behavior such as "do not kill" or "do not sell sex" function to promote welfare (Gigerenzer, 2008). However, because these rules pertain to behavior *per se*, they necessarily prohibit beneficial actions in moral dilemmas when forbidden acts can yield net benefits. This raises the question: How can *prohibiting* beneficial acts *generate* benefits?

Psychologists have argued that Kantian rules of behavior, contrary to appearance, maximize welfare in the long run, on average, even if they lead to occasional errors (e.g., Gigerenzer, 2010). The idea is that calculating welfare consequences for specific cases is too computationally demanding, necessitating simple rules. This theory resembles the position in moral philosophy of "rule utilitarianism," in which a set of inflexible rules is observed because it is the best feasible way to maximize welfare (Sunstein, 2005).

In one version of this argument, moral constraints are implemented by emotions (Greene, 2007). The reluctance to kill in the context of moral dilemmas, on this view, is due to emotional systems that guide behavior (Haidt, 2001). That is, these emotions, whose function is to "motivate altruistic behavior" (Pyysiäinen and Hauser, 2010, p. 105), inhibit the choice of a^* when it is in W. In sum, *the predominant explanation for nonconsequentialism is that these judgments reflect the operation of human altruism systems that are implemented via moral rules of behavior.*

In contrast, we propose the alternative hypothesis that human altruism systems are consequentialist, as in (1) above, just like altruism mechanisms in burying beetles. If this is true, then nonconsequentialism in moral dilemmas is not due to the operation of altruism mechanisms. Instead, we have argued elsewhere that moral nonconsequentialism might be designed for strategic interactions among perpetrators, victims, and third-party condemners (DeScioli & Kurzban, 2009a). Here, however, we focus on the nature of human altruism systems, specifically whether or not these systems are consequentialist.

These two possibilities are shown in Fig. 1. The first possibility, depicted in the top panel, is that altruism