ELSEVIER

Evolution and Human Behavior 29 (2008) 242–248

Evolution and Human Behavior

The price you pay: cost-dependent reputation effects of altruistic punishment

Rob M.A. Nelissen*

Tilburg University, The Netherlands

Initial receipt 24 September 2007; final revision received 1 January 2008

Abstract

Two studies were conducted to test reputation-based accounts of altruism which predict that the more people sacrifice to help others, the greater their ensuing benefits. We tested this prediction by varying the cost invested in altruistic behavior, here modeled as costly sanctioning of unfair behavior. Confirming this prediction, it was found that only altruists who invested most in the punishment of unfairness were preferred as partners and were transferred more money in a subsequent trust game. This implies that the benefits of behaving altruistically depend upon how much one is willing to pay. It is discussed that these results fit both an indirect reciprocity and a costly signaling framework.

Keywords: Altruism; Altruistic punishment; Reputation; Costly signaling; Indirect reciprocity

1. Introduction

Human altruism is unique because (a) it extends beyond the boundaries of kinship more profoundly than in any other species, and (b) it sometimes benefits individuals with whom future interactions are unlikely. Consequently, researchers have proposed new explanations for altruistic behavior in addition to traditional notions of kin selection (Hamilton, 1964) and direct reciprocity (Axelrod, 1984; Trivers, 1971). In search of additional explanations, reputation-based accounts for the evolution of altruism are currently gaining momentum (for overviews, see Fehr & Fischbacher, 2003; Nowak, 2006). Altruistic behavior establishes a reputation from which individuals may benefit in two ways. According to indirect reciprocity theory (IRT; Alexander, 1987; Nowak & Sigmund, 2005), other group members reward acts of altruism even though they did not personally benefit. Costly signaling theory (CST) holds that people signal favorable yet unobservable traits by behaving altruistically (Bliege Bird & Smith, 2005; Cronk, 2005; Gintis, Smith, & Bowles, 2001; Zahavi & Zahavi, 1997), making them attractive mates, allies, or partners.

Support for reputation-based accounts of altruism comes from two lines of research. First, studies have established that altruistic behavior may indeed have beneficial consequences such as indirect rewards and enhanced status. Second, in line with expectations that can be derived from a reputation-based view of altruism, research has demonstrated that varying cues as to the observability or visibility of behavior may elicit or enhance altruistic behavior. Empirical findings from both approaches will be briefly reviewed below. I will add to this growing body of research by combining both approaches. Specifically, I present results of two studies that show that beneficial consequences of altruistic behavior may vary as a result of information concerning the costliness that behavior.

1.1. Reputation-based effects of altruism

Some ethnographic (e.g., Bliege Bird & Smith, 2005; Gurven, 2004) and experimental (Barclay, 2006; Milinski, Semmann, & Krambeck, 2002a; 2002b) studies report that altruistic behavior is indeed rewarded by other group

* Department of Social Psychology and TIBER (Tilburg Institute of Behavioral Economics Research), Tilburg University, PO BOX 90153, 5000-LE Tilburg, The Netherlands.

E-mail address: r.m.a.nelissen@uvt.nl.

1090-5138/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.evolhumbehav.2008.01.001

members. What appears to be payback by third parties is consistent with indirect reciprocity accounts of altruism. One potentially puzzling fact about these findings is that rewarding altruism is itself an altruistic act that faces a second-order free-riding problem (e.g., Henrich & Boyd, 2001); however, recent modeling studies show that this can be a stable strategy under certain conditions (Panchanathan & Boyd, 2004).

Consistent with a costly signaling view of reputation-based altruism, several studies show that public displays of altruism increase people's status (Boone, 1998; Hardy & Van Vugt, 2006; Milinski et al., 2002b; Smith & Bliege Bird, 2000; Smith, Bliege Bird, & Bird 2003) and make them more attractive partners in subsequent interactions (Barclay, 2006). Costly signaling theory can be characterized as an account of apparent waste in animal life (Veblen, 1994/1899; Zahavi, 1975; Zahavi & Zahavi, 1997). It proposes that individuals incur costs in terms of time, energy, risk, or money to convey information about personal traits that are otherwise unobservable. The costs associated with the signal guarantee its reliability and thereby recipients' attention. (Notably, cost alone is not sufficient to guarantee reliability; the cost—or ensuing benefit—must be condition-dependent, such that those of lower quality signal at lower rates or not at all.) Hence, according to CST, altruistic behavior reflects desirable personal traits and therefore benefits signalers as it sways others' perceptions and preferences. Notably, costly signaling is advantageous to both signalers and receivers. Unlike indirect reciprocity, it does not require others to behave altruistically in return—others merely behave in their personal interest by acting on signals of quality.

1.2. Elicitors of reputation-based altruism

A second line of research focuses on antecedents and entails demonstrations that altruistic behavior is sensitive to cues that ought to affect its occurrence, if it served the proposed function (Andrews, Gangestad, & Matthews, 2002; Cosmides & Tooby, 1992; Smith, Borgerhoff-Mulder, & Hill, 2001; Williams, 1966). In line with reputation-based accounts, studies show that altruistic acts become more frequent in the presence of others (e.g., Hardy & Van Vugt, 2006; Kurzban, DeScioli, & O'Brien, 2007; Rege & Telle, 2004), in theory because an audience increases broadcasting efficiency. Nevertheless, it is unclear whether audience effects should be attributed to increased opportunities for gaining a good reputation or increased likelihood of being hurt by acquiring a bad one (e.g., Hoffman, McCabe, Shachat, & Smith, 1994). The latter implies that rather than seeking to boost their reputation, people want to avoid punishment.

In order to further explore the relationship between reputation and altruism, the present studies attempt to integrate both of the aforementioned lines of research by showing condition-dependent beneficial effects of altruistic behavior. Specifically, I studied how the costs invested in an altruistic act influence its interpersonal consequences.

1.3. Cost-dependent reputation effects of altruism

The present studies varied the costs invested by three (fictional) individuals in punishing an unfair distribution between two other people (cf., Barclay, 2006; Kurzban et al., 2007). Third-party sanctions present the paradigmatic case of altruistic punishment (Fehr & Fischbacher, 2004): A disinterested (third) party witnesses an unequal distribution between two other people and subsequently has an opportunity to punish the allocator for this unfair behavior. As punishment is costly and the third party was not personally affected by the unfair distribution—acts like these are referred to as altruistic punishment. Altruistic punishment is generally considered a form of second-order altruism (see also Boyd, Gintis, Bowles & Richerson, 2003; Fehr & Gächter, 2002). Research has demonstrated that people perceive the willingness to engage in altruistic punishment as a signal of a person's concerns for fairness and thereby his/her trustworthiness in cooperative interactions (see also Barclay, 2004, 2006; Fessler & Haley, 2003; Frank, 1988).

According to CST, the costs invested in a signal are a means of ensuring its reliability. By extension, the more that an individual is willing to invest in order to punish unfairness, the more reliably this conveys concerns for fairness and trustworthiness. Hence, not only will punishers be preferred over nonpunishers, but, moreover, more reliable signalers, who incur greater costs (i.e., strong punishers), will also be preferred over less reliable signalers, who incur smaller costs (i.e., weak punishers). Although these predictions are explicitly entailed by CST, interestingly, similar predictions can potentially also be derived from IRT: it is conceivable that, if punishment is regarded as a prosocial act, under some systems of indirect reciprocity, punishers would be rewarded in accordance with their sacrifice, and hence, the greater the costs incurred to punish, the greater the reward. Note, however, that whereas CST explicitly entails inferences about unseen traits (herein perceived fairness or trustworthiness), such inferences are not an intrinsic feature of this extrapolated version of IRT. In order to explore these possibilities, I used a trust game (Berg, Dickhaut, & McCabe, 1995) to examine perceptions of, and behavior toward, ostensible third-party punishers.

In a trust game, one player (the participant in the present studies) is endowed with a sum of money, of which they may transfer a proportion to another player (the partner). By doing so, the money triples in value. (If the player decides to keep all the money, the game ends.) Subsequently, the other player decides whether or not to return a proportion to the first player. Generally, the amount transferred by the first player is regarded as a measure of trust (for an overview, see Camerer, 2003). To my knowledge, this is the first study that actually tests reputation-based accounts of altruism by manipulating the costs invested in altruistic behavior. Specifically, it is predicted that people confer social benefits (both in terms of enhanced preference and financial rewards)