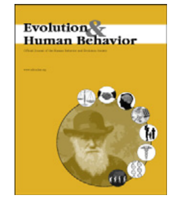
Evolution and Human Behavior 38 (2017) 27–38

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org



Original Article

A signal-detection approach to modeling forgiveness decisions

Jolene H. Tan *, Shenghua Luan, Konstantinos Katsikopoulos

Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195, Berlin, Germany

ARTICLE INFO

Article history:
Initial receipt 29 January 2016
Final revision received 7 June 2016

Keywords:
Cooperation
Forgiveness
Fast-and-frugal trees
Franklin's rule
Error management theory
Signal detection theory

ABSTRACT

Whether to forgive is a key decision supporting cooperation. Like many other evolutionarily recurrent decisions, it is made under uncertainty and requires the trade-off of costs and benefits. This decision can be conceptualized as a signal detection or error management task: Forgiving is adaptive if a relationship with the "harmdoer" will be fitness enhancing and not adaptive if the relationship will be fitness reducing, and the decision should be biased toward lowering the likelihood of the more costly error, which depending on the context may be either erroneously not forgiving or forgiving. Building on such conceptualization, we developed two cognitive models and examined how well they described participants' forgiveness decisions in hypothetical scenarios and predicted their decisions in recalled real-life incidents. We found that the models performed similarly and generally well—around 80% in describing and 70% in prediction. Moreover, this modeling approach allowed us to estimate the decision bias of each participant; we found that the biases were generally consistent with the prescriptions of signal detection theory and were directed at reducing the more costly error. In addition to testing mechanistic models of the forgiveness decision, our study also contributes to forgiveness research by applying a novel experimental method that studied both hypothetical and real-life decisions in tandem. These models and experimental methods could be used to study other evolutionarily recurrent problems, advancing understanding of how they are solved in the mind.

"Forgiveness is the bridesmaid; cooperation is the bride."
[Michael McCullough, *Beyond Revenge*: *The Evolution of the Forgiveness Instinct*]

1. Introduction

Recurrent cooperative relationships are widespread in humans and other social animals (Dugatkin, 2002). Because such relationships are often threatened by harm arising from conflicts of interests, communication errors, or mere random noise, choosing an appropriate action in the aftermath of harm is a crucial evolutionarily recurrent problem that social animals would have evolved mechanisms to solve (Aureli, Cords, & van Schaik, 2002). Harm can, but need not, result in revenge and the termination of cooperation; agents may instead choose to forgive the "harmdoer" and to continue the relationship (McCullough, Kurzban, & Tabak, 2013). Because forgiveness can be fitness enhancing by maintaining cooperation over time (Godfray, 1992), deciding whether to forgive is a key decision of cooperation.

Understanding cooperation among nonkin has received significant research attention in the last decades. The first part of this endeavor has been to explain *why* an agent performs costly actions to benefit another. Inspired by work on reciprocal altruism (Trivers, 1971) and game theoretical insights (Axelrod, 1980; Boyd & Richerson, 1992), researchers have made major advances in understanding how cheaters are curbed so that cooperation can be beneficial (Kurzban, Burton-Chellew, & West, 2015). The second part has been to clarify *how* agents cooperate, which has been referred to as creating "high-resolution maps" of the intricate proximate phenotypic mechanisms (Cosmides & Tooby, 1992). Unlike understanding why, mechanistic understanding is still nascent (Bshary & Oliveira, 2015). It is an open question what computational rules are used to process information in decisions regarding forgiveness and cooperation (e.g., Schacht & Grote, 2015).

Like most evolutionarily recurrent tasks, decisions about forgiveness are made under uncertainty and feature asymmetric costs and benefits (McCullough et al., 2013). One way of conceptualizing how this asymmetry may shape the decision process is through the lenses of signal detection theory (SDT; Green & Swets, 1966) and error management theory (EMT; Haselton & Buss, 2000; McKay & Efferson, 2010). Between the two theories, SDT provides a precise formulation of how cost–benefit tradeoffs should be made while EMT applies these principles to explain the existence of biases in evolved cognitive systems. Specifically, because errors are inevitable in uncertain environments

* Corresponding author.
E-mail addresses: tan@mpib-berlin.mpg.de, jolene.tan.h@gmail.com (J.H. Tan), shluan@mpib-berlin.mpg.de (S. Luan), katsikop@mpib-berlin.mpg.de (K. Katsikopoulos).

http://dx.doi.org/10.1016/j.evolhumbehav.2016.06.004
1090-5138/

Table 1
Possible outcomes of forgiveness decisions.

Decision	Nature of the harmdoer	
	Ally	Foe
Forgive	True positive	False positive
Do not forgive	False negative	True negative

Note. An ally denotes an agent with whom a relationship will bring more fitness benefits than costs, whereas a foe is the reverse.

and have different costs, adapted systems of cognition are biased to guide behavior toward incurring the less costly error (Haselton & Nettle, 2005). Thus, biases are design features rather than defects and should be calibrated by the relative effects of errors on fitness.

In this light, we investigated decisions about forgiveness as signal detection tasks and expected the decision process[i] to resemble that of similar tasks. This perspective allowed us to hypothesize about the characteristics of the decision process and the contexts in which biases toward or against forgiveness would occur. For example, given the same harm situation, we would expect agents to be biased toward forgiving those with whom they have fitness interdependencies and biased against forgiving others with whom they have unrewarding relations (McCullough et al., 2013).

The focus of our study is on forgiveness decisions and its potential interpersonal consequences.[ii] We first specified the structure of the task and its possible cognitive solutions, and then investigated how forgiveness decisions could be described and predicted by two models: a compensatory weighting-and-adding linear model and a noncompensatory fast-and-frugal heuristic. Both models incorporate the essentials required to make cost–benefit tradeoffs but make different assumptions about cognitive implementation. The parameters estimated in these models allow us to test predictions regarding the impact of error costs on the direction and magnitude of bias. Beyond forgiveness decisions, this modeling approach can potentially be applied to understand how agents solve other evolutionarily recurrent problems under uncertainty (e.g., Johnson, Blumstein, Fowler, & Haselton, 2013).

1.1. Forgiveness as signal detection

1.1.1. The framework

Forgiveness functions to maintain relationships after conflict and enable continued cooperation between the victim and the harmdoer (Burnette, McCullough, Van Tongeren, & Davis, 2012; McCullough, Fincham, & Tsang, 2003). From this perspective, forgiving is adaptive if the harmdoer is an ally but not a foe. We use the term *ally* to refer to an agent with whom a relationship will result in more fitness benefits than costs, and *foe* as one with whom a relationship will result in more costs than benefits (McCullough et al., 2013; Tan & Luan, 2015). Table 1 displays the four possible outcomes of forgiveness decisions: Correct decisions are when an ally is forgiven (*true positive*) and a foe is not (*true negative*) and incorrect decisions are when a foe is forgiven (*false positive*) and an ally is not (*false negative*). With a true positive, the victim gains the net benefits from the relationship with the ally, and with a false negative, the victim misses out on those benefits. On the flip side, with a false positive, the victim faces the net exploitation costs of the relationship with the foe, and with a true negative, the victim is spared those costs.

Informed by SDT, we assume that there are two subprocesses involved in the decision of whether to forgive: judging the *strength of evidence* that the harmdoer is an ally and setting an appropriate bias, or *decision criterion*. Forgiveness is chosen when the evidence strength exceeds the decision criterion (see Fig. 1). Setting a liberal criterion means forgiving even when the evidence is weak, indicating a bias toward forgiving, whereas a conservative criterion means forgiving only when the evidence is strong, indicating a bias against forgiving.

When an agent has a high prosocial concern for the other's welfare relative to its own, the agent is more likely to make sacrifices and provide fitness benefits to the other (e.g., Struthers, Eaton, Santelli, Uchiyama, & Shirvani, 2008; Tooby & Cosmides, 2008). Thus, the greater the harmdoer's inferred prosocial concern for the victim, the stronger the evidence that the harmdoer is an ally. However, strong evidence is no guarantee that the harmdoer is indeed an ally, because evidence is inferred from past and current observations, which are imperfectly linked to the future, and it is likely to be perceived with noise. As such, there is inherent uncertainty in the decision, which is illustrated in Fig. 1 by the overlapping evidence-strength distributions of allies and foes.

Given this uncertainty, where should the decision criterion be set? In other words, how strong must the evidence strength be for the harmdoer to be forgiven? The selection of the criterion reflects a trade-off: Assuming that the two distributions are fixed, a liberal criterion reduces the likelihood of a false negative (i.e., not forgiving an ally) at the expense of increasing that of a false positive (i.e., forgiving a foe), and a conservative criterion has the opposite effect. To lower the total cost of errors and increase the expected benefits of the decision, a liberal criterion should be adopted when false negatives are costlier, and a conservative one when false positives are costlier.

1.1.2. Predictors of forgiveness

Given that forgiveness decisions are made in a wide variety of contexts that vary in cost–benefit asymmetry, decision makers need to judge the evidence strength and cost of errors by processing information in the environment that is predictive of the two. In this section we review some predictors examined in the present study (summarized in Table 2).

Evidence strength. Judging the harmdoer's prosocial concern or the strength of the evidence that the harmdoer is an ally requires insight into the mental state of the harmdoer. To this end, the victim may consider the harmdoer's intent to harm, whether the harmdoer was to blame for the harm, and whether a sincere apology was offered. These three cues or predictors were taken from a meta-analysis of forgiveness involving 175 studies (Fehr, Gelfand, & Nag, 2010). They were the variables identified with the strongest main effects on forgiveness within the category related to making sense of a harm and the harmdoer. The other two variables in this category, harm severity and rumination, were not indicative of the harmdoer's mental state and hence were not included in the present study. Of the three, intent was found to have the strongest effect on forgiveness, followed by apology and blame. In addition to being well studied, the level of abstraction of these cues makes them relevant across a wide range of forgiveness contexts.

With an intent to harm, the harmdoer is inferred to have the goal of reducing the victim's fitness, or at the very least, to be indifferent to the impact the action would have on the victim's welfare (e.g., Malle & Knobe, 1997; Struthers et al., 2008; Weiner, 1995). Intention to harm is thus a strong cue that the harmdoer is likely to repeat the harm and that the strength of the evidence that the harmdoer is an ally is low (Petersen, Sell, Tooby, & Cosmides, 2012).

The concept of blame is closely related to attributions of responsibility and accountability (Weiner, 1995). Blame is assigned when a harmdoer's actions directly led to the harm done or when the harmdoer could have prevented the harm (Alicke, 2000). Blame is generally less indicative

[i] There has been debate about whether bias occurs on the level of perception or decision (e.g., Marshall, Trimmer, Houston, & McNamara, 2013; McKay & Efferson, 2010). A recent study on gender differences in the perception of sexual interest has identified bias as occurring on the level of decision (Perilloux & Kurzban, 2015). Following this and other modeling studies in SDT (e.g., Pleskac & Busemeyer, 2010), we assume that biases in forgiveness decisions occur at the decision level.

[ii] Forgiveness also has many intrapersonal consequences on the individual's physical and psychological health (e.g., Wade, Hoyt, Kidwell, & Worthington, 2014; Worthington, 2005). Nevertheless, the intrapersonal consequences are likely to be side effects of the interpersonal consequences rather than the main driver for the evolution of forgiveness.