Evolution and Human Behavior 38 (2017) 39–47

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

One-shot reciprocity under error management is unbiased and fragile

Jarid Zimmermann [a,*], Charles Efferson [b,*]

[a] Department of Economics, University of Cologne, Cologne, Germany
[b] Department of Economics, University of Zurich, Zurich, Switzerland

ARTICLE INFO

Article history:
Initial receipt 27 November 2015
Final revision received 9 June 2016

Keywords:
Cognitive biases
Error management theory
Evolution of cooperation
Anonymous one-shot games

ABSTRACT

The error management model of altruism in one-shot interactions provides an influential explanation for one of the most controversial behaviors in evolutionary social science. The model posits that one-shot altruism arises from a domain-specific cognitive bias that avoids the error of mistaking a long-term relationship for a one-shot interaction. One-shot altruism is thus, in an intriguingly paradoxical way, a form of reciprocity. We examine the logic behind this idea in detail. In its most general form the error management model is exceedingly flexible, and restrictions about the psychology of agents are necessary for selection to be well-defined. Once these restrictions are in place, selection is well defined, but it leads to behavior that is perfectly consistent with an unbiased rational benchmark. Thus, the evolution of one-shot reciprocity does not require an evoked cognitive bias based on repeated interactions and reputation. Moreover, in spite of its flexibility in terms of psychology, the error management model assumes that behavior is exceedingly rigid when individuals face a new interaction partner. Reciprocity can only take the form of tit-for-tat, and individuals cannot adjust their behavior in response to new information about the duration of a relationship. Zefferman (2014) showed that one-shot reciprocity does not reliably evolve if one relaxes the first restriction, and we show that the behavior does not reliably evolve if one relaxes the second restriction. Altogether, these theoretical results on one-shot reciprocity do not square well with experiments showing increased altruism in the presence of payoff-irrelevant stimuli that suggest others are watching.

1. Introduction

Error management theory (Haselton & Nettle, 2006) has provided a number of provocative hypotheses about the evolution of human behaviors in different domains. Error management mechanisms all share the assumption that asymmetric error costs in the ancestral past drove the genetic evolution of domain-specific mechanisms responsible for strong biases in behavior. These behavioral biases often persist and can thus be observed among contemporary humans. To recount perhaps the most well-known example (Haselton & Buss, 2000; Haselton & Nettle, 2006; Perriloux & Kurzban, 2015), consider a man in a bar. The man is curious about whether various women in the bar might have sex with him. The man can make two types of error. He can approach a woman who rejects him, or he can fail to approach a woman who would have responded positively had he approached her. The hypothesis proposes that for men, for most of human evolutionary history, missed mating opportunities were more costly than rejections. Because of this selective regime in the ancestral past, our representative man in a bar will show a strong tendency to approach women for sex. Though the details vary by decision-making domain, other error management hypotheses follow the same basic logic.

In general, one of the challenges in error management theory is determining whether a given bias in behavior involves an associated cognitive bias (Marshall, Trimmer, Houston, & McNamara, 2013; McKay & Efferson, 2010). If decision makers face asymmetric error costs and maximize expected utility or fitness, decision makers will exhibit behavioral biases even with Bayesian beliefs. The man in the bar, for example, might overestimate the woman's interest in him relative to what the evidence suggests, but this is not necessary. If the cost asymmetry is sufficiently extreme, he will approach the woman even if he has an exceedingly weak belief that he will be successful. Moreover, this is true even if he has integrated all relevant information in an unbiased and theoretically justifiable way, which means he has posterior beliefs equivalent to a Bayesian. The upshot is that biases in behavior under cost asymmetries may often be perfectly consistent with ordinary optimization and unbiased beliefs. Error management accounts, in contrast, emphasize the hypothesis that asymmetric error costs in a given domain in the ancestral past have generated adaptive domain-specific cognitive biases (Haselton & Nettle, 2006; Johnson, Blumstein, Fowler, & Haselton, 2013). Because error management often predicts the same behavior, for example, as maximizing expected utility under Bayesian beliefs, identifying effects specifically due to biased cognition can be difficult (Marshall et al., 2013; McKay & Efferson, 2010).

* Corresponding authors.
E-mail addresses: jarid.zimmermann@uni-koeln.de (J. Zimmermann), charles.efferson@econ.uzh.ch (C. Efferson).

http://dx.doi.org/10.1016/j.evolhumbehav.2016.06.005
1090-5138/

These challenges are especially relevant for the error management account of anonymous one-shot altruism. Anonymous one-shot altruism has been documented experimentally many times (Camerer, 2003), but providing an evolutionary explanation has proven to be a caustic and controversy-filled area of research (Henrich, 2004; Raihani & Bshary, 2015). One highly influential hypothesis argues that subjects who are altruistic in one-shot experiments are managing errors. Specifically, they are somehow treating the one-shot interaction as repeated because repeated interactions were a crucial part of social life for ancestral humans. As a result, humans have evolved cognitive biases that are extremely sensitive to signals suggesting one's prosocial reputation might be at stake. After observing such a signal, the relevant psychology can become active, and individuals behave prosocially in order to protect their reputations in implicitly repeated interactions (Burnham, 2013; Hagen & Hammerstein, 2006; Haley & Fessler, 2005; Raihani & Bshary, 2015, but see Zefferman, 2014). Anonymous one-shot altruism in this case is more appropriately thought of as one-shot reciprocity. Though the explicit structure of the social interaction is anonymous and one-shot, the implicit structure hinges on an evoked psychology involving repeated interactions, reciprocity, and reputation management.

Empirical studies of one-shot reciprocity have largely tested whether altruistic giving increases in the presence of payoff-irrelevant signals suggesting the subject is being observed. A typical signal, for example, is some kind of stylized face that appears in the background without explanation. Studies of this sort have produced a fascinating mix of findings both for and against the one-shot reciprocity hypothesis (Nettle et al., 2013; Sparks & Barclay, 2013), and we even have conflicting results from studies using exactly the same stylized face and similar experimental protocols (Fehr & Schneider, 2010; Haley & Fessler, 2005; Vogt, Efferson, Berger, & Fehr, 2015). Given recent studies showing that experimental results on social behavior do not replicate as often as we might like (Camerer et al., 2016; Open-Science-Collaboration, 2015; Shanks et al., 2013), we should approach mixed empirical results with some skepticism, and future experimental research on one-shot reciprocity would benefit greatly from pre-registration. Furthermore, even if results supporting one-shot reciprocity prove reliable in the long run, the appropriate evolutionary interpretation is far from obvious (Vogt et al., 2015).

Nonetheless, the fact remains that several experiments have found that payoff-irrelevant cues increase altruism, and the interpretation that reciprocity and reputation affect one-shot behavior has been extremely influential (Hagen & Hammerstein, 2006; Haley & Fessler, 2005; Raihani & Bshary, 2015). Understanding the evolution of psychological mechanisms that might support one-shot reciprocity is our objective in this paper. In particular, when payoff-irrelevant cues increase altruism, payoff-irrelevance and the minimal nature of the stimuli (e.g., Rigdon, Ishii, Watabe, & Kitayama, 2009) suggest that a cognitive bias could be at work. A recent evolutionary model has provided a theoretical foundation for this idea by demonstrating how past cost asymmetries could have selected for a psychology that supports one-shot reciprocity (Delton, Krasnow, Cosmides, & Tooby, 2011a). The model assumes that agents are uncertain about whether social interactions are one-shot or repeated. Agents receive cues that provide information about this critical distinction, and they then commit to a strategy. Agents can thus make two types of error. They can treat a one-shot interaction as repeated, or they can treat repeated interactions as one-shot. When agents are playing a social dilemma with potential efficiency gains, the latter error can be much more costly. This cost asymmetry can lead to the evolution of a cognitively biased tendency to cooperate "irrationally" (Delton et al., 2011a, p. 13336) in one-shot interactions.

The link to experiments showing that payoff-irrelevant cues can increase altruism is the following. In ancestral settings, cues of observability were conceivably important sources of information indicating repeated interactions and the need to manage one's reputation. The error management model of one-shot reciprocity shows that under appropriate conditions selection can render agents extremely sensitive to such cues. Specifically, a population can evolve so that agents behave prosocially even if available cues provide only weak evidence that interactions are repeated. This hypersensitivity is what the contemporary experimentalist identifies when she finds that a stylized face, for example, increases altruism in a setting that is otherwise described as one-shot. Experimental participants may or may not be aware of how they respond to a stylized face. Regardless, the error management account argues that ancestral cost asymmetries led to a cognitive bias exceedingly prone to yield altruistic behavior even when observable cues only weakly signal that one's prosocial reputation is at stake.

The error management model of one-shot reciprocity raises two fundamental questions, and we take up both in this paper. First, does one-shot reciprocity actually require a cognitive bias? As we have argued, cost asymmetries can generate tremendous biases in behavior without cognitive distortions. To identify a cognitive bias, one must have an unbiased benchmark. We provide exactly such a benchmark below and compare it to the error management model of one-shot reciprocity.

Second, regardless of the cognitive underpinnings, how robust is the evolution of one-shot reciprocity as a behavior? A growing body of theory has shown that the evolution of reciprocal strategies can be quite fragile (Boyd, 2006; Boyd & Lorberbaum, 1987; Le & Boyd, 2007; van Veelen, García, Rand, & Nowak, 2012; Wahl & Nowak, 1999; Zefferman, 2014). In particular, repeated interactions create many equilibria. As a result, a population can evolve such that any given reciprocal strategy, once common, will collapse and open the door for some other reciprocal strategy to invade. Reciprocal strategies come and go, and the population spends a conspicuous amount of time at the uncooperative equilibrium along the way (van Veelen et al., 2012). Without assortment, preventing this outcome usually requires one to arbitrarily exclude certain strategies from consideration, and this leads to model results that seem equivalently arbitrary (Henrich, 2004).

Importantly, if these problems exist when interactions are actually repeated, they could also exist for the implicitly repeated interactions of one-shot reciprocity. Zefferman (2014) has recently shown that this is indeed the case. We come to the same conclusion in a different way. Specifically, Zefferman (2014) allowed for various forms of reciprocity that are hesitant, repentant, and forgiving. We simply allow agents to update how they play as they receive new information about whether a relationship is one-shot or repeated. Intuitively, if error management agents choose defection or reciprocity given beliefs in the face of uncertainty (Delton et al., 2011a), we allow them to update their choice when uncertainty is removed. This is a minute and compelling modification of the error management model because it represents a simple extension of the logic inherent in the model itself.

Throughout the paper we show in detail how our approach relates to both Delton et al. (2011a) and Zefferman (2014). As a brief prelude, like Delton et al. (2011a) but unlike Zefferman (2014), we focus on proximate psychology. Accordingly, we consider a question ubiquitous in error management theory, the question of whether evolution leads to adaptive cognitive biases. In addition, like Zefferman (2014) but unlike Delton et al. (2011a), we find that intuitive and compelling modifications of the error management model dramatically reduce cost asymmetries and limit the evolution of one-shot reciprocity as a consequence.

2. Uncertainty and the cost asymmetry

Agents are randomly paired to play a simultaneous prisoner's dilemma with two possible actions. Cooperating brings a private cost, $c > 0$, and generates a benefit, $b > c$, for the other player. Defecting does not bring a cost or generate a benefit. Interactions can be repeated or one-shot, which we indicate with the variable R. R is a random variable with support {0,1}. This simply means that R takes each of the two values in the set {0,1} with some probability. Once R takes a specific