One-shot reciprocity under error management is unbiased and fragile*

Jarid Zimmermann[1] and Charles Efferson[2]

[1]Department of Economics, University of Cologne

[2]Department of Economics, University of Zurich

Running title: One-shot reciprocity is unbiased and fragile

Word count: c. 6500

*__Corresponding authors:__ Jarid Zimmermann (*jarid.zimmermann@uni-koeln.de*) and Charles Efferson (*charles.efferson@econ.uzh.ch*)

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.

Key words: cognitive biases, error management theory, evolution of cooperation, anonymous one-shot games

1 Introduction

Error management theory (Haselton and 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 and Buss, 2000; Haselton and Nettle, 2006; Perriloux and 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 (McKay and Efferson, 2010; Marshall *et al.*, 2013). 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 bi-

ases 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 and Nettle, 2006; Johnson *et al.*, 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 (McKay and Efferson, 2010; Marshall *et al.*, 2013).

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 and 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 (Haley and Fessler, 2005; Hagen and Hammerstein, 2006; Burnham, 2013; Raihani and 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 and Barclay, 2013), and we even have conflicting results from studies using exactly the same stylized face and similar experimental protocols (Haley and Fessler, 2005; Fehr and Schneider, 2010; Vogt *et al.*, 2015). Given recent studies showing that experimental results on social behavior do not replicate as often as we might like (Shanks *et al.*, 2013; Open-Science-Collaboration, 2015; Camerer *et al.*, 2016), 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 (Haley and Fessler, 2005; Hagen and Hammerstein, 2006; Raihani and 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 *et al.*, 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 *et al.*, 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 and Lorberbaum, 1987; Wahl and Nowak, 1999; Boyd, 2006; Le and Boyd, 2007; van Veelen *et al.*, 2012; 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 value from the support, we refer to the realization of R, which we denote generically as r.

$R = 0$ indicates a one-shot interaction, and thus $R = 0$ means the agents in a pair play a typical one-shot prisoner's dilemma. Defection is always better from the individual's perspective, but mutual cooperation is better for both agents than mutual defection. $R = 1$ indicates repeated interactions, which occurs with an ex ante probability of $P(R = 1) \in (0, 1)$. If $R = 1$, the continuation probability is $\omega \in (0, 1)$, and thus the expected number of interactions is $k = 1/(1 - \omega) > 1$. $R = 1$ is a standard repeated prisoner's dilemma in dyads (Axelrod and Hamilton, 1981). As ω increases, the expected number of repeated interactions increases, and this can increase the gains from mutual cooperation when reciprocators are paired.

Before paired agents play, each receives a private signal providing information about whether the game will be one-shot or repeated. For individual i, this signal is a random variable, S_i, with the real numbers as a support and realizations s_i. Importantly, a realized signal leaves an agent with some degree of uncertainty. How agents respond to this posterior uncertainty is a key question of interest. The cumulative distribution functions for signals are denoted $F(s_i \mid R = r)$. Private signals are informative but noisy. Specifically, in keeping with the error management model (Delton *et al.*, 2011a), we assume that the $F(\cdot \mid R = r)$ are continuous and strictly monotonic, and for all finite s_i they satisfy $F(s_i \mid R = 0) > F(s_i \mid R = 1)$. The latter condition ensures that relatively small values of s_i provide relatively strong evidence that a pair will have a one-shot interaction, while relatively large values of s_i provide relatively strong evidence that a pair will have a relationship with $k > 1$ expected interactions.

After receiving a signal, an agent commits to one of two options, either always defect

(D) or tit-for-tat (T). We call these options "sub-game strategies" because they refer to the strategies available after an agent receives a private signal and reaches the associated sub-game. The set of sub-game strategies is thus $\{\mathrm{D}, \mathrm{T}\}$. To denote sub-game strategies for i, X_i is a random variable with support $\{\mathrm{D}, \mathrm{T}\}$ and realizations x_i.

When ω is close to one, a relationship in a dyad is long-lasting in the sense that ex ante the expected number of interactions is large. In this case a radical cost asymmetry results. Specifically, if interactions are repeated and i commits to defection when paired with a partner playing tit-for-tat, i will receive b in lieu of $k(b-c)$. This error involves a cost of $k(b-c)-b$, which becomes arbitrarily large as $\omega \to 1^-$. If, however, the interaction is one-shot and i commits to tit-for-tat when paired with a partner playing unconditional defection, i will receive $-b$ instead of 0. This error involves a cost of $0-(-b)$, which is constant. By extension, the asymmetry in expected error costs becomes arbitrarily large as the length of relationships increases when interactions are repeated. Intuitively, defecting on someone you will interact with over and over again can be boundlessly costly, while cooperating with someone you will never see again cannot.

3 Degrees of freedom under error management

We begin by examining the flexibility of decision making under the error management model. In doing so, we focus on the psychological basis of behavior because the error management model specifies the psychology of decision making in a way that has potential implications for evolution. An analysis of population dynamics comes later.

Selection in the error management model is based on fitness values that depend on phenotypes. In general, if an error management model has many degrees of freedom, it will admit different psychological pathways for producing a given phenotype (McKay and Efferson, 2010). This implies, in turn, the possibility that a selected phenotype can be produced in many different ways. In such cases, selection on psychology will not be well defined, and random drift will play an outsize role in the evolution of cognition. As

we now explain, the error management model of one-shot reciprocity exhibits this kind of excess flexibility.

Specifically, each agent processes information, which leads to some belief about whether a relationship will involve repeated interactions, and given a belief each agent has some motivation to choose tit-for-tat. The flexibility of the error management model stems from the fact that a single phenotype is often consistent with multiple combinations of information processing and motivation. Consequently, even if selection favors a unique phenotype, it may not favor a unique form of cognition. In practice, however, the error management model of one-shot reciprocity imposes restrictions that eliminate this possibility. The result is two different versions of the model that represent two different views of how cognition can evolve.

To see this, let agent i have a signal threshold, E_i. If a signal is below or equal to the threshold ($s_i \leq E_i$), the agent plays tit-for-tat with probability $\alpha_i \in [0, 1]$ and always defect with probability $1 - \alpha_i$. If the signal is above the signal threshold ($s_i > E_i$), the agent plays tit-for-tat with probability $\beta_i \in [0, 1]$ and always defect with probability $1-\beta_i$. Altogether, the quantities E_i, α_i, and β_i specify the psychology of decision making under the error management model. E_i represents information processing, while α_i and β_i represent the agent's motivation to play tit-for-tat given a processed signal.

The phenotype of an agent comprises both the probability of playing tit-for-tat if an interaction is one-shot (i.e. $P(X_i = \mathrm{T} \mid R = 0)$) and the probability of playing tit-for-tat if interactions are repeated (i.e. $P(X_i = \mathrm{T} \mid R = 1)$). Many phenotypes available under the error management are consistent with multiple combinations of α_i, β_i, and E_i. Consider, for example, any phenotype that plays tit-for-tat with a constant probability regardless of whether or not interactions are repeated. If $\alpha_i = \beta_i$, such a phenotype results, and the signal threshold in these cases can take any value whatsoever. Because the signal threshold can take any value, the error management model can generate the phenotype in question in an infinite number of ways. Phenotypes that always choose

T with the same probability are important because they include classic strategies as special cases. Specifically, $\alpha_i = \beta_i = 0$ means the agent always defects, and $\alpha_i = \beta_i = 1$ means the agent always chooses tit-for-tat. Moreover, we show (Fig. 2 and electronic supplementary material, § 2.3) that one or both of these strategies can be evolutionarily stable when the error management model is restricted to reduce the importance of drift. For present purposes, the important point is that any phenotype given by $\alpha_i = \beta_i$ can be produced in an infinite number of ways.

Apart from phenotypes given by $\alpha_i = \beta_i$, other phenotypes might also be consistent with multiple underlying psychologies. The details depend on both the phenotype being considered and the probability distributions for private signals (electronic supplementary material, § 1). In Fig. 1A, we show an example of the important case involving normally distributed signals (e.g. Delton *et al.*, 2011a). As Fig. 1A shows, the space of phenotypes allowed by the error management model is a strict subset of all possible phenotypes, and we show a sample of curves within this subset. Importantly, we show only a sample of curves because doing so clearly reveals that the curves routinely overlap and intersect each other. Where two or more curves overlap or intersect, multiple selectively neutral psychologies (i.e. combinations of α_i, β_i, and E_i) can generate the phenotypes in question. Consequently, in terms of cognition, drift can be an extremely important evolutionary force. We suspect that the error management model, for this reason, eliminates excess flexibility by providing additional structure in two different ways.

Specifically, one can fix E_i exogenously for all agents and allow the distribution of α_i and β_i values to evolve (Delton *et al.*, 2011a, Model 1). This restriction ensures that the model can generate a given phenotype in only one way (electronic supplementary material, § 1), and drift will be less important than in the full three-dimensional model. The (α_i, β_i) cognitive architecture allows for strategies that seem especially suggestive of a cognitive bias. The reason is because positive α_i values mean that agents play tit-for-tat given a "one-shot belief" (Delton *et al.*, 2011a, p. 13337), which intuitively

suggests some kind of distortion in the way agents process and respond to information. We will see momentarily if this intuition is correct.

Alternatively, one can set $\alpha_i = 0$ and $\beta_i = 1$ for all agents, and allow the distribution of E_i values to evolve (Delton *et al.*, 2011a, Model 2). This restriction also ensures that the model can generate a given phenotype in only one way (electronic supplementary material, § 1). Fig. 1B shows how these two restrictions affect the space of admissible phenotypes.

Having two versions of the error management model solves a potential problem. By focusing separately on the evolution of either motivation in the (α_i, β_i) model or information processing in the signal threshold (E_i) model, one can reduce the importance of drift. Crucially, however, the two versions of the error management model are effectively redundant in an important sense. Evolution under the (α_i, β_i) architecture leads to three possible outcomes at the population level (Fig. 2 and electronic supplementary material, § 2.3 and § 2.5). Selection leads to either (i) $\overline{\alpha} = 0$ and $\overline{\beta} = 0$, (ii) $\overline{\alpha} = 0$ and $\overline{\beta} = 1$, or (iii) $\overline{\alpha} = 1$ and $\overline{\beta} = 1$. These three steady states are all available under the signal threshold model provided one allows for limiting phenotypes as signal thresholds become infinitely large or small. Although the (α_i, β_i) architecture allows for phenotypes that the signal threshold architecture cannot produce in general (Fig. 1B), the phenotypes unique to the (α_i, β_i) architecture do not seem to arise under selection. In terms of evolutionary outcomes, the (α_i, β_i) architecture thus adds little if anything to a simpler cognitive architecture based only on signal thresholds.

In spite of this redundancy, our first task is to see if the error management model supports the evolution of cognitive biases, and to do so we must check both versions of the model. Accordingly, in the next section we provide rational benchmarks for both the (α_i, β_i) architecture and the signal threshold architecture.

4 Does error management support a cognitive bias?

To provide rational benchmarks, we work with both versions of the error management model. In each case, we identify the Bayesian Nash equilibria of the game in terms of the appropriate cognitive architecture (electronic supplementary material, § 2). For the first cognitive architecture, the signal threshold is exogenously fixed, and strategies are defined in terms of α_i and β_i values. For the second cognitive architecture, α_i and β_i values are exogenously fixed, and strategies are defined in terms of the signal threshold. For comparison, we use evolutionary game theory to find equilibrium strategies for the associated error management models. First, we fix the signal threshold and let the distribution over α_i and β_i values evolve (Delton *et al.*, 2011a, Model 1). Second, we fix α_i and β_i and let the distribution of signal threshold values evolve (Delton *et al.*, 2011a, Model 2).

For both types of cognitive architecture, we show analytically that the evolutionary stable strategies for the error management model (electronic supplementary material, § 2.3 and § 2.4) match symmetric Bayesian Nash equilibria (electronic supplementary material, § 2.1 and § 2.2). To provide more general results, we also used agent-based simulations to simulate evolutionary dynamics for the error management models under both cognitive architectures (electronic supplementary material, § 2.5). Unlike an analysis of evolutionary stable strategies, this approach allows for arbitrary distributions of strategies in the population. Fig. 2 compares the error management model and the rational benchmark under the (α_i, β_i) architecture for a wide range of parameter values. Fig. 3 shows the analogous comparison under the signal threshold architecture. In both cases the correspondence is striking. The typical outcome is equivalence between error management agents and unbiased optimizers. When agents face uncertainty about whether a new relationship will involve a one-shot interaction or repeated interactions, error management and a rational benchmark are effectively the same.

These results demonstrate that evolutionary stable strategies are a refinement of Nash

equilibria (Weibull, 1995; Samuelson, 1998). The more important message relates to one-shot reciprocity specifically and error management theory more broadly. Simply put, we have no reason to conclude that the error management model supports a domain-specific bias in how agents process and respond to signals related to reputation and repeated interactions. We can just as well say that the model supports one-shot reciprocity via run-of-the-mill optimization given Bayesian beliefs. Discriminating between these two generic possibilities remains one of the principal challenges of error management theory (McKay and Efferson, 2010; Marshall *et al.*, 2013).

5 Is one-shot reciprocity robust?

In this section, we examine the extent to which the evolution of one-shot reciprocity is robust to a simple and intuitive modification of the error management model. We restrict attention to the signal threshold model. We do so because, as discussed above, the two versions of the error management model are effectively redundant in the sense that the (α_i, β_i) architecture always leads to the evolution of phenotypes available under the signal threshold architecture.

The evolution of one-shot reciprocity under error management involves two key restrictions (Delton *et al.*, 2011a). First, as a sub-game strategy, each agent chooses either unconditional defection or tit-for-tat. Second, each agent has some sensitivity to private signals, a sensitivity summarized by E_i, that commits the agent fully to a sub-game strategy *before* the agent's first interaction with her partner.

Zefferman (2014) relaxed the first restriction but maintained the second. In his model, each agent must fully commit before the initial interaction with a partner, but the set of sub-game strategies includes options that allow agents to repent for their past defections, to forgive the past defections of others, and to hesitate before cooperating for the first time. With these modifications to the set of sub-game strategies, one-shot reciprocity does not evolve reliably (Zefferman, 2014).

We maintain the first restriction and relax the second. Specifically, we maintain always defecting and tit-for-tat as the only two sub-game strategies. However, if an agent reaches the second interaction with a specific partner, all uncertainty is removed. The agent at this point knows with certainty that interactions are repeated. We thus introduce a responsive strategy that allows the agent to update her sub-game strategy[1] at this point in time. This is the only change we introduce relative to the error management model of one-shot reciprocity.

Specifically, let G_i (electronic supplementary material, § 3) indicate whether i is unresponsive (U) or responsive (R). If $G_i = \mathrm{U}$, i is unresponsive. If i's realized signal is above her threshold, i plays tit-for-tat. Otherwise, i defects unconditionally. Unresponsive types do not update their sub-game strategies if they reach the second interaction with a partner. In particular, this means unresponsive agents can be locked into always defecting even when they are certain interactions are repeated. If $G_i = \mathrm{R}$, in contrast, i is responsive. Responsive types are like unresponsive types for the first interaction. If, however, a responsive agent reaches a second interaction with a given partner, she responds to the fact that she no longer faces uncertainty about whether interactions are repeated and plays tit-for-tat.

Importantly, one can in some cases translate between our approach and an approach that modifies the set of sub-game strategies (e.g. Zefferman, 2014). For example, when paired with certain types of partner, a responsive agent is behaviorally equivalent to an agent who pre-commits fully before play, who chooses tit-for-two-tats for a signal above her threshold, and who chooses hesitant tit-for-tat otherwise. Our model compares this type of agent to the original error management agents (i.e. Delton *et al.*, 2011a). Although Zefferman (2014) analyzed two models involving hesitant tit-for-tat or tit-for-two-tats, he did not analyze the particular combination we consider. Our model, in effect, represents the minimum conceivable change to the error management model.

[1] We retain the term sub-game strategy to refer to strategies from the set $\{\mathrm{D}, \mathrm{T}\}$, but we now additionally use the term to refer to the sub-game reached after a single interaction with a given partner.

In addition, in a reply to McNally and Tanner (2011), Delton *et al.* (2011b) argue that hesitant strategies are largely irrelevant for understanding one-shot reciprocity. Our comparison with Zefferman (2014), however, shows that hesitant strategies are a natural result of allowing error management agents to respond to all relevant information about the duration of a social relationship. Responsive error managers respond to the uncertainty that obtains after observing private signals, as in the original error management model. They additionally respond to the certainty that obtains when they reach a second interaction with someone. Accepting the first response but rejecting the second is tantamount to positing an error manager who is sensitive to cues of observability but cannot look a partner in the face and recognize that they have interacted previously. Accepting both responses, however, effectively means accepting hesitant strategies. For this reason, in spite of claims to the contrary, hesitant strategies are a natural outgrowth of error management.

We present an analysis of unresponsive and responsive types in the electronic supplementary material (§ 3). The key result shows that unresponsive types should be extremely vulnerable to invasion by responsive types in precisely those cases that matter most. Specifically, if we exclude responsive types from consideration, long-lasting relationships under repeated interactions, as implied by large values of ω, lead to the evolution of low values of E_i. Low values of E_i allow agents to avoid the especially costly error of defecting when paired with a reciprocating partner for many interactions. This is the basic error management result (electronic supplementary material, § 3.1). In a population of unresponsive types, however, a mutant responsive type has strictly higher expected fitness (electronic supplementary material, § 3.2) than the resident unresponsive type if and only if $(\omega b - c)(1 - F(E_i \mid R = 1))/(1 - \omega^2) > c$. This condition is satisfied if the continuation probability, ω, is sufficiently large and the signal threshold, E_i, is sufficiently small. In effect, large values of ω have two countervailing effects. They support the evolution of one-shot reciprocity by unresponsive agents when responsive types are

excluded. They also, however, render a population of unresponsive types increasingly vulnerable to invasion when responsive types are allowed.

For more general results, we used agent-based simulations to simulate evolutionary dynamics in a model that allows for unresponsive agents, responsive agents, and agents who always defect without conditioning on their private signals (electronic supplementary material, § 3.3). Our simulations show that, even for relationships that last a long time under repeated interactions, one-shot reciprocity is rarely observed. In particular, for a wide range of parameter combinations favoring cooperation, responsive agents displace both unresponsive agents and unconditional defectors (e.g. Fig. 4A). Moreover, signal thresholds for responsive types evolve to relatively high values (e.g. Fig. 4B), which ensures that agents rarely cooperate on the first interaction. Altogether, the combination of responsiveness and high signal thresholds means that agents can enjoy the mutual gains of prolonged cooperation once they know they face repeated interactions. Doing so, however, does not require agents to run a significant risk of exploitation in one-shot interactions. The error management model, in contrast, forces agents to risk one-shot exploitation in order to enjoy the gains of long-run cooperation. This is what drives the evolution of one-shot reciprocity, but the result does not persist when agents can respond to both the uncertainty that always holds before play and to the certainty that sometimes arises during play.

6 Discussion and conclusion

Our results suggest potential challenges for the error management model of one-shot reciprocity and for error management theory in general. With respect to error management theory in general, we have demonstrated that asymmetric costs lead to the evolution of behavioral biases perfectly consistent with unbiased optimization. Accordingly, one cannot offer biases in behavior, biases like men stubbornly approaching women for sex or reciprocators risking one-shot exploitation, as evidence for adaptive domain-specific

cognitive biases.

McKay and Efferson (2010) have offered cognitive constraints as one possible route to the evolution of genuine error management biases. Our results clarify what this mechanism would require. In Fig. 5, we have translated evolutionary outcomes from both versions of the error management model (Figs. 2 and Fig. 3) into phenotype space. Fig. 5 shows that the two versions of the model do not always lead to the evolution of identical phenotypes. As explained above, the (α_i, β_i) model always leads the evolution of phenotypes available to the signal treshold model (Fig. 2), but the opposite is not always true. When comparing evolutionary outcomes, disparities between the two models are typically small, but they can occur precisely because the two cognitive architectures allow for different phenotypic possibilities (Fig. 1B).

One could potentially describe such disparities as situations in which constraints lead to bias. To do so, however, we must choose one architecture and the associated set of possible phenotypes as a reference. We judge "bias" relative to the benchmarks for this reference. We must then, however, force evolution to take place under the *other* architecture. If evolution under the latter architecture cannot lead to a benchmark outcome under the reference architecture, we can invoke the notion of constraints producing bias. Why, though, would we pick one architecture as a reference, only to require that selection and evolution take place under a different architecture? As we have shown, restricting attention to one architecture at a time means that evolution leads to outcomes consistent with a rational benchmark.

If cognition is sufficiently flexible, pinning down the notion of an adaptive cognitive bias can be challenging for other reasons. If selection favors a phenotype consistent with multiple psychologies, what then is an adaptive cognitive bias? Imagine, for example, that ancestral men looking for sex received signals from women and processed the associated information to infer the interest these women had in sex. In addition, given a processed signal, ancestral men had some motivation for sex. Missed mating opportuni-

ties were costly, but rejections were not. Consequently, unless a woman signaled extreme distaste, approaching women was usually the best choice.

How can selection produce a bias in favor of men approaching women for sex? In general, many options might be possible. A man could process information to over-infer a woman's interest while having a relatively weak motivation for sex that exactly offsets his biased inferences in the optimal way. Just as good, however, might be a man who under-infers a woman's interest while having a strong motivation for sex that compensates optimally. Finally, and perhaps once again just as good, a man could make unbiased inferences, implying posterior beliefs equivalent to a Bayesian, and given these beliefs he has an optimal motivation for sex. If a situation like this obtains, a specific phenotype is optimal, and an array of psychologies can generate the phenotype in question. Some forms of adaptive cognition may indeed be biased. A specific bias, however, would be neither more nor less adaptive than any other cognition, biased or otherwise, that produces the same phenotype. We do not know how plausible this example might be, but we would like to propose that high-dimensional cognition will often allow multiple routes to a given phenotype.

Apart from potential implications for error management theory, our results demonstrate the theoretical fragility of one-shot reciprocity. Several experiments have demonstrated that payoff-irrelevant face-like stimuli can increase altruism in anonymous one-shot settings (Nettle *et al.*, 2013; Sparks and Barclay, 2013). This finding places us in the middle of an interpretive puzzle. The stimuli in question do not affect payoffs, and the stimuli suggest, in a quite specific way, scrutiny by others. For these reasons, a compelling interpretation of experimental results is to posit a domain-specific cognitive bias based on repeated interactions, reciprocity, and reputation management. In light of recent modeling efforts, we have no obvious articulation of what this means.

In behavioral terms, uncertainty about the duration of a social relationship can lead to the evolution of altruism in one-shot interactions in order to avoid the large oppor-

tunity costs associated with treating repeated interactions as one-shot (Delton *et al.*, 2011a). Our results, however, in conjunction with those of Zefferman (2014), show that key restrictions have to be in place for this kind of evolutionary process to work. Zefferman (2014) shows that the set of sub-game strategies has to be appropriately restricted. We show that the agent's ability to respond to new information has to be appropriately restricted. Relax either restriction, and one-shot reciprocity under error management does not reliably evolve. Moreover, we show that, even if the necessary restrictions hold, the evolution of one-shot reciprocity is perfectly consistent with unbiased optimization. Consequently, observing one-shot reciprocity provides an argument neither for nor against a cognitive bias.

This leaves at least three possibilities for making sense of experimental results on one-shot reciprocity. One possibility is that one-shot reciprocity evolved under uncertainty and asymmetric error costs in the ancestral past. In the ancestral past, the behavior that evolved was perfectly consistent with unbiased optimization, but the cognition involved was domain-specific. In a contemporary one-shot experiment with a face in the background (e.g. Haley and Fessler, 2005), observed altruism is due to this vestigial, domain-specific psychology. In particular, observed altruism in the present is not consistent with unbiased optimization because the ancestral cognition prevents an optimal response to the explicit material incentives in a contemporary one-shot experiment. A second possibility is that the relevant cognition was and is unbiased, and subjects who see payoff-irrelevant faces in contemporary experiments actually update their beliefs about whether the game is one-shot. This explanation would be bad news for experimentalists because it would suggest that subjects routinely discredit the explicit incentive structure stipulated in an experiment. Finally, in light of the recent replication crisis in experimental social science (Open-Science-Collaboration, 2015), a third possibility may simply be that one-shot reciprocity is not a robust experimental result.

In any case, whatever one's favorite interpretation of empirical studies, uncertainty

and asymmetric error costs only lead in theory to the evolution of one-shot reciprocity under extreme restrictions. One has to assume that people are astonishingly narrow when dealing with new interaction partners. Specifically, they consider very few options when initially deciding how to play. They also do not update how they play as they acquire new information, and they know they will not update how they play as they acquire new information. These are strong assumptions, but relaxing one or more of them attenuates expected cost asymmetries dramatically. If a person sees sufficient scope for recuperating from an initial defection once she knows a relationship will last, one-shot reciprocity does not pay.

Critically, we would like to emphasize the following. We are not arguing that prosocial behavior in one-shot interactions is completely unrelated to reciprocity. We are also not arguing that cues of observability in a one-shot experiment (e.g. Haley and Fessler, 2005) have nothing to do with reputational concerns. The number of studies showing effects from payoff-irrelevant faces suggests that, for the moment at least, one-shot reciprocity remains a viable hypothesis. We are, however, arguing that uncertainty and associated asymmetric error costs do not provide a robust theoretical basis for the evolution of a cognitive bias that supports one-shot reciprocity. When the behavior evolves, it is unbiased. With a minimum degree of realism, the behavior does not evolve.

This brings us to our interpretive puzzle. Aside from error management, we have other explanations for one-shot altruism (e.g. Richerson *et al.*, 2015). Whatever the role these other mechanisms may or may not play, none of them to our knowledge imply that a person should become more altruistic because a stylized face of no material consequence suddenly appears without explanation in one's surroundings. In spite of several studies showing such an effect, the only candidate explanation for this finding, namely the error management model of one-shot reciprocity, also does not imply this behavioral regularity. This is the paradox arising from the unbiased fragility of one-shot reciprocity.

7 Acknowledgements

We would like to thank two anonymous referees for several helpful comments on an earlier draft of the paper.

References

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, **211**(27 March), 1390–1396.

Boyd, R. (2006). Reciprocity: you have to think different. *Journal of Evolutionary Biology*, **19**(5), 1380–1382.

Boyd, R. and Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game. *Nature*, **327**, 58–59.

Burnham, T. C. (2013). Toward a neo-darwinian synthesis of neoclassical and behavioral economics. *Journal of Economic Behavior & Organization*, **90, Supplement**(0), S113 – S127. Evolution as a General Theoretical Framework for Economics and Public Policy.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*.

Delton, A. W., Krasnow, M. M., Cosmides, L., and Tooby, J. (2011a). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, **108**(32), 13335–13340.

Delton, A. W., Krasnow, M. M., Cosmides, L., and Tooby, J. (2011b). Reply to mcnally and tanner: Generosity evolves when cooperative decisions must be made under uncertainty. *Proceedings of the National Academy of Sciences*, **108**(44), E972.

Fehr, E. and Schneider, F. (2010). Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity? *Proceedings of the Royal Society B: Biological Sciences*, **277**(1686), 1315–1323.

Hagen, E. H. and Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*, **69**(3), 339 – 348.

Haley, K. J. and Fessler, D. M. T. (2005). Nobody's watching?: subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, **26**(3), 245–256.

Haselton, M. G. and Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, **78**(1), 81–91.

Haselton, M. G. and Nettle, D. (2006). The paranoid optimist: an integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, **10**(1), 47–66.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization*, **53**(1), 3–35.

Johnson, D. D., Blumstein, D. T., Fowler, J. H., and Haselton, M. G. (2013). The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*, **28**(8), 474 – 481.

Le, S. and Boyd, R. (2007). Evolutionary dynamics of the continuous iterated prisoner's dilemma. *Journal of Theoretical Biology*, **245**(2), 258 – 267.

Marshall, J. A., Trimmer, P. C., Houston, A. I., and McNamara, J. M. (2013). On evolutionary explanations of cognitive biases. *Trends in Ecology & Evolution*, **28**(8), 469 – 473.

McKay, R. and Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior*, **31**(5), 309 – 319.

McNally, L. and Tanner, C. J. (2011). Flexible strategies, forgiveness, and the evolution of generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, **108**(44), E971.

Nettle, D., Harper, Z., Kidson, A., Stone, R., Penton-Voak, I. S., and Bateson, M. (2013). The watching eyes effect in the Dictator Game: it's not how much you give, it's being seen to give something. *Evolution and Human Behavior*, **34**(1), 35–40.

Open-Science-Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, **349**(6251).

Perriloux, C. and Kurzban, R. (2015). *Do* Men Overperceive Women's Sexual Interest? *Psychological Science*, **26**(1), 70–77.

Raihani, N. J. and Bshary, R. (2015). Why humans might help strangers. *Frontiers in Behavioral Neuroscience*, **9**(39).

Richerson, P. J., Baldini, R., Bell, A., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E., Narr, N., Newson, L., Ross, C., Smaldino, P., Waring, T., and Zefferman, M. (2015). Cultural group selection plays an essential role in explaining human cooperation: a sketch of the evidence. *Behavioral and Brain Sciences*, ***In press***.

Rigdon, M., Ishii, K., Watabe, M., and Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology*, **30**(3), 358–367.

Samuelson, L. (1998). *Evolutionary games and equilibrium selection*, volume 1. MIT Press.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., and Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, **8**(4), 1–10.

Sparks, A. and Barclay, P. (2013). Eye images increase generosity, but not for long: the limited effect of a false cue. *Evolution and Human Behavior*, **34**(5), 317–322.

van Veelen, M., García, J., Rand, D. G., and Nowak, M. A. (2012). Direct reciprocity in structured populations. *Proceedings of the National Academy of Sciences*, **109**(25), 9929–9934.

Vogt, S., Efferson, C., Berger, J., and Fehr, E. (2015). Eye spots do not increase altruism in children. *Evolution and Human Behavior*, **36**(3), 224 – 231.

Wahl, L. M. and Nowak, M. A. (1999). The continuous prisoner's dilemma: II. linear reactive strategies with noise. *Journal of Theoretical Biology*, **200**(3), 323 – 338.

Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge: The MIT Press.

Zefferman, M. R. (2014). Direct reciprocity under uncertainty does not explain one-shot cooperation, but demonstrates the benefits of a norm psychology. *Evolution and Human Behavior*, **35**(5), 358 – 367.

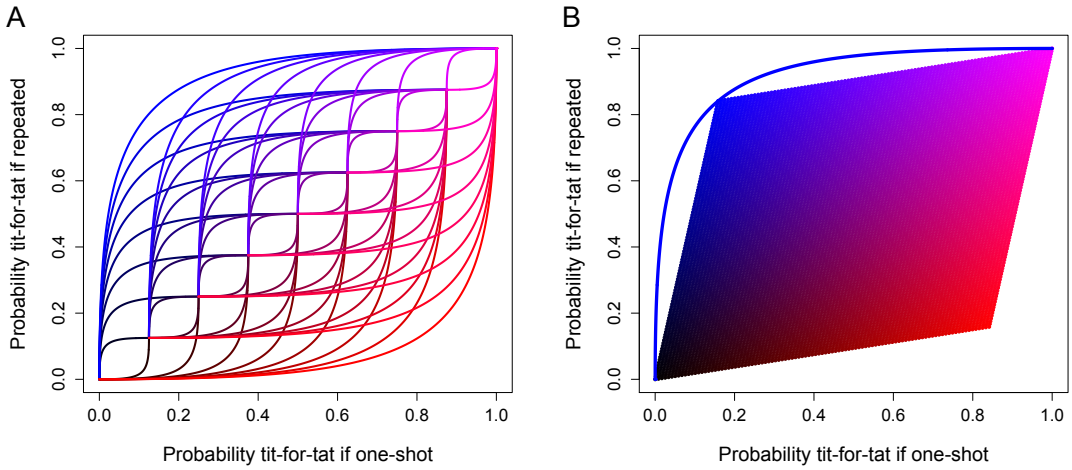

Figure 1: Possible phenotypes under different cognitive architectures. A phenotype consists of two probabilities, the probability of choosing tit-for-tat when an interaction is one-shot ($P(X_i = \mathrm{T} \mid R = 0)$) and the probability of choosing tit-for-tat when interactions are repeated ($P(X_i = \mathrm{T} \mid R = 1)$). To generate this figure, signals were normally distributed (e.g. Delton *et al.*, 2011a) with means at -1 ($R = 0$) and 1 ($R = 1$) and standard deviations of 1. Panel **A** shows a small selection of curves from the set of possible phenotypes for the full three-dimensional error management model. We generated the curves by fixing values for α_i and β_i, which are the motivations to choose T for low signals and high signals respectively, and then letting the signal threshold vary continuously. Pure blue means that $\alpha_i = 0$ and $\beta_i = 1$, while pure red means that $\alpha_i = 1$ and $\beta_i = 0$. Intermediate colors reflect intermediate values of α_i and β_i. The overlapping and intersecting curves in **A** show that, for many phenotypes, the full error management model can produce the phenotype in question via multiple combinations of α_i, β_i, and the signal threshold. Panel **B** shows what happens when restrictions are put in place to eliminate this flexibility and by extension the importance of drift. The polygon shows the set of possible phenotypes for the (α_i, β_i) architecture when the signal threshold is fixed at 0. The blue line shows the set of possible phenotypes for the signal treshold architecture $(\alpha_i = 0, \beta_i = 1)$.

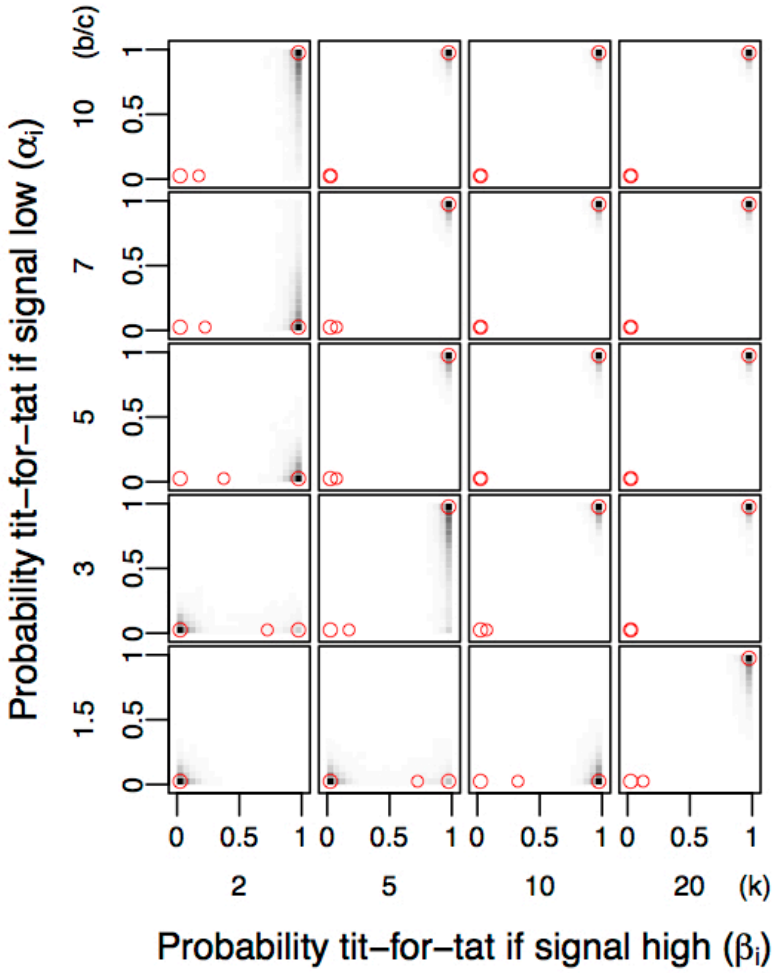

Figure 2: Equivalence between a rational benchmark and the evolution of an error management psychology under the (α_i, β_i) architecture. Each cell shows the space of α_i and β_i values, where α_i represents the motivation to play tit-for-tat when a private signal suggests a one-shot interaction, and β_i represents the motivation to play tit-for-tat when a private signal suggests repeated interactions. The cells differ by values of $k = 1/(1-\omega)$, which is the expected number of interactions when interactions are repeated, and b/c, which is the benefit-to-cost ratio in the prisoner's dilemma. Red circles show Bayesian Nash equilibria, which serve as unbiased rational benchmarks. Gray-scale histograms show the steady-state distributions of strategies from evolutionary simulations of the error management model. The overlap between the rational benchmarks and the error management model shows we cannot conclude that error management supports the evolution of a domain-specific cognitive bias associated with one-shot reciprocity.

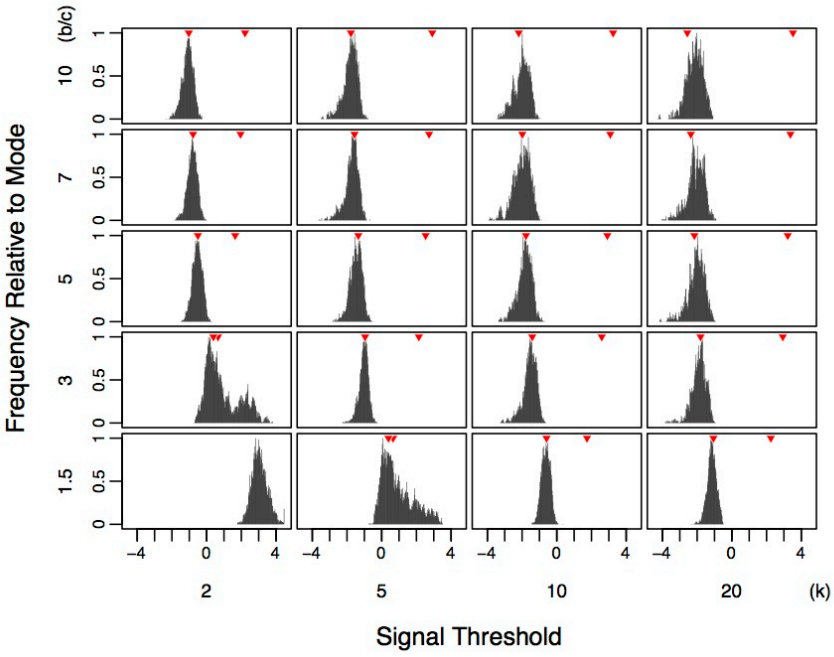

Figure 3: Equivalence between a rational benchmark and the evolution of an error management psychology under the signal threshold architecture. Signal thresholds separate the space of signals into two parts. If an agent receives a signal at or below the threshold, the agent plays always defect. If an agent receives a signal above the threshold, the agent plays tit-for-tat. We specified signal distributions so that an agent with a signal threshold at zero plays always defect and tit-for-tat with equal probabilities (electronic supplementary material, § 3.3.1). Each cell shows a histogram indicating the distribution of strategies after simulating evolution under the error management model. Red triangles show Bayesian Nash equilibria, which serve as unbiased rational benchmarks. The cells differ by values of $k = 1/(1-\omega)$, which is the expected number of interactions when interactions are repeated, and b/c, which is the benefit-to-cost ratio in the prisoner's dilemma. The overlap between the rational benchmarks and the error management model shows we cannot conclude that error management supports the evolution of a domain-specific cognitive bias associated with one-shot reciprocity. The lower left cell has only one equilibrium in which agents always defect, which is analogous to an infinitely large threshold. The associated histogram shows the distribution of signal threshold values after 25,000 generations of directional selection.

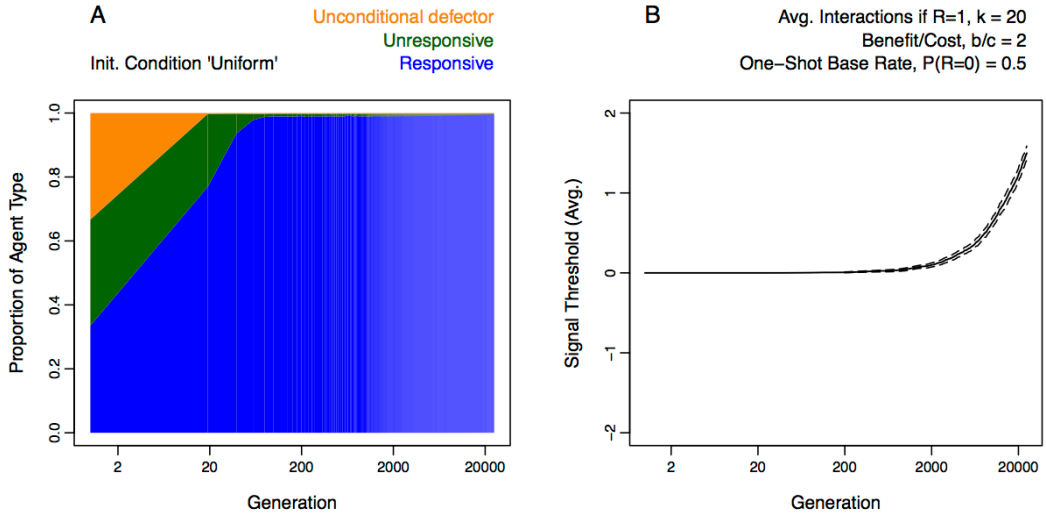

Figure 4: The fragility of one-shot reciprocity. The evolution of one-shot reciprocity hinges on the restriction that agents are unresponsive, which means they must commit fully to always defecting or tit-for-tat before an initial interaction with a new partner. Here we introduce responsive types. Responsive agents can update their choice to tit-for-tat once they reach a second interaction with a partner and thus know interactions are repeated. Panel **A** shows the simulated evolution of strategies with a uniform initial mix of responsive types, unresponsive types, and unconditional defectors. Responsive types quickly become the most common type. Panel **B** shows the associated evolution of the average signal threshold in the population with 95% bootstrapped confidence intervals. Higher values imply that one-shot reciprocity is rare. Signal thresholds become increasingly large as responsive types take over, which means that one-shot reciprocity almost never occurs. For this example, the expected number of interactions under repeated interactions is $k = 20$, while the benefit-to-cost ratio is $b/c = 2$. The ex ante probability of repeated interaction is $P(R = 1) = 0.5$. Additional details and further analyses are available in the electronic supplementary material (Figure S4 and § 3).

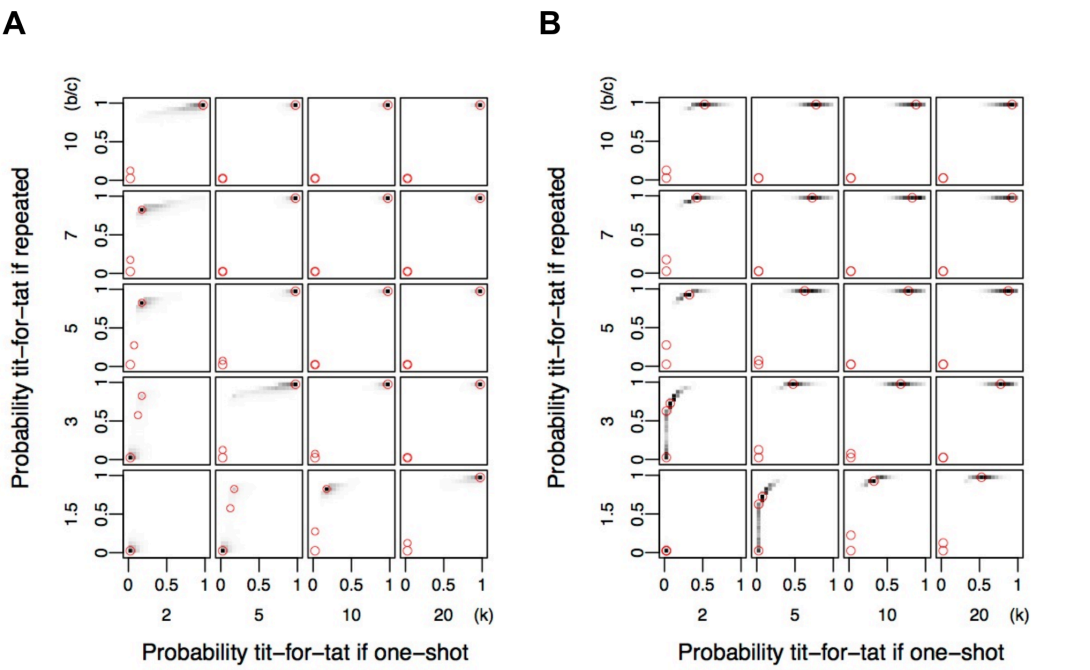

Figure 5: A cognitive bias requires one cognitive architecture for reference and another cognitive architecture for evolution. The figure shows results from Fig. 2 and Fig. 3 in phenotype space. A phenotype consists of two probabilities, the probability of choosing tit-for-tat when an interaction is one-shot ($P(X_i = \mathrm{T} \mid R = 0)$) and the probability of choosing tit-for-tat when interactions are repeated ($P(X_i = \mathrm{T} \mid R = 1)$). Panel **A** shows the (α_i, β_i) architecture (Fig. 2), and **B** shows the signal threshold architecture (Fig. 3). The cells differ by values of $k = 1/(1-\omega)$, which is the expected number of interactions when interactions are repeated, and b/c, which is the benefit-to-cost ratio in the prisoner's dilemma. Red circles show Bayesian Nash equilibria, which serve as unbiased rational benchmarks given an architecture. Gray-scale histograms show the steady-state distributions from evolutionary simulations of the error management models. Because the two architectures do not allow for the same possible phenotypes (Fig. 1B), the rational benchmarks and evolutionary outcomes are not always identical in both cases, e.g. the lower right cells of **A** and **B**. Inferring a bias based on such disparities requires us to choose one architecture as a reference while assuming that evolution occurs with respect to the other architecture.