Selective Advantages of Guilt

Sarita Rosenstock and Cailin O’Connor

Abstract
Using results from evolutionary game theory, we analyze the conditions under which guilt can provide individual fitness benefits to actors, and so evolve. In particular, we focus on the individual benefits of guilty apology. We find that guilty apology is more likely to evolve in cases where actors interact repeatedly over long periods of time, where the costs of apology are low or moderate, and where guilt is hard to fake.

1 Introduction
Evolutionary game theory is a branch of mathematics used to model the evolution of strategic behavior in humans and animals. This framework is not traditionally employed to understand the evolution of emotions because emotions, simpliciter, are not behaviors. O’Connor (2016) argues, however, that the enormous body of literature from evolutionary game theory on the evolution of cooperation, altruism, and apology can be used to inform the evolution of guilt by showing where and when guilt can provide individual fitness benefits to actors. She also presents novel modeling work clarifying how guilt can benefit individuals by prompting apology. In this paper, we extend O’Connor’s analysis to analytic, rather than computational models. We also give further insight into the conditions under which guilty apology can evolve.

In section 2 we describe the inferential strategy by which we use evolutionary game theoretic results to provide insight into the evolution of guilt. We also discuss O’Connor’s basic insights into the conditions under which guilt provides individual fitness benefits to actors. In section 3 we present our evolutionary model of guilty apology, and clarify conditions under which guilt is likely to evolve to play this strategic role.

2 Evolutionary Game Theory and Guilt
Evolutionary game theoretic models involve two basic elements—games and dynamics. Games, in the game theoretic sense, are simplified representations of strategic interactions. Dynamics, on the other hand, specify how a population of actors playing a game will change, or evolve. Which behaviors (or strategies) in the game will become more prevalent as evolution progresses? Which will disappear?
Games, in evolutionary models, explicitly represent three things—players, strategies, and payoffs. These correspond to the agents involved in an interaction, their possible behaviors, and what they get for their behaviors, respectively. Note that there is no resource, here, for representing the emotional state of an actor. Inasmuch as emotions in humans are causally connected to behaviors, however, we can use these models to gain insight into what functional role emotions might play. Guilt, our focus here, is associated with three types of behaviors in humans. First, the anticipation of guilt prevents social transgression (Tangney et al., 1996). It is correlated, for this reason, with altruistic and cooperative behavior in humans, as well as decreases in norm violation (Regan, 1971; Ketelaar and Tung Au, 2003; Malti and Krettenauer, 2013). Second, the actual experience of guilt leads to a suite of reparative behaviors including apology, gift giving, acceptance of punishment, and self punishment (Silfver, 2007; Ohtsubo and Watanabe, 2009; Nelissen and Zeelenberg, 2009). Lastly, expressions of guilt seem to lead to decreased punishing behaviors, and forgiveness, by group members (Gold and Weiner, 2000; Fischbacher and Utikal, 2013; Eisenberg et al., 1997). If we find, in evolutionary models, that these sorts of behaviors provide selective advantages to individuals, we identify a situation in which guilt can provide a selective advantage as well. By dint of leading to selected behavior, guilt is also selected.

O’Connor (2016) identifies three sets of evolutionary game theoretic results that can inform the evolution of guilt. The first employs the famous prisoner’s dilemma game to model the evolution of altruism, by which we mean any behavior in which an agent decreases their own payoff in order to increase another agent’s payoff. In this literature, the mechanisms which have been identified that can create individual level benefits for altruism are reciprocity and punishment.\(^1\) If one is in a group where actors can remember past actions and reciprocate—by behaving altruistically towards altruists and selfishly towards the selfish—altruism can be directly beneficial to the individual. Emotions that promote altruism, such as guilt, are likewise beneficial. When actors punish those who fail to behave altruistically, likewise altruism, and guilt, are directly beneficial to the individual. Notably, human groups engage both in reciprocity and in punishment suggesting that guilt will tend to provide a selective advantage by preventing failures of altruism in these groups (Boyd et al., 2003; Boyd and Richerson, 2009).

Secondly, in models that employ the stag hunt to represent mutually beneficial, but risky, cooperation, guilt can benefit actors by stabilizing such cooperative behavior. In these types of interactions, it always benefits actors to cooperate when their partners do as well, even in the face of transient temptation to do otherwise. An emotion, like guilt, that promotes cooperation will then provide individual benefits to any actor in a generally cooperative group.\(^2\)

\(^1\)See Nowak (2006) for an overview of the evolution of altruism in the prisoner’s dilemma. We do not address potential group-level benefits of guilt, here. For theoretical discussion of such possible benefits of guilt, see Deem and Ramsey (2015); Ramsey and Deem (2015). Nor do we address kin selection.

\(^2\)See Skyrms (2004) for more on the stag hunt and the evolution of cooperation.
Alexander (2007) shows that cooperation in the stag hunt is especially likely to evolve in groups where the same actors tend to keep interacting.

Lastly, it has been observed that apology can benefit individuals playing the iterated prisoner’s dilemma—a version of the game where the same actors repeatedly are engaged in an opportunity for altruism. In this game, strategies that reciprocate by refusing to behave altruistically towards selfish types can do well, but they suffer a problem when faced with accidental bad behavior by a partner. These strategies can become locked in a spiral of mutual negative reciprocation, which hurts all involved. Actors who apologize, and accept the apologies of group members, can gain an advantage in such conditions. These apologies can work if they are costly (Okamoto and Matsumura, 2000; Ohtsubo and Watanabe, 2009; Ho, 2012; Han et al., 2013), if they are unfakeable, or if they combine elements of costly and unfakeable apology (O’Connor, 2016). This indicates that the often costly apologies generated by the experience of guilt may, paradoxically, provide individual fitness benefits in the long run by convincing group members to accept guilty actors into the social fold after bad behavior. In the rest of the paper, we present modeling results expanding and supporting this claim.

3 Model and Results

A prisoner’s dilemma is a two-player game in which each player has two possible strategies: “cooperate” and “defect”. If both players cooperate, they both get a moderate payoff (2, in our model). If one cooperates and one defects, the cooperator gets nothing (0) and the defector gets a large payoff (3). If they both defect, they both get a small payoff (1). In other words, mutual cooperation is preferable to mutual defection, but each player does best to defect regardless of the other player’s choice. While we follow the literature in using the term “cooperation” for the prosocial strategy, it is in fact a case of altruism because players who choose it incur a cost and increase their partner’s payoff.

This game is a dilemma because the strategy pair where both players defect is the only Nash equilibrium of the game—the only set of strategies for which no player can benefit by switching. The expected outcome is then for both players to defect, despite the fact that mutual cooperation is preferred by everyone.

In an iterated prisoner’s dilemma, as mentioned, agents repeat the prisoner’s dilemma round after round with some probability, and as a result cooperation can get a foothold in the game via reciprocation. One such reciprocating strategy is called the “grim trigger”—players begin by cooperating, but if their partner defects they immediately switch to defection for the rest of the interaction.3 In this way, they benefit themselves by cooperating with cooperators and defecting with defectors.

---

3One can equivalently think of this probability, n, as encoding the average number of rounds each encounter is expected to last, as this is given by \( \frac{1}{1-n} \). For example, if \( n = 0.95 \), there will be an average of 20 rounds played.
As described in the last section, this strategy runs into problems when players have a chance of accidentally performing the wrong action—defecting instead of cooperating, or vice versa. Accidental defections cause grim triggers to permanently defect on good cooperative partners. Mutual negative reciprocation of this sort is mutually damaging. In such an environment, an apologetic strategy, which we call the “guilt-prone grim trigger”, or just guilt-prone, for short, can outperform a punitive one. Guilt-prone players act as grim triggers, but apologize after accidental defection. Upon receipt of such an apology, they forgive and forget, and so return to playing cooperate.

The problem with this strategy is that an apology will not effectively signal guilt if defectors can use it to convince their partners to cooperate, even though they intend to continue to defect. Another way of putting this is that guilty apology might not be evolutionarily viable if “fakers” can take advantage of forgiveness among guilt-prone players.

There are two lines of defense against such fakers. One is for guilty apologies to be unfakeable. This relates to arguments by Frank (1988) that moral emotions, such as guilt, evolve as honest signals of cooperative intent in humans. Empirical evidence suggests that humans do trust signals of guilt from group members to some degree when deciding whether to forgive and forget, but that guilt, unlike some emotions, is not associated with stereotyped facial and body postures (Deem and Ramsey, 2015). In other words, it is not entirely unfakeable. For this reason, in our models we assume that guilt-prone types always manage to successfully apologize and fakers are still successful with some probability.

Another way to discourage fakers is to impose a cost for apologizing. When guilt-prone types apologize to each other, they are able to re-enter a potentially long cooperative engagement where they both reap the benefits of mutual aid. This means that the expected benefit to apologizing is high. When fakers apologize, they defect the next round, necessitating another costly apology if they wish to re-enter the social fold. This means that the benefit to fakers of apologizing is a short period of defection, which yields a relatively small payoff. These differential benefits means that paying an identical cost will be less worthwhile for fakers than guilt-prone types under many conditions.

Our models work as follows. We assume that a population of actors plays the iterated prisoner’s dilemma where every round the game continues with probability, $n$. Each round, there is a probability, $a$, that actors accidentally perform the wrong action. The strategies in the population are cooperate unconditionally, defect unconditionally, grim trigger, guilt-prone, and faker. There is some probability $p \leq 1$ that fakers manage to signal their guilt. And in order to successfully apologize, actors pay a cost. To allow for the possibility that actually guilty types pay a lower cost than fakers to convince others of their

---

4O’Connor (2016) also looks at “tit-for-tat”, another reciprocating strategy. She finds similar results in models that include it, instead of the grim trigger, although the analysis is more complicated for various reasons.

5In some of the models we consider, stochastic effects mean that this benefit is slightly higher or lower.
guilt we define $c \geq d$ where $c$ is the cost of apology for guilt-prone types and $d$ for fakers.\footnote{After choosing values for our parameters, we can generate a payoff table for each strategy based on the expected outcome for playing an iterated prisoners dilemma under these conditions. For the details of these calculations see Rosenstock (2016).}

### 3.1 When Can Guilt Evolve?

In this section we will address the conditions under which guilt-prone is an evolutionarily stable strategy (ESS). ESSes are strategies where populations playing them cannot be invaded by a small number of actors using a different strategy. It is also the case that for the replicator dynamics, the most commonly used model of evolutionary change in evolution game theory, ESSes are stable. If populations evolve to them, they stay there, absent other forces.\footnote{ESSes are strategies that do better against themselves than other strategies do against them. Or, if another strategy does equally well against an ESS, the ESS does better, when they meet, than the strategy does against itself.} For this reason, ESSes are often likely to evolve.\footnote{They are not the only stable states under the replicator dynamics, and sometimes ESSes are quite unlikely to evolve. Huttegger and Zollman (2013) discuss problems with ESS methodology as opposed to dynamical analyses. O’Connor (2015) discusses a particular example, the evolution of learning, where ESS analysis is misleading. In the next section we will move to a dynamical analysis for these reasons.}

In the models we consider, unconditional defection is always an ESS. Guilt-proneness is sometimes an ESS, though it can be destabilized by successful fakers (who themselves are eventually replaced by defectors). As it turns out, the guilt-prone strategy is evolutionarily stable against fakers in a sizable portion of the parameter space. As indicated above, both higher cost, $c$, and lower probability of fake apologies working, $p$, helps protect guilt-prone players against fakers. For now, we will assume that $c = d$, or that both types pay the same cost for apology.

In order for the guilt-prone strategy to be an ESS given a fixed error rate, $a = 0.01$, and chance of repeat encounter, $n$, figure 1 shows that the harder an apology is to fake, the cheaper the cost of apology needs to be. Alternatively, the higher the cost, the less fakeable an apology needs to be. This figure also shows that for some $p$, the guilt-prone strategy can be an ESS against fakers even when the cost of apology is negative. Moreover, $p$ doesn’t need to be small at all, nor $c$ large, in order for the guilt-prone strategy to be an ESS against fakers if $n$ is sufficiently large. Note that these figures only show conditions under which guilt is stable against fakers—more on other strategies in a minute.

This graph also indicates that for larger $n$, guilt is stable under wider conditions, i.e., with lower cost, $c$, and higher fakeability, $p$. This makes sense, since the more rounds that are played on average, the more likely a guilt-prone player is to reap benefits of long interactions with other guilty types, and the more likely they catch on and disbelieve a fake apology, thus depriving the faker of the benefits of defecting against a cooperator for the rest of the encounter.

An increase in the error rate, $a$, makes it marginally more difficult for the
Figure 1: Minimum cost, \( c \), for guilt-prone to be an ESS vs. faker, for each fakeability value \( p \), error rate \( a = 0.01 \).

guilt-prone strategy to be an ESS against faking. This is because as \( a \) increases, guilt-prone players accidentally defect and have to pay the apology cost more frequently, and fakers are more likely to accidentally cooperate, thus delaying their chances of getting caught. However, the effect of \( a \) is relatively minor compared to the effects of \( p \), \( c \) and \( n \), especially where \( a \leq 0.1 \), which we consider a reasonable assumption.

One might also be interested in determining under which conditions the guilt-prone strategy is an ESS versus other strategies. When not playing against fakers, the parameter \( p \) no longer matters. Figure 2 shows the minimum \( n \) for which the guilt-prone strategy is an ESS versus grim trigger, unconditional cooperation, and unconditional defection when the error rate is \( a = 0.01 \). The guilt-prone strategy is an ESS for most of the parameter space we’ve been looking at, i.e., low costs and high chance of repetition. As \( a \) increases, \( n \) needs to be a bit larger for the guilt-prone strategy to be an ESS, but little changes in the ranges of \( n \) and \( c \) that we focus on. Note that in early human groups, we expect \( n \) to be very high, meaning that guilt-prone should be an ESS under these conditions.

3.2 The Robustness of Guilty Apology

We have now seen that guilt-proneness can evolve in order to promote costly and/or honest apology. In this section, we will describe, in some greater detail, the conditions under which guilt is likely to evolve for this function, or likely to be stable if it does. ESS analyses are useful because they tell us something about which strategies have the potential to evolve. Some ESSs, however, have very small basins of attraction under the replicator dynamics. A basin of attraction, for an equilibrium, is the set of population states that evolve to that equilibrium.
The size of a basin of attraction can tell us something about the evolvability of a strategy. Equilibria with large basins are more likely to evolve from a random starting place. Also, mutations, or noise, in evolutionary processes can move populations from one equilibrium to another. Equilibria with large basins of attraction tend to be harder to disrupt, while those with small ones are easy to move away from. In models explicitly representing this sort of noise, populations tend to spend most of their time at equilibria with large basins of attraction.

We thus want to ask: under what conditions does guilt-proneness, as represented in our models, have a large basin of attraction? What are the factors that make it likely to evolve and be stable for the purpose of promoting apology? There are a few parameters to consider in answering this question. We can ask what happens to the basin of attraction for guilt-proneness when we vary $c$, the cost of apology for guilt-prone players, $p$, the probability that fakers successfully trick others into trusting their apologies, and $d$, the cost of apology for fakers.

Let us start with $p$. In models without fakers, guilt-prone types do very well, because their apologies are trustworthy. Fakers can be thought of as siphoning away the benefits of guilty apology. For this reason, holding other conditions fixed, guilt-proneness has a larger basin of attraction whenever $p$ is smaller. If guilt is hard to fake, it is more likely to evolve.

The role of $c$ and $d$, the costs for apology, are a little more subtle. First, consider the case where $c = d$, or fakers and guilty types pay the same cost. When $p$ is low, guilt-prone types do well against fakers. For this reason, increasing $c$ actually makes guilt-proneness less likely to evolve. It simply decreases the payoffs to guilty types, while failing to significantly help them differentiate themselves from fakers. When $p$ is higher, cost can help guilt-prone types. It
allows them to prove their cooperative intent, compared to faker types. Figure 3 shows the sizes of the basin of attraction for guilt-proneness, as opposed to defection, in games where $p$ and $c$ vary, $a = .01$ and $n = 0.95$. The x-axis tracks cost, $c = d$, which ranges from .005 to 1. The y-axis shows the likelihood that guilt evolves. For $p = 0.95$, the optimal cost for the evolution of guilty apology, of those explored, is 0.4. For $p = 0.9$, the optimal cost is 0.2. For the smaller values of $p$, costs make guilt less likely to evolve.

![Basins of attraction for guilt-prone](image)

Figure 3: Sizes of basin of attraction for guilt-prone strategy as $c$ and $p$ vary.

The other situation worth considering here is the one where $d > c$, or where fakers must pay some greater cost to apologize. The idea is that their apologies are less convincing and so social partners exact an extra cost before trusting their apologies. Figure 4 shows basins of attraction for guilt in these models with $p$ held fixed at 0.95 and $a = 0.01$. Two data sets are pictured here. For the first, $c = 0$—guilt-prone types pay no cost to apologize—and $d$ ranges from 0.01 to 0.9. For the second, $c = 0.2$—a small cost for guilty apology—and $d$ ranges from 0.21 to 0.9. In both cases, increasing $d$, the cost to fakers, while holding $c$ fixed, increases the likelihood that guilt evolves. When there is a cost for guilt, this generally decreases the likelihood it will evolve. Both these results should be unsurprising, costs for fakers make faking a less successful strategy, and stabilize guilt. Costs for guilty apology make guilty types less successful and allow defection to evolve more often.

In all the results just shown, we hold $a$ and $n$ fixed. As in the ESS analysis, variations in $a$, within a sensible parameter range, have little effect on outcomes.

---

9The basins of attraction were measured using the discrete time replicator dynamics. Results are from 10000 simulations run until the population was clearly converging to one of the two rest points—all play guilt-prone, or all play defect. The strategies included were unconditional cooperation, unconditional defection, guilt-prone grim trigger, and faker.
Variations in $n$ have much more significant effects. For larger $n$, generally, guilt proneness will be more evolvable. Also, adding other strategies can shift evolutionary outcomes of these models significantly. If the grim trigger is included, for example, it is also an ESS under many parameter values. For some parameter values, the presence of this strategy increases the basin of attraction for guilt proneness because it disproportionately hurts defectors and fakers. For other parameter values, especially when costs are higher, grim trigger is so successful itself that it decreases the basin of attraction for guilty apology. We do not include a full analysis of models with grim trigger strategies for space reasons.

To sum up, the more difficult it is for faker types to convince social partners that they truly feel guilt, the more stable and evolvable guilt (for the function of apology) is. Costs to apology can help promote the evolution of guilt by differentially harming guilty versus faker types, but these costs also come at, well, a cost in terms of stability in the face of defection.

4 Conclusion

Results from evolutionary models indicate that there are many conditions that can make guilt-proneness individually beneficial for actors. When it comes to benefits to guilt before bad behavior, these include the presence of reciprocating, or punishing group members, and the presence of established, mutually beneficial patterns of cooperation. When it comes to benefits after bad behavior, guilt can help actors if it allows for unfakeable apology, costly apology, or some combination of the two of these. Guilt is particularly likely to evolve and be stable for this function if it is harder to fake, either in the sense that group members do not believe fake apologizers, or in the sense that they levy higher costs to ensure the apologies of faker types.
One might object that the models presented here do not explicitly represent the role of culture in guilt. Culture seems likely to have played a role in the evolution of guilt, and clearly plays a role in the production of guilt in modern societies. We do not mean to downplay the importance of cultural elements in the evolution of guilt. Nonetheless, the models here still provide insight in that they clarify the conditions, whether cultural or environmental, under which guilt, whether culturally produced or not, can provide a selective advantage and evolve.

References


Michael Deem and Grant Ramsey. Guilt by association. 2015.


