Guilt, Games, and Evolution

Cailin O'Connor, Department of Logic and Philosophy of Science, UC Irvine

May 2016 – Evolutionary game theory is a branch of mathematics used to model the evolution of strategic behavior in humans and animals. This framework is not typically used to shed light on the evolution of emotions because emotions are not themselves behaviors.[1] I have previously argued, however, that the huge amount of literature from evolutionary game theory on the evolution of cooperation, altruism, and apology can be used to study the evolution of guilt by showing where and when guilt can provide individual fitness benefits to actors by dint of causing adaptive behaviors (O'Connor, 2016). In that paper, and in subsequent evolutionary game theoretic work, Sarita Rosenstock and I have focused in particular on potential benefits accruing to guilt-prone individuals as the result of costly apology (Rosenstock and O'Connor, ms).

In this article, I will summarize findings from these papers. As I will point out, there are three main sets of results in evolutionary game theory that shed light on the evolution of guilt. First, work on altruism in the prisoner's dilemma game indicates that in environments where actors engage in reciprocation and punishment, guilt can provide individual benefits by promoting altruism. Second, work on the stag hunt shows that when actors are in groups of relatively cooperative partners, and especially when they are engaged in repeated interactions with neighbors, guilt can promote fitness by leading to cooperation. And lastly, results on costly apology show that, perhaps unintuitively, paying costs can allow actors to successfully apologize and to reap the cooperative benefits of doing so. These findings do not fully explain the evolution of guilt, but they do clarify the conditions under which it provides evolutionary benefits. For this reason, they have the potential to help emotions researchers in building more detailed pictures of the evolution of moral emotions.

In the next section, I'll describe the modeling strategy used throughout the paper, and the behaviors typically associated with guilt. Following that, I will describe the models and results from each of the three branches of research mentioned above. In the conclusion, I will argue that these models can help determine conditionals of the form 'if *X* obtains, guilt can provide individual fitness benefits' that emotion researchers can employ in forming and assessing more detailed accounts of the evolution of guilt.

Evolutionary Game Theory and Guilt

Evolutionary game theoretic models represent populations of actors whose behaviors evolve over time. In particular, there are two elements of these models. Games represent the sorts of strategic interactions that individuals, including humans, encounter in social contexts. (There are games, for example, to represent bargaining, coordination, communication, public goods dilemmas, etc.) Dynamics are rules for how actors who play these games will change their behaviors over time. For example, if actors learn to repeat behaviors that benefit them, dynamics can be used to model this sort of change. In a biological evolutionary scenario, if actors who make successful choices have more offspring, and these offspring have similar behavioral genes, dynamics can be used to represent the spread of these beneficial genes. Throughout this paper, results described will typically be from models using the replicator dynamics – a model of change assuming, simply, that strategies that do better on average tend to spread while those that do worse on average tend to die off.

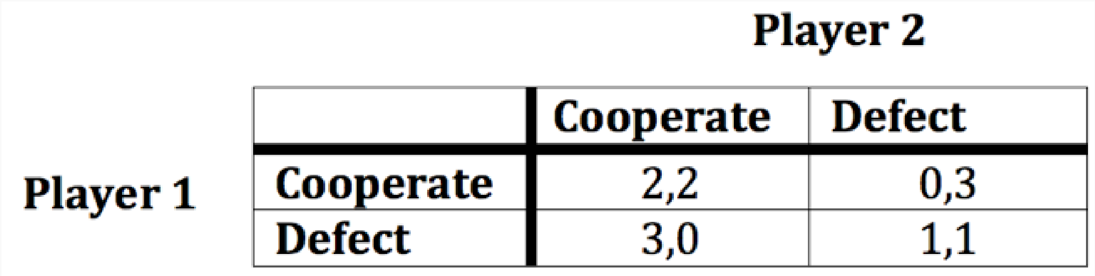
Games, the first element of evolutionary game theoretic models, have explicit features intended to represent three things: players, strategies, and payoffs. Players correspond to the actors in an evolving population. For our purposes, these will be humans in populations where guilt can potentially evolve. Strategies correspond to the behavioral choices these actors can make. In the models discussed below, strategies will include choices like ‘behave altruistically’ or ‘apologize’. Payoffs correspond to what the players get for combinations of strategies. Two players who coordinate their actions, for example, may receive good payoffs compared to actors who mis-coordinate. Notice that games do not include explicit representations of emotions. In O’Connor (2016), however, I describe a strategy for modeling the evolution of guilt using evolutionary game theoretic models. Rather than focusing on the emotion itself, I focus on the behaviors associated with guilt. When these behaviors provide selective advantages, we can assume that guilt also comes under some level of positive selection pressure by dint of causing them.

To employ this strategy to model the evolution of guilt, then, it will first be necessary to say something about the behaviors associated with guilt. There are three general classes of behavior that are of interest. First, the anticipation of guilt prevents transgressive behavior in humans (Tangney et al., 1996). In particular, guilt-proneness is associated with higher levels of altruism and cooperation, and with decreased norm violation (Regan, 1971; Ketelaar and Tung Au, 2003; Malti and Krettenauer, 2013). Second, the experience of guilt leads humans to engage in reparative behaviors including apology, acceptance of punishment, gift giving, and self punishment (Silfver, 2007; Ohtsubo and Watanabe, 2009; Nelissen and Zeelenberg, 2009). And lastly, humans are impacted by signs of guilt in others. Apology and expressions of guilt and remorse are often met with a decrease in punishing behavior (Gold and Weiner, 2000; Fischbacher and Utikal, 2013; Eisenberg et al., 1997). In the next sections, I will describe evolutionary models of these sorts of behaviors. The idea is that in cases where these behaviors provide evolutionary benefits, guilt has the potential to do so as well.

Before continuing, I should note that this article will not discuss potential group level benefits from guilt. Much work in evolutionary game theory has focused on group level benefits of altruism, and, as noted, altruism is associated with guilt. As such, this may be a promising avenue for using evolutionary modeling to inform the evolution of guilt. For more on group selection, see Deem and Ramsey (2016a) and Deem and Ramsey (2016b) (this issue of *Emotion Researcher*).

The Prisoner’s Dilemma and Altruism

The first relevant set of results employs the famous prisoner’s dilemma game to model the evolution of altruism. By altruism I mean a behavior in which an actor decreases their own payoff to increases another’s. The game works as follows – two players each have two strategies, to ‘cooperate’ (or behave altruistically) or to ‘defect’. Players who jointly cooperate outperform actors who jointly defect. The game is a dilemma, however, because regardless of what the other player does, each player is incentivized by her payoffs to defect. Figure 1 shows a payoff table representing this game. Rows represent the strategies of player one and columns the strategies of player two. Entries to the table show payoffs for any combination of strategies, with player one listed first. Joint cooperation yields a payoff of 2, joint defection 1, defecting against a cooperator yields 3, and cooperating against a defector 0.



		Player 2	
		Cooperate	Defect
Player 1	Cooperate	2,2	0,3
	Defect	3,0	1,1

Figure 1: A payoff table for the prisoner’s dilemma. Rows represent player one’s strategies and columns represent player two’s strategies. Entries represent payoffs with player one listed first.

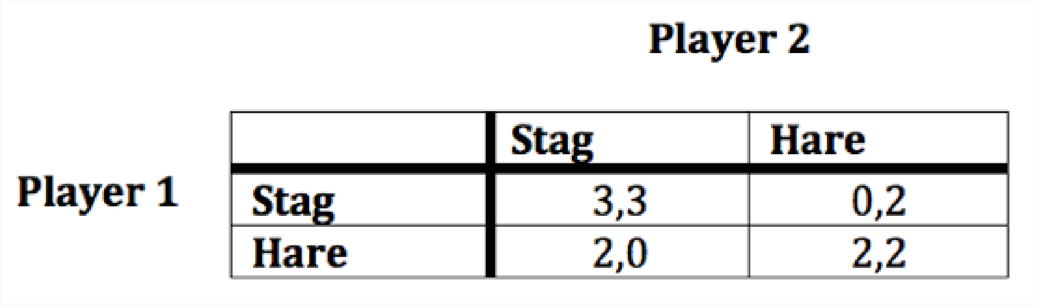
A Nash equilibrium is a set of strategies in a game where neither player can switch and improve their payoff. For this reason, Nash equilibria are important as predictions of actors’ behavior. Once actors play them, they have no incentive to change strategies. The single Nash equilibrium of this game is for both players to defect. Of course, in the real world, humans do exhibit altruistic behavior under many circumstances. This observation has turned the evolution of altruism into a puzzle – although altruistic behavior is *prima facie* not rational from an individual choice perspective, the evolutionary game theoretic literature seeks to explore the conditions under which it can evolve nevertheless.

In this literature, the mechanisms which have been identified that can create individual level benefits for altruism are *reciprocity* and *punishment*. (See Nowak (2006) for an overview of the evolution of altruism in the prisoner's dilemma. Other mechanisms – like group selection, kin selection, and network structure – can also lead to the evolution of altruism, but not via individual benefits, so I do not discuss them here.) In environments where actors reciprocate, when an actor defects group members will respond by defecting right back. In contrast, a cooperator will be met with cooperation. This gives cooperators an evolutionary edge. As for punishment, in groups where actors punish those who defect, the payoff for defection is decreased. This means that, for the right level of punishment, it will no longer be a Nash equilibrium to defect, and so defection is not expected to evolve.

When it comes to guilt, these observations mean that in groups where actors reciprocate and punish, guilt can provide individual fitness benefits by promoting altruistic behaviors. Guilt-prone actors are less likely to behave selfishly, and as a result they reap the benefits of reciprocal altruism. For example, suppose a hunter feels guilty if they keep all the meat they catch, and so they share it with the group. In a reciprocating group, this hunter will receive enough meat in the future to more than make up for the loss, and so ultimately benefits as a result of her prosocial emotion. Likewise, guilt-prone actors can avoid punishment by group members and benefit in that way. If a hunter is expected to share meat, and guilt leads her to do so dependably, she will avoid the negative social consequences of failing to share and at least occasionally getting caught and punished. It is noteworthy that both reciprocity and punishment are observable in human groups and were probably part of the social evolutionary environment for modern humans (Boyd, 2009; Boyd et al. 2003). This means that in this social environment guilt should be expected to come under positive selection pressure by dint of promoting altruistic behavior.

The Stag Hunt and Cooperation

The second set of relevant results employs the stag hunt to model mutually beneficial cooperation. (To be perfectly clear, although 'cooperate' is the conventional name of the prosocial strategy in the prisoner's dilemma, it is more properly thought of as altruistic, while 'stag' in the stag hunt is a truly cooperative strategy.) Imagine two hunters who each have the choice to either hunt for stag or hare. If they hunt stag, they both capture a great deal of meat. Hunting hare generates less bounty. But stag hunting is risky in that it requires both actors to attend to the task for success. As the payoff table in figure 2 shows, mutual stag hunting yields a payoff of 3, hare hunting always yields a payoff of 2, and hunting stag while an opponent hunts hare yields a payoff of 0.



		Player 2	
		Stag	Hare
Player 1	Stag	3,3	0,2
	Hare	2,0	2,2

Figure 2: A payoff table for the stag hunt.

Unlike the case of the prisoner's dilemma, mutual cooperation is a Nash equilibrium of the stag hunt. When two players both hunt stag, they have no incentive to switch strategies. But stag hunters do not do particularly well in populations composed of hare hunters. In other words, for cooperation, and emotions leading to cooperation, to be beneficial, actors must be in a group with other cooperators. To give an example, imagine an actor who has agreed to work on a joint project – say building a dike – but feels tempted to simply laze around instead. Suppose further that this actor is guilt-prone, and decides to engage in the joint work in spite of the temptation to shirk it. If her partner likewise works on the dike, the guilt-prone actor benefits as a result of her emotion. If her partner does not work on the dike, she suffers. This means that once some level of cooperation is off the ground in human groups, an emotion like guilt can come under positive selection pressure to stabilize and improve cooperation.

Alexander (2007) shows that cooperation in the stag hunt is especially likely to evolve in groups where the same actors tend to interact repeatedly, as in early human groups. See Skyrms (2004) for more on the stag hunt and the evolution of cooperation.

The Iterated Prisoner’s Dilemma and Costly Apology

There is one last body of work to discuss, and this is work on the evolution of costly apology. The iterated prisoner’s dilemma is a game where actors play the prisoner’s dilemma again and again over the course of some number of rounds. In this game, reciprocating strategies, briefly mentioned above, can help actors do well. Reciprocators tend to cooperate with cooperators and defect with defectors. In this way, they gain the benefits of mutual altruism, and avoid being taken advantage of by selfish partners.

What happens, though, when a normally altruistic partner accidentally behaves selfishly? This, in fact, happens all the time in human groups. A normally well behaved child sometimes steals a cookie, a normally fair-dealing work partner cheats because she is under great financial pressure, a normally dependable co-author fabricates data in desperation to publish.

In such cases, a reciprocator will respond by defecting as well, which may lead the original defector to reciprocate against this new defection, etc. In other words, actors who reciprocate can sometimes get stuck in spirals of mutual revenge, leading them both to poor outcomes. For example, suppose a hunter fails to share meat because they are especially hungry, or in a bad mood, or distracted. If a partner then refuses to share meat on the next hunt as a result, the mutually altruistic relationship they have may unravel, hurting them both. One solution to this problem is apology. In models of the iterated prisoner’s dilemma, apologies can help solve the retribution problem just described. But, in order to be successful, these apologies must be costly (Okamoto and Matsumura, 2000; Ohtsubo and Watanabe, 2009; Ho, 2012; Han et al., 2013), hard to fake, or some combination of costly and hard to fake (O’Connor, 2016).

Why is this the case? When actors can apologize cost-free, selfish types can defect against a partner, apologize, and defect again the next day. I will call these actors fake apologizers, or ‘fakers’. When enough fakers are in a population, there is no reason to trust an apology one receives, because such apologies do not necessarily mean that your partner will behave altruistically next time. If actors pay some cost to apologize things are slightly different, though. For an altruistic type engaged in mutual altruism, the benefits of apologizing are enormous. Upon apologizing they, and their partner, will continue to reap the benefits of mutual altruism over the course of a long, fruitful engagement. For a selfish type, the benefits of issuing an apology are relatively small. They will manage to take advantage of their partner again, but after doing so will have to apologize once more to avoid negative reciprocation. This means that altruists gain a disproportionate benefit for apologizing. When the right costs are introduced, only altruists will still be willing to apologize, because it will no longer be worthwhile for fakers to bother paying the costs to regain the trust of their partner.

Consider the following (silly) example to drive this home. Bob works for Allison making cupcakes, and, as a payment, receives one cupcake every day. On Friday he steals an extra cupcake. Allison is tempted to fire him, but Bob begs forgiveness. He insists that he will clean the entire cupcake factory to prove his remorse is real. If he were planning to steal again the following day, it would not be worthwhile for Bob to pay this cost. (One extra cupcake is not worth the benefit of cleaning an entire factory.) If, however, Bob plans to keep to the agreement and receive a cupcake every day for the rest of his working life, cleaning the factory is a small price to pay. Under these conditions, strategies for costly apology can benefit individuals, and so can evolve.

How does guilt factor in? Remember that empirical work on guilt shows that it causes a suite of costly behaviors such as self-punishment, gift giving, and acceptance of punishment. All of these can count as costs that ensure genuine apology. If a group member observes such costly behavior, she has reason to trust the apologizer. So, unintuitively, paying these costs may actually benefit guilty individuals. Under these conditions, costly apology, and guilt, can evolve.

Note, though, that these costs are still detrimental to those who pay them. What if guilt-prone actors, who intend to behave altruistically in the future as a result of their moral emotional tendencies, could somehow convince others that their apologies are real without paying costs to do so? This possibility is related to work by economist Robert Frank. He argued that moral emotions evolved as honest signals of cooperative intent in humans (Frank, 1988). The idea is that moral emotions are hard to fake, and so can be used to pick good altruistic partners. In O'Connor (2016), I point out that if guilt allows actors to honestly signal their remorse upon defection, it can evolve to promote cost free apology as well. The idea here is that after Bob steals the extra cupcake, he apologizes and professes his guilt. If Allison can actually tell that he really feels guilty, i.e., if his emotional signal is unfakeable, she can trust him because guilt-prone people do, in fact, tend to behave altruistically.

The problem with this account is that guilt, unlike many other emotions (joy, fear, anger), is not associated with stereotypical facial and body postures. In other words, it is at least somewhat fakeable. For this reason in O'Connor (2016) and in Rosenstock and O'Connor (ms), we explore models where apologies may be both costly *and* somewhat hard to fake. The idea is that if people are even a little bit able to read each others' guilt, this may allow for the evolution of guilty apology that is only a little bit costly. (This builds off of work by Huttegger et al. (2015) arguing that costly signalers may employ lower costs if their signals are also somewhat hard to fake.)

To make this third possibility for guilty apology clear, imagine that Bob again steals the cupcake and apologizes profusely, saying that he feels guilty. And suppose that if Bob really does feel guilty, he will be somewhat better at convincing Allison that he feels guilty. If so, a strategy where he pays a smaller cost – cleaning just the floor in the cupcake room instead of the entire factory – and she accepts his apology, will be able to evolve.

In Rosenstock and O'Connor (ms), we look in detail at the conditions under which such partially costly and partially honest apologies are likely to evolve. In doing so, we clarify the conditions under which guilt will come under positive selection pressure by dint of leading to apology. When guilt is easy to fake, costs can, indeed, allow guilty apology to evolve by stopping fakers. But these costs always inhibit the evolution of guilt, at least to some degree. When guilt is harder to fake, costs are unnecessary for the evolution of guilty apology. The better humans are at reading each other, the better the evolvability of guilty apology. In all cases, guilty apology provides bigger benefits when actors are involved in long, repeated interactions. This is because those that apologize are situated to reap larger benefits from doing so.

Conclusion

The body of work just described can be employed by emotions researchers to clarify, improve, and assess their detailed and more realistic accounts of the evolution of guilt in humans. In particular, game-theoretic research on guilt has individuated circumstances in which guilt can be individually advantageous. These circumstances can be summarized through a set of conditionals of this form: "Ceteris paribus, if circumstances X apply, guilt produces individual fitness benefits and is more likely to evolve'. To summarize once more, these circumstances are that actors reciprocate, punish, are generally cooperative, repeat interactions with the same neighbors, and, for the purposes of apology, engage in long repeated interactions and are somewhat able to read each others emotions. It is suggestive that many of these conditions held in early hominid groups.

There is one last thing that must be mentioned. Cultural evolution was almost certainly a factor shaping the biological evolution of guilt. Furthermore, many researchers have pointed out that the production of guilt is culturally mediated in that it is sensitive to cultural norms. The models described here do not include any sort of gene-culture co-evolution, nor do they explicitly include cultural features. This, however, does not really pose a problem given the proposed purpose of these models. This is because the conditions under which guilt provides individual fitness benefits can be a result of either the natural or the social environment. Furthermore, even if guilt is culturally produced, and spreads via cultural evolution, the same models provide insight into where and whether this is expected to occur.

References

J McKenzie A. (2007). *The structural evolution of morality*. Cambridge University Press.

Boyd, R. and Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3281–3288.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535.

Deem, M. J., & Ramsey, G. (2016a). Guilt by association? *Philosophical Psychology*, 29(4), 570-585

Deem, M. J., & Ramsey, G. (2016b). The Evolutionary Puzzle of Guilt: Individual or Group Selection?, *Emotion Researcher*, ISRE's Sourcebook for Research on Emotion and Affect, Andrea Scarantino (ed.), http://emotionresearcher.com/the-evolutionary-puzzle-of-guilt-individual-or-group-selection/, accessed May 27, 2016.

Eisenberg, T., Garvey, S. P., and Wells. M. T. (1997). But was he sorry? The role of remorse in capital sentencing. *Cornell Law Review*, 83:1599-1637.

Fischbacher, U. and Utikal, V.. On the acceptance of apologies. *Games and Economic Behavior*, 82:592–608, 2013.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. WW Norton & Co, 1988.

Gold, G. J. and Weiner, B. (2000). Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology*, 22(4):291–300.

Han, T. A., Pereira, L. M., Santos, F. C. and Lenaerts, T. (2013). Why is it so hard to say sorry? Evolution of apology with commitments in the iterated prisoner's dilemma. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 177–183. AAAI Press.

Ho, B. (2012). Apologies as signals: with evidence from a trust game. *Management Science*, 58 (1):141–158.

Huttegger, S. Bruner, J. P. and Zollman, K. (2015). The Handicap Principle is an Artifact. *Philosophy of Science*. 82(5): 997-1009.

Ketelaar, T. and Au, W. T. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition & Emotion*, 17(3):429–453.

Malti, T. and Krettenauer, T. (2013). The relation of moral emotion attributions to prosocial and antisocial behavior: A meta-analysis. *Child development*, 84(2):397–412.

Nelissen, R. and Zeelenberg, M. (2009). When guilt evokes self-punishment: evidence for the existence of a dobby effect. *Emotion*, 9(1):118-122.

Nowak, M. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563.

O'Connor, C. (2016). The evolution of guilt: a model-based approach. *Philosophy of Science*, 83(5).

Ohtsubo, Y. and Watanabe, E. (2009). Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evolution and Human Behavior*, 30(2):114–123.

Okamoto, K. and Matsumura, S. (2000). The evolution of punishment and apology: an iterated prisoner's dilemma model. *Evolutionary Ecology*, 14(8):703–720.

Regan, J. W. (1971). Guilt, perceived injustice, and altruistic behavior. *Journal of Personality and Social Psychology*, 18(1):124-132.

Rosenstock, S. and O'Connor, C. (ms). Selective Advantages of Guilt . Philosophy of Science Archive.

Silfver, M. (2007). Coping with guilt and shame: A narrative approach . *Journal of Moral Education* , 36(2):169–183.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Tangney, J. P., Miller, R. S., Flicker, L. and Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, 70(6): 1256-1269.

[1] The indirect evolutionary approach in economics focuses on the evolution of preferences, but these are not exactly emotions.