Title : The Evolution of Guilt: A Model-Based Approach

Author : Cailin O'Connor
Address : Department of Logic and Philosophy of Science, University of California, Irvine, 3151 Social Science Plaza A, Irvine, California 92697, USA
Electronic address : cailino@uci.edu

Acknowledgements : Many thanks to Justin Bruner, Edouard Machery, Jay Odenbaugh, Brian Skyrms, James Weatherall, and various conference and colloquium audiences for comments on this work. Special thanks to Michael Deem and Grant Ramsey for their continued feedback and the original inspiration.

Abstract

Using evolutionary game theory, I consider how guilt can provide individual fitness benefits to actors both before and after bad behavior. This supplements recent work by philosophers on the evolution of guilt with a more complete picture of the relevant selection pressures.

The Evolution of Guilt: A Model-Based Approach

Cailin O'Connor
Department of Logic and Philosophy of Science
University of California, Irvine, CA 92697
cailino@uci.edu

Abstract

Using evolutionary game theory, I consider how guilt can provide individual fitness benefits to actors both before and after bad behavior. This supplements recent work by philosophers on the evolution of guilt with a more complete picture of the relevant selection pressures.

1 Introduction

Moral emotions, such as shame and guilt, are deeply important to human moral behavior. Although few ethicists think the 'is' of evolved moral emotions should be directly translated to an 'ought' of ethical imperative, evidence from psychology and biology has increasingly made clear that at very least a full picture of human ethics must take these emotions into account.

This paper will focus on the evolution of guilt specifically. The goal is to provide an analysis of how guilt can be individually beneficial to actors, drawing on extensive literature from evolutionary game theory regarding the evolution of prosocial behavior. In this way, work by philosophers on the evolution of guilt (like that of Joyce (2007), Deem and Ramsey (2015), and Ramsey and Deem (2015)) can be supplemented by a more detailed picture of the relevant evolutionary pressures. As I will show, this literature suggests a number of ways that guilt can provide individual fitness benefits, both by preventing transgression in the first place, and by leading to reparative behaviors after transgression. In an attempt to better understand this latter role of guilt, I present novel modeling work on the evolution of apology.

In section 2, I discuss guilt in humans focusing on how it influences behavior. In section 3, I describe how evolutionary game theory can be used to inform the evolution of emotion. In section 4, I use evolutionary game theoretic models to shed light on the evolution of guilt.

2 The Role of Guilt in Behavior

Since this paper aims to connect insights from evolutionary game theory to the evolution of guilt, I will focus here on the behaviors guilt induces in humans, rather than the psychology behind it. This is because, as will be elaborated in section 3, game theoretic models represent agents through behavior.

Guilt is a negative emotion focused on one's past behavior, and, in particular, on social transgression (Tangney et al. 1996). Guilt seems to shape human behavior in two ways. First, the anticipation of experiencing guilt can influence actors' choices as to whether to commit a transgression. Empirical work demonstrates that guilt proneness in humans decreases the likelihood of social transgression (Svensson et al. 2013), and increases prosocial behavior, including altruism and cooperation (Malti and Krettenauer 2013; Regan 1971; Ketelaar and Tung Au 2003). Secondly, the actual experience of guilt after committing a transgression can lead to confession and to reparative behaviors like apology, gift giving, acceptance of punishment, and self punishment (Silfver 2007; Ohtsubo and Watanabe 2009; Nelissen and Zeelenberg 2009). Expressions of guilt also influence the behavior of interactive partners. Actors who express remorse are more likely to be judged guilty of committing a crime (Bornstein et al. 2002; Jehle et al. 2009), but their punishments tend to be reduced (Gold and Weiner 2000; Fischbacher and Utikal 2013; Eisenberg et al. 1997).

3 Modeling the Evolution of Guilt

Evolutionary game theory considers the evolution of strategic behavior in populations. *Games*—mathematical models of strategic interactions—are usually defined by three things: *players*, or actors in the game, *strategies*, the things actors can do, and *payoffs*, outcomes for the actors. One further game theoretic concept that must be introduced is that of *Nash equilibrium*—a set of strategies where no player can deviate and improve her payoff. Because no one wants to switch from a Nash equilibrium, these strategies are thought of as stable and are often evolutionarily significant.

Evolutionary game theorists employ what are called *dynamics* to games—rules that determine population change as a function of the payoffs actors receive. The *replicator dynamics* are the most commonly used model of evolutionary change in evolutionary game theory, and they will be the primary dynamics employed here. They assume that strategies that garner good payoffs will proliferate in a population while those that do not will tend to die out.

Evolutionary game theory deals with the evolution of behavioral traits in a social context, and has previously focused on prosociality, making this methodology an appropriate one to study the evolution of guilt (which, as mentioned, is often associated with prosocial behavior). This said, emotions simpliciter are not behaviors, and evolutionary game theoretic models represent actors through behaviors. What one can do is to model the evolution of a behavior associated with a particular emotion, show that this behavior is a successful one, and then argue that this may explain the evolution of said emotion. A tendency towards certain emotional states, then, is selected for by dint of causing certain types of behaviors.[1]

The prisoner's dilemma models two agents who may choose either to cooperate with each other or to 'defect'. While defection is better for the individual, two defectors do poorly in comparison to two cooperators. This seems to capture the strategic character of many real world human interactions—cooperation provides benefits to interactive partners, but it is also beneficial to take advantage of others. Figure 1 shows a typical prisoner's dilemma. The rows and columns model the choices—cooperate or defect—of the two interactive partners. Each cell in the chart represents one possible outcome with payoffs to the row actor listed first and column second. The unique Nash equilibrium of the game is Defect v Defect.

		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.

The stag hunt is a model of cooperation under risk. Suppose that two hunters can either choose to hunt for hare, or for stag. Taking down a stag is preferable because it provides more meat. But two hunters are needed to hunt stag, whereas a solo hunter can catch a hare. For this reason, stag hunting is risky. If one interactive partner does not choose to cooperate, the solo stag hunter gets nothing. The payoff table for this game is shown in figure 2. The Nash equilibria are Hare v Hare and Stag v Stag. The stag hunt may seem like a sub-ideal model for guilt, as there is no temptation to defect against a cooperator. In real world stag hunts, though, humans may be tempted in the moment to hunt hare (i.e., lazing around instead of building that shelter) and in such cases moral emotions, like guilt, might influence behavior.

4 Models

I will now look at evolutionary models that employ these two games in an attempt to understand how guilt can provide fitness benefits to individuals. This discussion will be divided into two parts. First, I consider how guilt can provide individual fitness benefits by *preventing* antisocial behavior both in the stag hunt and the prisoner's dilemma. The

[1]This method is similar to the 'indirect evolutionary approach', where actors evolve preferences that lead them to behave in ways that are beneficial overall, though they may be detrimental in a narrow interactive context (Güth 1995). I do not use this method because while emotions shape preferences, they also influence behavior in other ways (by creating states of arousal, for example).

		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.

second part considers whether guilt can provide individual fitness benefits by helping actors re-enter the social fold after behaving badly. Note that there is a very large literature on the evolution of prosocial behavior—here I point to the most relevant results from this literature, but do not attempt a survey.[2]

4.1 Guilt Before Defection

As discussed in section 2, empirical results indicate that guilt can influence human behavior by helping prevent failures of cooperation before they occur. If guilt is an underlying trait that leads to cooperation, any model where cooperative behavior provides individual benefit is a model where guilt could do so as well.

In the Stag Hunt

Suppose a population plays a stag hunt and that some significant proportion of the population plays stag (cooperates) when interacting with another agent. In such a case, any trait that promotes the choice of stag (cooperation) will benefit an individual agent. The reason for this is that each actor is more likely to meet a stag hunter than a hare hunter and if this occurs, the actor does best to choose to hunt stag as well. In such a scenario, the evolution of cooperation is not particularly mysterious, and neither is the evolution of any trait that promotes cooperation. Cooperation directly benefits fitness, with no further structure to the model (Skyrms 2004).

If one assumes that the ancestral state for early humans was a *lack* of cooperation, the stag hunt model seems less helpful. In a population where all actors are hunting hare, there is no individual incentive to hunt stag, and underlying traits that lead to more stag hunting will not be selected for. Mechanisms that lead to stag hunting from such a state have been investigated, however. Social structure, as modeled by interaction with neighbors in a social network, can allow cooperation to emerge in the stag hunt

[2] I ignore work on other games of cooperation. Discussions of group and kin selection models for prosocial behavior are outside the scope of this paper, which focuses on individual selection.

due to individual benefits as can the ability of actors to coordinate behavior with signals (Skyrms 2004; Alexander 2007). In these cases, an emotion like guilt that promotes prosocial behavior is directly beneficial.

In the Prisoner's Dilemma

The key to evolving cooperation in the prisoner's dilemma is *correlated interaction* (Axelrod and Hamilton 1981; Hamilton 1963). If cooperators meet cooperators, and defectors meet defectors in the prisoner's dilemma, two outcomes of the game—Cooperate v Cooperate and Defect v Defect—become more important than the rest. As figure 3 illustrates, if actors always (or often) meet their own types, it becomes beneficial to cooperate rather than to defect.

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

Figure 3: A payoff table for a prisoner's dilemma with correlated interaction.

Many mechanisms have been proposed whereby correlated interaction can occur in the prisoner's dilemma. Most of these fall under the broad categories of kin selection, group selection, indirect or direct reciprocity, and network reciprocity (Nowak 2006). Reciprocity can shape selective scenarios so that cooperation is individually beneficial.[3] If an actor in the right reciprocity type scenario switches to defection, he or she will no longer continue to meet cooperative partners, making cooperation (and thus guilt) individually beneficial.

In Response to Punishment

It should be noted that there is something a bit funny about the discussion just provided. Guilt, at least in modern humans, is evoked when actors break norms. In the models above, I suppose that guilt is evolving because it is tied to cooperative behavior *before* there are normative expectations for this behavior. The results above, then, are better thought of as applying to something like proto-guilt.

[3]Under group selection cooperators meet cooperators and so evolve, but in any particular case switching strategies to defect will be individually beneficial. Ditto for cases of network reciprocity. In kin selection, cooperation is individually beneficial from an inclusive fitness standpoint, but I do not consider inclusive fitness here.

The evolution of normative punishment has been supported by evolutionary models (Boyd et al. 2003; Boyd and Richerson 1992; Okamoto and Matsumura 2000). More importantly, for our discussion, it is an empirical fact that humans punish norm violators (Fehr and Gächter 2002; Ostrom et al. 1994). In a population that punishes defectors, defection becomes individually costly. Any trait, such as guilt proneness, that prevents accidental defection (or decreases temptation to defect on the part of the individual) will provide an individual selective advantage in such a social environment (Boyd et al. 2003; Boyd and Richerson 2009).

4.2 Guilt After Defection

I now turn to the question of whether guilt can provide individual fitness benefits to actors who have *already* defected. As discussed in section 2, in these cases guilt seems to harm individual actors by leading to confession and an increased chance of being caught, as well as to costly reparative behaviors and punishment. On the other hand, it can lead to apology and forgiveness, and to decreased punishment from other individuals. I explore the possibility that guilt is actually beneficial in such cases because it allows future potential partners to recognize underlying cooperative tendencies despite recent anti-social behavior and so forgive guilt prone types.

Costly Apology

The evolutionary game theoretic literature on behavior after defection focuses on a game called the iterated prisoner's dilemma (IPD). In this game, two agents play the standard prisoner's dilemma some number of times.[4] Strategies in this game include choices like the well studied 'grim trigger' (GT)—cooperate until my partner defects, and defect after that. Players may also just choose to cooperate unconditionally (C) or defect unconditionally (D). Another strategy that has been widely considered in this game is tit-for-tat (TFT) where actors cooperate on the first round of interaction and after that copy whatever their interactive partner did the round before.

This literature also commonly employs models where actors sometimes err. For example, an actor may be inclined towards unconditional cooperation, but defect in each round with some probability. This aspect of the model captures the idea that otherwise prosocial individuals may behave badly by accident, because of temptation, or due to exigencies of a particular situation.

Both GT and TFT are strategies where actors correlate interaction through reciprocity. In both of these strategies, actors will tend to cooperate with other cooperators and defect with other defectors, and for this reason both strategies can be evolutionarily

[4] Or else they play it for an unspecified length of time where at each round there is some probability that the game ends.

successful in the IPD (Axelrod and Hamilton 1981; Axelrod 2000). Both strategies, however, have problems when their interactive partners are prone to error. Suppose two GTs are playing the IPD and one accidentally defects. Her partner will immediately enter a state of permanent defection, and she will likewise do so. If two actors playing TFT interact, and one accidentally defects, the partner will defect, causing the original defector to defect again, and so on. In these cases although both actors are cooperative, they enter a spiral of defection where they lose payoff (Nowak 2006).[5] On an intuitive level, it makes sense that retaliation is useful as a way to punish and avoid bad cooperative partners. But, as these results suggest, it is good to have a way out of retaliation.

Theorists have attempted to solve this problem through the introduction of apology to these games (Okamoto and Matsumura 2000; Han et al. 2013; Ohtsubo and Watanabe 2009; Ho 2012). In apologetic strategies, actors who are otherwise cooperatively inclined, but defect through error, apologize to their interactive partners and are readmitted to the social fold. One necessity for these apologizing strategies to be stable in a population is that the apologizers pay a cost either directly or through punishment (Han et al. 2013; Okamoto and Matsumura 2000). These costs are necessary to prevent the invasion of 'faker' strategies where one apologizes, is readmitted to the social fold, and continues to defect. If apology bears a cost, it will not be worthwhile for fakers to apologize because the benefits of defecting again in the next round will not be high enough. For those with cooperative intent, the costly apology is worth paying in exchange for a long, fruitful cooperative interaction.[6]

Given these results, it is striking that after defection guilt in humans leads to a suite of behaviors—reparation, a willingness to accept punishment, and self punishment—that are individually costly. This points at a way that guilt, perhaps surprisingly, provides individual fitness benefits. Guilt prone types provide costly signals of their cooperative intent that would not be worthwhile to send unless they actually wanted to continue to cooperate in the future.

Cost-Free Apology

While this literature seems to shed light on the function of guilt after defection, the models discussed do not perfectly match empirical observations. As discussed, expressions of guilt tend to lead to decreased punishment by group members rather than increased punishment. Note that for guilt prone types in the models just discussed, if their interactive

[5] There are a number of TFT variants that avoid this issue (Wu and Axelrod 1995; Nowak and Sigmund 1992; Nowak et al. 1993). Apology strategies can be thought of as alternatives to these solutions to the retaliation problem.

[6] Experimental evidence indicates that humans indeed make costly apologies and that these are more successful than cost free ones in many cases (Ohtsubo and Watanabe 2009; Ho 2012; Nelissen 2012). Guilt may play a role in motivating costly apology (Ohtsubo and Watanabe 2009).

partners could trust their apologies *without* exacting some cost, this would obviously be preferable.

There is a literature in evolutionary game theory on this sort of trustworthy signal of cooperative intent. For example, Robson (1990) uses models to show that if actors can establish a 'secret handshake', a behavioral signal correlated with a tendency to cooperate in the prisoner's dilemma, cooperation can evolve. Of course, these signals are vulnerable to fakers in the same way cost-free apologies are vulnerable to fakers. Frank (1988, 1987) argues that moral emotions, such as guilt, can be thought of as a special sort of signal of cooperative intent because moral emotions are, in fact, correlated with cooperative behavior, and, Frank argues, are difficult to employ for non-cooperators.

Frank focuses on one-shot prisoner's dilemmas where actors use signals of emotion to choose cooperative partners in the first place. But this also seems to point to a way that guilt could be individually beneficial after defection. Perhaps actors can use honest signals of guilt to convince wronged partners of their future cooperative intent without paying some cost to guarantee it.

Consider an IPD where actors play for some number of rounds, n. During each round, an actor errs with probability α. Consider the following available strategies: unconditional cooperation (C), unconditional defection (D), grim trigger (GT), tit for tat (TFT), and guilt prone versions of either TFT or GT. In a guilt prone grim trigger (GPGT), actors behave like grim triggers, but after each defection they apologize. They also accept apologies and continue to cooperate with others who send them. This means that in practice, when guilt prone types meet they behave as unconditional cooperators. For now, I assume it is impossible to fake these apologies, because they are guaranteed by emotional signals.

I consider replicator dynamics simulations both of populations where actors can play C, D, GT, or GPGT and where actors can play C, D, TFT, or GPTFT.[7] Under all parameter values considered ($\alpha = .01, .05$ and $n = 10, 100$), GP types were by far the most likely to evolve. For very low n, D types evolve, and in all runs of models with TFT and GPTFT, population states with a combination of TFT and C sometimes evolve.[8] In other words, when it can act as an honest signal of apology, guilt evolves. This result is robust even when signals of guilt are not always trusted by recipients, and when guilty types are more likely to be caught defecting than other types.

The results just presented, however, are not entirely convincing. As Deem and Ramsey (2015) point out, guilt does not seem to fit well into Frank's picture of moral emotions

[7] The discrete time replicator dynamics, employed here to generate simulation results, are formulated as $x_i' = x_i(\frac{f_i(x)}{\sum_{j=1}^{n} f_j(x)x_j})$ where x_i is the proportion of a population playing strategy i, $f_i(x)$ is the fitness of type i in the population state x and $\sum_{j=1}^{n} f_j(x)x_j$ is the average population fitness in this state.

[8] For more details on any of the simulations presented in this paper, and related simulations, please contact the author.

as reliable indicators of cooperative intent. Unlike many other emotions, guilt is not associated with stereotypical facial expressions or body positions. But, on the other hand, humans do spend effort signaling their guilt verbally. And there is evidence that humans are, at least to some degree, able to 'read' the cooperative intent of others (Brosig 2002; Frank et al. 1993). In other words, the pictures of guilt leading to costly-apology or to cost-free apology do not seem to entirely fit. In the next section, I discuss an intermediate possibility that may help.

Low Cost Apology

Huttegger et al. (2015) point out that the distinction between cases where signals are trustworthy because they are costly (like costly apology), and cases where signals are trustworthy because only certain types of senders are able to generate them (like unfakeable emotional signals) is a spurious one. These authors show that even if a signal is only somewhat hard to fake, this can decrease the necessary costs that those employing this signal must pay to guarantee that it is genuine.[9] This work may help unify the two ways discussed that signals of guilt can be trustworthy, and so help account for empirical observations of guilt after defection.

Again, consider an IPD where actors sometimes err. Suppose that actors can be C, D, GT, or GPGT.[10] Lastly, suppose that faker types (F) exist who act like defectors but are able to send signals of guilt. When GPGT types receive these signals from faker types, they forgive the fakers and continue to cooperate.

Assume the following: actors pay a cost, C, to attempt to signal their guilt, and even when such attempts are made, they are not always successful. Also assume that because GP types really do experience guilt, the probability that they are successful when signaling their guilt, P_{GP}, is generally higher than the probability that fakers successfully signal, P_F. In these models, as P_F decreases, the signal cost to ensure that fakers cannot invade a guilt prone population also decreases. In other words, even if signals of guilt are only somewhat trustworthy, this can change the level of punishment or reparation needed for apology to work. Figure 4 demonstrates this for games where $\alpha = .05$ and $n = 10$ or $n = 100$. This result holds as long as n is not too low and α is not too high.[11]

Suppose that instead of sometimes failing to signal, faker types experience a higher

[9] I am equivocating a little bit here. Huttegger et al. (2015) are referring to cases where a difference in signal cost between high and low types ensures that only high types send the signal. In the costly apology literature, costs for apologies are generally the same for fakers and non-fakers, but non-fakers get a greater benefit for signaling, meaning there is still a discrepancy in the signal benefit for the different types.

[10] I did not consider TFT in this case for simplicity sake, but there is reason to think the results should extend to TFT and GPTFT.

[11] To be more precise, for each P_F this figure shows the lowest cost C such that GPGT remains an ESS of the game.

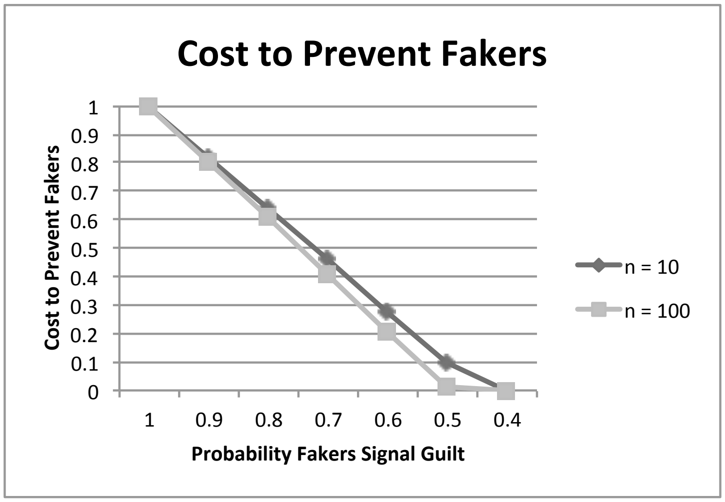

Figure 4: Costs necessary to stabilize populations of GP types against invasion by F types as the probability that F types successfully signal their guilt (P_F) drops.

direct cost when signaling their guilt. Because their verbal assertions of guilt are less convincing, they must spend more effort on reparative behavior, or accept greater punishments from group members to successfully signal. In the models at hand, this small change in believability also can stymie faker types.[12]

The take-away is that guilt after defection may function either as an honest signal of cooperative intent, as a mechanism leading to costly signals of cooperative intent, or as something in between. This in-between area seems to fit best with empirical observations of guilt.

5 Conclusion

It should be noted that these models do not explicitly account for the important role of cultural evolution and gene-culture coevolution in the evolution of guilt (Chudek and Henrich 2011; Henrich and Henrich 2006). This reservation noted, the models described give a broad set of cases where guilt might be individually selected for, whether or not the selective environment was shaped by culture and whether or not guilt itself is culturally created. Also, the work here involves limited runs of simulations, rather than more detailed analysis of the parameter space of these models, and so should be taken

[12]In models where GPGT and F types pay equal costs, GPGT generally has a small basin of attraction for the replicator dynamics. This is because F types still do fairly well against GPGT types and then D types outperform them. When F types pay higher costs than GPGT types, the basin of attraction for GPGT can be very large.

to indicate possible benefits for guilt rather than something deeper. Rosenstock and O'Connor (2015) are exploring similar analytic results, with similar outcomes.

Although an emotion like guilt may seem to be mainly group-beneficial[13], there are a number of plausible selective environments in which guilt, or something like it, can be individually beneficial in evolutionary settings. These selective environments, involving reciprocity, punishment, and apology, fit well with our empirical picture of human societies. As this paper illustrates, guilt may have adapted (or exapted) to play different roles in a complex and multifaceted developing human social environment.

References

Alexander, J McKenzie. 2007. *The structural evolution of morality.* Cambridge University Press Cambridge.

Axelrod, Robert. 2000. "On six advances in cooperation theory." *Analyse & Kritik* 22(1): 130–151.

Axelrod, Robert and William D Hamilton. 1981. "The evolution of cooperation." *Science* 211(4489): 1390–1396.

Bornstein, Brian H, Lahna M Rung, and Monica K Miller. 2002. "The effects of defendant remorse on mock juror decisions in a malpractice case." *Behavioral sciences & the law* 20(4): 393–409.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J Richerson. 2003. "The evolution of altruistic punishment." *Proceedings of the National Academy of Sciences* 100(6): 3531–3535.

Boyd, Robert and Peter J Richerson. 1992. "Punishment allows the evolution of cooperation (or anything else) in sizable groups." *Ethology and sociobiology* 13(3): 171–195.

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

Brosig, Jeannette. 2002. "Identifying cooperative behavior: some experimental results in a prisoner's dilemma game." *Journal of Economic Behavior & Organization* 47(3): 275–290.

Chudek, Maciej and Joseph Henrich. 2011. "Culture–gene coevolution, norm-psychology and the emergence of human prosociality." *Trends in cognitive sciences* 15(5): 218–226.

[13] See Deem and Ramsey (2015) for a discussion of group level benefits of guilt.

Deem, Michael and Grant Ramsey. 2015. “Guilt by Association.”.

Eisenberg, Theodore, Stephen P Garvey, and Martin T Wells. 1997. “But was he sorry? The role of remorse in capital sentencing.” *Cornell L. Rev.* 83:1599.

Fehr, Ernst and Simon Gächter. 2002. “Altruistic punishment in humans.” *Nature* 415(6868): 137–140.

Fischbacher, Urs and Verena Utikal. 2013. “On the acceptance of apologies.” *Games and Economic Behavior* 82:592–608.

Frank, Robert H. 1987. “If homo economicus could choose his own utility function, would he want one with a conscience?.” *The American Economic Review* 593–604.

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

Frank, Robert H, Thomas Gilovich, and Dennis T Regan. 1993. “The evolution of one-shot cooperation: An experiment.” *Ethology and sociobiology* 14(4): 247–256.

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

Güth, Werner. 1995. “An evolutionary approach to explaining cooperative behavior by reciprocal incentives.” *International Journal of Game Theory* 24(4): 323–344.

Hamilton, William D. 1963. “The evolution of altruistic behavior.” *American naturalist* 354–356.

Han, The Anh, Luís Moniz Pereira, Francisco C Santos, and Tom Lenaerts. 2013. “Why is it so hard to say sorry? evolution of apology with commitments in the iterated Prisoner’s Dilemma.” *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence.* 177–183.

Henrich, Joseph and Natalie Henrich. 2006. “Culture, evolution and the puzzle of human cooperation.” *Cognitive systems research* 7(2): 220–245.

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

Huttegger, Simon M, Justin P Bruner, and Kevin JS Zollman. 2015. “The handicap principle is an artifact.”.

Jehle, Alayna, Monica K Miller, and Markus Kemmelmeier. 2009. "The influence of accounts and remorse on mock jurors' judgments of offenders.." *Law and human behavior* 33(5): 393.

Joyce, Richard. 2007. *The evolution of morality.* Mit Press.

Ketelaar, Timothy and Wing Tung Au. 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, Tina and Tobias Krettenauer. 2013. "The Relation of Moral Emotion Attributions to Prosocial and Antisocial Behavior: A Meta-Analysis." *Child development* 84(2): 397–412.

Nelissen, Rob and Marcel Zeelenberg. 2009. "When guilt evokes self-punishment: evidence for the existence of a Dobby Effect.." *Emotion* 9(1): 118.

Nelissen, Rob MA. 2012. "Guilt-induced self-punishment as a sign of remorse." *Social Psychological and Personality Science* 3(2): 139–144.

Nowak, Martin, Karl Sigmund, et al. 1993. "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game." *Nature* 364(6432): 56–58.

Nowak, Martin A. 2006. "Five rules for the evolution of cooperation." *science* 314(5805): 1560–1563.

Nowak, Martin A and Karl Sigmund. 1992. "Tit for tat in heterogeneous populations." *Nature* 355(6357): 250–253.

Ohtsubo, Yohsuke and Esuka Watanabe. 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, Kyoko and Shuichi Matsumura. 2000. "The evolution of punishment and apology: an iterated prisoner's dilemma model." *Evolutionary Ecology* 14(8): 703–720.

Ostrom, Elinor, Roy Gardner, and James Walker. 1994. *Rules, games, and common-pool resources.* University of Michigan Press.

Ramsey, Grant and Michael Deem. 2015. "Empathy, Culture, and the Function of Guilt.".

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

Robson, Arthur J. 1990. "Efficiency in evolutionary games: Darwin, Nash and the secret handshake." *Journal of theoretical Biology* 144(3): 379–396.

Rosenstock, Sarita and Cailin O'Connor. 2015. "Selective Advantages of Guilt.".

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

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

Svensson, Robert, Frank M Weerman, Lieven JR Pauwels, Gerben JN Bruinsma, and Wim Bernasco. 2013. "Moral emotions and offending: Do feelings of anticipated shame and guilt mediate the effect of socialization on offending?." *European Journal of Criminology* 10(1): 22–39.

Tangney, June Price, Rowland S Miller, Laura Flicker, and Deborah Hill Barlow. 1996. "Are shame, guilt, and embarrassment distinct emotions?." *Journal of personality and social psychology* 70(6): 1256.

Wu, Jianzhong and Robert Axelrod. 1995. "How to cope with noise in the iterated prisoner's dilemma." *Journal of Conflict resolution* 183–189.