Title : When it's Good to Feel Bad: Evolutionary Models of Guilt and Apology

Author : Sarita Rosenstock and 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 : Thanks to Michael Deem, Grant Ramsey, and especially Andrea Scarantino for comments and feedback on this work.

Abstract

We use results from evolutionary game theory to analyze the conditions under which guilt can provide individual fitness benefits, and so evolve. In particular, we focus on the benefits of guilty apology. We find that in evolutionary models 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. Philosophers interested in naturalized ethics, and emotion researchers, can employ these results to assess the plausibility of fuller accounts of the evolution of guilt.

When it's Good to Feel Bad: Evolutionary Models of Guilt and Apology

Abstract

We use results from evolutionary game theory to analyze the conditions under which guilt can provide individual fitness benefits, and so evolve. In particular, we focus on the benefits of guilty apology. We find that in evolutionary models 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. Philosophers interested in naturalized ethics, and emotion researchers, can employ these results to assess the plausibility of fuller accounts of the evolution of guilt.

1 Introduction

Some emotions provide fairly straightforward fitness advantages. Fear, for example, yields obvious evolutionary benefits, like avoiding predation by tigers. Guilt poses more of an evolutionary puzzle. Deem and Ramsey (2016b) point out that this emotion is associated with behaviors that, initially, seem maladaptive. Guilt-prone individuals behave altruistically, even in cases where they do not expect to be caught. Those who feel guilty after a transgression accept punishment readily, and even punish themselves. For this reason, one might think that guilt evolved for its benefits to human groups, rather than individuals. But the evolution of group beneficial and individually harmful traits is notoriously fraught. How do we account for the evolution of this puzzling emotion? Philosophers interested in the role of moral emotions as underpinnings of naturalized ethics have recently gotten interested in this question (Deem and Ramsey, 2016b,a; Ramsey and Deem, 2015; Joyce, 2007).

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, simpiciter, are not behaviors. O'Connor (2016) argues, however, that the extensive bodies

of literature from evolutionary game theory on the evolution of cooperation, altruism, and apology can be used to shed light on 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 also presents novel modeling work clarifying how guilt can benefit individuals by prompting apology. In this paper, we provide a much more thorough analysis of the benefits of guilty apology focusing, especially, on 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. As we will show, guilty apology is more likely to evolve when guilt is hard to fake, actors interact repeatedly over long periods of time, and the costs to apology are not too high. We argue that these models can help determine conditionals of the form 'ceterus paribus, if x obtains, guilt provides greater individual fitness benefits' that philosophers and emotions researchers can employ in forming and assessing more detailed accounts of the evolution of guilt.

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. Here we will give just a quick overview of these results. The first employs the famous prisoner's dilemma game, which we will describe at length in the next section, to model the evolution of altruism. By altruism 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*. (See Nowak (2006) for an overview of the evolution of altruism in the prisoner's dilemma.)

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. If a guilt-prone agent does not take advantage of a group member because they anticipate feeling badly about it, and they then escape ostracism by that group member and others in the group, guilt can provide a benefit to that agent.

When actors *punish* those who fail to behave altruistically, likewise altruism, and guilt, are directly beneficial to the individual. To give an example, if an actor anticipates the experience of guilt, and so chooses not to steal from a friend, this might be beneficial if that actor then subsequently escapes group punishment for their behavior. 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. For example, suppose two actors have agreed to hunt a stag together, or cooperate, but one is tempted to hunt a hare instead, i.e., to seek short term, less risky, payoff. If anticipation of guilt keeps this actor focused on the stag hunt, and their partner pulls through, they will

eventually receive greater rewards for sticking to the larger, if riskier, joint project. Alexander (2007) shows that cooperation in the stag hunt is especially likely to evolve in groups where the same actors tend to keep interacting, as in early human groups. (See Skyrms (2004) for more on the stag hunt and the evolution of cooperation.)

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. Imagine, for example, interactive partners who regularly share meat. Suppose that after one hunt a partner fails to do so because they are especially hungry. If these actors reciprocate, the slighted partner will fail to share meat after the next hunt, leading the other partner to fail to share in following interactions, and so on. 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). (In the next section, we will explain at length why this is so.) These results indicate 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.

Before continuing, we should note that the above analysis, and our work here, ignores potential group level benefits from guilt. Nor do we address kin selection models of altruism and cooperation. The idea behind such group and kin selection models is that while individuals may suffer costs from a trait, like altruism, or guilt-proneness, their group may benefit, so that ultimately the proportion of the trait increases despite individual costs. For theoretical discussion of such possible benefits of guilt, see Deem and Ramsey (2016b).

3 Model and Results

We now turn to models of guilty apology. 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" instead of "altruism" here 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. It should not be confused with cooperation as modeled in the stag hunt, where actors obtain mutual benefits.

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

Figure 1: A payoff table of the prisoner's dilemma. Strategies for player one are represented by rows. Strategies for player two are represented by columns. Entries to the table show payoffs for each combination of strategies with player one listed first.

Figure 1 shows what is called a payoff table for this game. Possible strategies for player one are represented by rows. Strategies for player two are represented by columns. Entries to the table show payoffs for each combination of strategies with player one listed first. This game is a dilemma because the expected outcome is for both players to defect, despite the fact that mutual cooperation is preferred by everyone. The strategy pair where both players defect is the only set of strategies for which no player can benefit by switching, or the only Nash equilibrium of the game.

In an iterated prisoner's dilemma agents repeat the prisoner's dilemma round after round with some probability n. One can equivalently think of 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 a result of such repeated interactions, 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. In this way, they cooperate with cooperators and defect with defectors, gaining the benefits from mutual cooperation and mitigating harm from defectors.

As briefly 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. These sorts of accidents occur for many reasons in the real world. Actors under difficult circumstances, or duress, may behave antisocially despite general prosocial tendencies. Actors might forget a cooperative agreement. Or

actors might simply not feel like being cooperative today. Accidental defection causes grim triggers to permanently defect on good cooperative partners. Mutual negative reciprocation of this sort is mutually damaging.[1] 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. Note that guilt-prone, in this context, is a behavioral strategy, rather than an explicit representation of actors who, in fact, experience some emotion. The idea is that the behavioral strategy corresponds to the actions that an actor who did experience guilt would take, and so investigating the success of this strategy can tell us something about the circumstances under which guilt is adaptive by dint of causing these behaviors.

A problem with the guilt-prone strategy is that an apology will not effectively signal guilt if defectors can also 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 faker apologizers, which we will call fakers, can take advantage of forgiveness among guilt-prone players.

As we mentioned in the last section, 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, 2016b). 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 successful with some probability less than one but greater than zero.

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. For example, imagine two actors, each of whom has stolen a cupcake and is deciding whether to issue a costly apology. The actor who is planning a long, cooperative

[1] O'Connor (2016) also looks at "tit-for-tat", another reciprocating strategy, and finds similar results in models that include it, instead of the grim trigger.

life with the cupcake maker will receive a large benefit from doing so. The actor who will steal a cupcake tomorrow only receives a small benefit before having to pay the cost again.

To summarize, 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—that those who usually cooperate in fact defect, or vice versa. (This, remember, is the condition under which reciprocation can be harmful, and apology is potentially useful.) The strategies in the population are:

C – Unconditional cooperation, or always cooperate in every round

D – Unconditional defection, or always defect in every round

GT – Grim trigger, cooperate unless your partner defects, and then defect for every following round

GP – Guilt prone, play grim trigger, apologize upon defection, and cooperate with those who apologize

F – Faker, always defect, apologize upon defecting

There is some probability $p \leq 1$ that fakers manage to signal their guilt. And in order to successfully apologize, actors pay a cost, as described above. To allow for the possibility that actually guilty types pay a lower cost than fakers to convince others of their guilt we define $c \geq d$ where c is the cost of apology for guilt-prone types and d for fakers. 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 appendix A.

To be explicit, the guilt-prone strategy in this model matches empirical observations of guilt after transgression in the following ways. Guilt-prone actors are more likely to apologize. They are willing to pay a cost to do so. Upon receipt of this apology, group members decrease their punishing behavior. And guilt-prone individuals actually are likely to behave prosocially in the future. By looking at the evolution of this strategy in the model, then, we hope to gain insight into the actual conditions under which guilty apology evolves.

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. This is the case because 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.

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. For this reason, an ESS analysis is a way of identifying strategies that have the potential to evolve in a model. Because of this potential, ESS analysis has been employed extensively in biology and the social sciences, to gain insight into evolutionary properties of many sorts of populations. Readers interested in learning more should start with Smith and Price (1973).[2]

In the models we consider, unconditional defection is always an ESS, because it always does better against itself than any other strategy does against it. Guilt-proneness is sometimes an ESS, because it too does better against itself than other strategies, though it is destabilized by successful fakers (who themselves are eventually replaced by defectors) under some parameter values. As it turns out, the guilt-prone strategy is evolutionarily stable against fakers in a sizable portion of the parameter space. This means that there are many conditions under which guilt-proneness can potentially evolve. As we will see, both higher cost of apology, 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 2 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. Note that this figure only shows conditions under which guilt is stable against fakers—more on other strategies in a minute.

This graph also indicates that for longer interactions (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 guilt-prone 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.

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, fakeability (p) no longer matters. Figure 3 shows the minimum length of interaction (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

[2]All this said, ESSes 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.

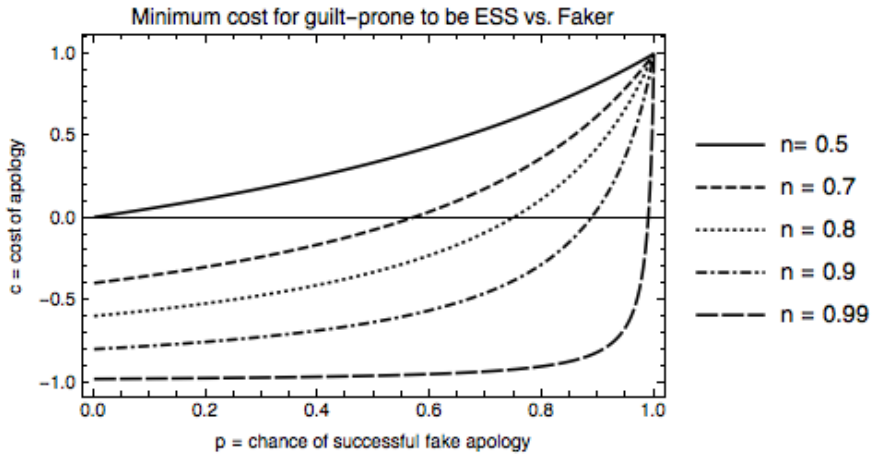

Figure 2: Minimum cost, c, for guilt-prone to be an ESS vs. faker, for each fakeability value p, with error rate $a = 0.01$. When faking is easier, as p goes to one, the cost for guilt-prone to be an ESS is higher. As n, increases, this cost is lower, because guilt-prone types do well in longer repeated interactions.

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 error rate, a, increases, length of play needs to be a bit larger for the guilt-prone strategy to be an ESS, but little changes in the range that we focus on. Note that in early human groups, we expect the length of repeated interaction to have been very high, meaning that guilt-prone should do well under these conditions.

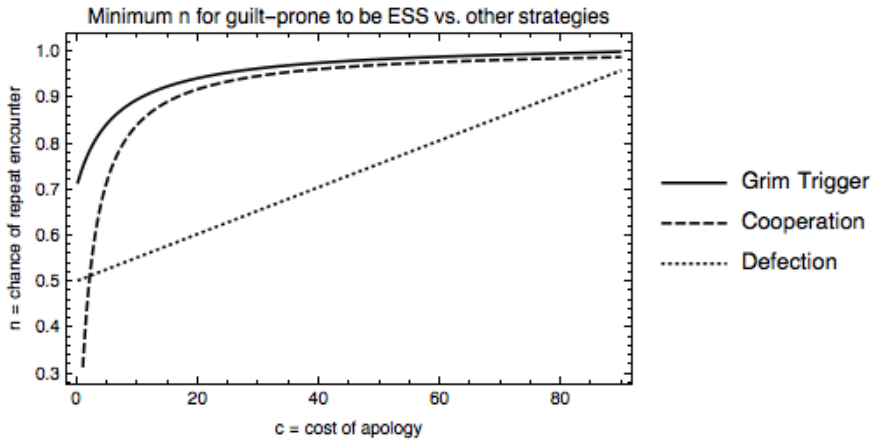

Figure 3: Minimum likelihood of repeated encounter, n, for guilt-prone to be an ESS against defectors, cooperators, and grim triggers, error rate $a = 0.01$. As the costs of apologizing increase, longer interactions are necessary for guilt-prone to be an ESS against all strategies.

Before continuing, it will be useful to take a moment to discuss the results just presented. The take-away from these models is an understanding of the conditions under which guilt, for the purposes of promoting

costly apology, can potentially evolve. This model is best understood not as a 'how-possibly' or just-so story for the evolution of guilt, but rather as a tool for researchers who do more detailed work on this evolution to evaluate the plausibility of various evolutionary pathways to modern, human guilt. It yields for us conditionals along the lines of 'if X, then guilt provides individual benefits/comes under positive selection pressure/will evolve ceterus paribus', that can be taken as evidence used to support or deny proper accounts of the evolution of guilt.

To summarize these conditionals, when humans interact for long periods of time, guilty apology is more evolvable, because in these cases there are long, fruitful interactions to be gained by those who can forgive and forget. When guilt is hard to fake, perhaps because humans are good at reading the emotions of other humans, again it has more potential to evolve. (In the same cases, note, there might be selection pressure to read these very emotions, if actors who can do so have a route to successful apology.) But actors need not be perfect emotion readers for guilt to be successful at promoting apology. For various levels of fakeability, there will be costs to apology that allow guilt to evolve. In other words, if guilt creates some individual costs to the agent, which, as discussed, occurs in real human populations in the form of self-punishment, acceptance of punishment from others, and reparations, this can help guilt evolve even in populations with fakers. As described, guilt-prone types reap a disproportionately high benefit from apologizing, which makes the costs worth their while, but not worthwhile for fakers.

In the next section, we will expand this analysis by looking, in detail, not just at *whether* guilt can evolve to promote apology under certain conditions, but how likely this evolution is.

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. ESS analyses are useful because they tell us something about which strategies have the potential to evolve. Some ESSes, 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.

What does this mean? The states of an evolutionary population correspond to the possible proportions of strategies in that population—50% fakers, 10% guilt-prone, and 40% cooperators, for example, or 100% grim triggers. In an evolutionary model using the replicator dynamics, each such state will evolve based on which strategies are doing relatively well. Successful strategies will expand, while less successful ones decline. Eventually this process will (usually) lead the population to an equilibrium, or a state where is

does not evolve anymore.[3] In our models, these equilibria are the ESSes of the game. Basins of attraction tell us what proportion starting states eventually end up at each ESS. For this reason, the size of a basin of attraction tells us something about the evolvability of a strategy. Equilibria with large basins are more likely to evolve, in a sense.

Also, mutations, or noise, in evolutionary processes, which tend to occur regularly in the real world, 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. Again, this means we should think of ESSes with large basins of attraction as likelier to evolve.

We thus want to ask: under what conditions does guilt-proneness, as represented in our models, have a larger 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 p, the probability that fakers successfully trick others into trusting their apologies, c, the cost of apology for guilt-prone players, 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 guilt-prone types pay the same cost. When p is low, guilt-prone types do well against fakers. For this reason, increasing costs actually makes guilt-proneness less likely to evolve. It simply decreases the payoffs to guilt-prone types, while failing to significantly help them differentiate themselves from fakers. When p is higher, cost can help guilt-prone types evolve. It allows them to prove their cooperative intent, compared to faker types. Figure 4 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 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 (C), unconditional defection (D), guilt-prone grim trigger (GP), and faker (F). The x-axis tracks cost, $c = d$, which ranges from .005 to 1. The y-axis shows the likelihood

[3]See O'Connor (2016) for a bit more on the replicator dynamics in these models.

that guilt evolves. For $p = 0.95$, when fakers are able to almost always convince others of their cooperative intent, the optimal cost for the evolution of guilty apology, of those explored, is 0.4. For $p = 0.9$, when fakers are slightly less successful, the optimal cost is 0.2. For the smaller values of p, costs make guilt less likely to evolve. In other words, the easier it is for fakers to convince others they are telling the truth, the higher the costs that make guilt-prone most likely to evolve.

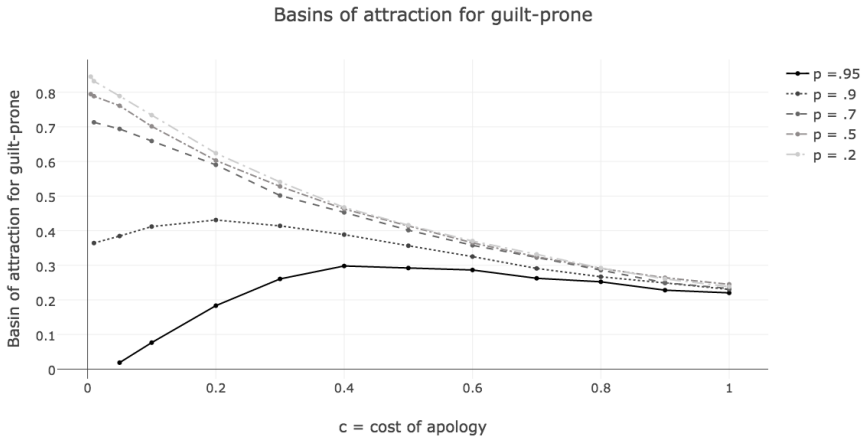

Figure 4: Sizes of basin of attraction for guilt-prone strategy as $c = d$, cost of apology for guilt-prone and fakers, and p, likelihood that fakers successfully apologize, vary. For high values of p, costs increase the basins of attraction for guilt-prone to a point. For low values of p, costs only hurt the evolvability of guilt-prone. Lower p always makes the basin of attraction of guilt-prone larger.

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 5 shows basins of attraction for guilt in these models with probability of successful faking (p) held fixed at 0.95 and error rate, $a = 0.01$. 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, meaning fakers pay various costs to apologize. 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, error rate, and n, probability of repeated interaction, fixed. For larger n, generally, guilt proneness will be more evolvable. As long as a is relatively small, changes to this parameter value do not significantly alter results. 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.

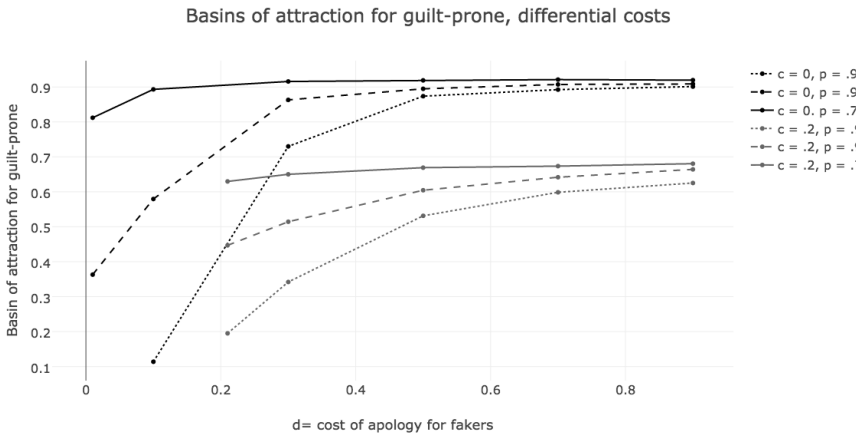

Figure 5: Sizes of basin of attraction for guilt-prone where $d \geq c$, or costs to fakers are higher than costs to guilt-prone to apologize. When p, probability of faking, is lower, or when c, cost for apology, is smaller, or when d, cost for fake apology, is larger, guilt-prone has a larger basin of attraction.

Again, these models generate a set of conditionals that now tell us something like, 'if X, then guilt-prone will be more *likely* to evolve, and to be stable, for the purposes of apology'. When guilt is easier to honestly convey, it is more likely that guilty apology will evolve and be stable. As in the last section, this makes sense. Guilt-proneness helps both players in the case of apology, one to identify a good cooperative partner even in the face of defection, and the other to convince group members to accept them back into the fold after messing up. As long as fakers can be kept at bay, this benefit obtains, and reading emotions helps this happen. Extra costs to fakers to apologize help guilty apology evolve, for the same reasons. These could obtain if, for example, group members are somewhat able to read emotions, and level extra costs for apology on those who do not seem genuine enough.

And lastly, costs for apologizing have a less straightforward impact on the likelihood that guilty apology evolves. They improve its chances when there are tricky fakers about by disincentivizing fake apology. On the other hand, they make it harder to sustain guilt because the costs straightforwardly harm the guilty individuals. The best conditions for the evolution of guilty apology would be those where actors can simply tell who is genuine and who faking. Of course, we should be so lucky.

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. It is also especially beneficial in repeated interactions. Costs for apology improve the evolvability of guilt when fakers are more successful, but hamper the success of guilt prone types otherwise.

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. Rather, we think these models provide insight whether or not guilt, and the environment that it evolved in, are culturally evolved. To put it another way, if we, as suggested, think of these models as giving us conditions under which guilt provides significant benefits, and so is more evolvable, these conditions may be produced by a culturally evolved social environment or a more straightforwardly biologically evolved one, and furthermore, they will provide benefits for culturally produced guilty behaviors as well as biological ones. Our if-then statements are broadly applicable. This is, of course, especially useful given that the details of the evolutionary environment of humans are sometimes murky. The mathematical models presented here are one more tool to use to gain clarity.

References

Alexander, J. M. (2007). *The structural evolution of morality.* Cambridge University Press.

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

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

Deem, M. and G. Ramsey (2016a). The evolutionary puzzle of guilt: Individual or group selection? *Emotion Researcher, ISREs Sourcebook for Research on Emotion and Affect, Andrea Scarantino (ed.)*.

Deem, M. and G. Ramsey (2016b). Guilt by association. *Philosophical Psychology 29*(4), 570–585.

Eisenberg, T., S. P. Garvey, and M. T. Wells (1997). But was he sorry? the role of remorse in capital sentencing. *Cornell L. Rev. 83*, 1599–1637.

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

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

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

Han, T. A., L. M. Pereira, F. C. Santos, and T. Lenaerts (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*, pp. 177–183. AAAI Press.

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

Huttegger, S. and K. Zollman (2013). Methodology in biological game theory. *The British Journal for the Philosophy of Science 64*(3), 637–658.

Joyce, R. (2007). *The evolution of morality*. MIT Press.

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

Nelissen, R. and M. Zeelenberg (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. (2015). Evolving to generalize: Trading precision for speed. *The British Journal for the Philosophy of Science*.

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

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

Ramsey, G. and M. Deem (2015). Empathy, culture, and the function of guilt. working paper.

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

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.

Smith, J. M. and G. Price (1973). The logic of animal conflict. *Nature 246*, 15–18.

Tangney, J. P., R. S. Miller, L. Flicker, and D. H. Barlow (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of personality and social psychology 70*(6), 1256–1269.

A Calculations of Expected Return

Here we show how we calculated the payoffs used in the above analysis for the iterated prisoner's dilemma.

The payoff table in figure 1 gives the expected return for a single round based on each of four possible action pairs performed by the players. However, this does not take in to account the fact that both players have a chance a of performing the wrong action, and that guilt-prone players and fakers pay a cost c when they defect, which changes their expected return from defecting. Moreover, we're primarily interested in total expected return over some unknown number of repeated interactions as determined by n. For some strategies, the action performed in a given round depends on actions of the other player in previous rounds: whether they made mistakes, and (for guilt-prone players vs. fakers) whether they effectively apologized for defecting.

We'll first calculate expected single-round payoffs for each intended action pair, and then use this information to calculate total expected return for a complete interaction between each pair of strategies. This total does not denote actual return from an interaction, as interactions involve probabilistic elements in the form of n, a, and p, but can be thought of as an average over a large number of such interactions.

Expected return each round

Here we show how to calculate each single-round action's expected return, by taking the sum of entry in the payoff table for player 1 multiplied by the probability that they get that payoff when performing that action, considering player 2's action, the error rate a, and the cost of apologizing c (if the strategy involves apology).

CC denotes the expected return for each player for a round in which both players intend to cooperate. This is given by:

$$CC = 2(1-a)^2 + 3a(1-a) + 0(1-a)a + 1a^2 = 2 - a$$

CC^* : expected return for a player who intends to cooperate this round, but will pay a cost of c to apologize if they accidentally defect, playing another player who intends to cooperate.

$$CC^* = 2(1-a)^2 + (3-c)a(1-a) + 0(1-a)a + (1-c)a^2 = 2 - a - ca$$

CD : expected return for a player who intends to cooperate this round, playing a player who intends to defect.

$$CD = 0(1-a)^2 + 1a(1-a) + 2(1-a)a + 3a^2 = 3a$$

CD^* : expected return for a player who intends to cooperate this round, but will pay a cost of c to apologize if they accidentally defect, playing a player who intends to defect.

$$CD^* = 0(1-a)^2 + (1-c)a(1-a) + 2(1-a)a + (3-c)a^2 = 3a - ca$$

DC : expected return for a player who intends to defect this round, playing a player who intends to cooperate.

$$DC = 3(1-a)^2 + 2a(1-a) + 1(1-a)a + 0a^2 = 3 - 3a$$

DC^* : expected return for a player who intends to defect this round but apologize for defecting, playing a player who intends to cooperate.

$$DC^* = (3-c)(1-a)^2 + 2a(1-a) + (1-c)(1-a)a + 0a^2 = 3 - 3a - c(1-a)$$

DD : expected return for a player who intends to defect this round, playing a player who also intends to defect.

$$DD = 1(1-a)^2 + 0a(1-a) + 3(1-a)a + 2a^2 = 1 + a$$

DD^* : expected return for a player who intends to defect this round but apologize for it, playing a player who also intends to defect.

$$DD^* = (1-c)(1-a)^2 + 0a(1-a) + (3-c)(1-a)a + 2a^2 = 1 + a - c(1-a)$$

Total expected return

In what follows, $Exp(A, B)$ denotes the total return payoff for a player using strategy A against a player using strategy B. The expected return each round is the sum of the each of the above single-round expected payoffs times the probability that they will be instantiated. When grim trigger players are involved, this probability is based on the error rate a for their opponent the previous round, as well as probability p of a faker's apology working in the case of guilt-prone players vs. fakers. The total expected return, then, is the

sum of the expected return each round i, times the probability n^i that there will be an ith round.[4]

$$
\begin{aligned}
Exp(C,C) &= \sum_{i=0}^{\infty} CC(n^i) = (2-a)\sum_{i=0}^{\infty} n^i = \frac{2-a}{1-n} \\
Exp(C,D) &= Exp(C,F) = \sum_{i=0}^{\infty} CD(n^i) = 3a\sum_{i=0}^{\infty} n^i = \frac{3a}{1-n} \\
Exp(D,C) &= \sum_{i=0}^{\infty} DC(n^i) = (3-3a)\sum_{i=0}^{\infty} n^i = \frac{3-3a}{1-n} \\
Exp(D,D) &= Exp(D,F) = \sum_{i=0}^{\infty} DD(n^i) = (1+a)\sum_{i=0}^{\infty} n^i = \frac{1+a}{1-n} \\
Exp(GP,GP) &= \sum_{i=0}^{\infty} CC^*(n^i) = (2-a-ca)\sum_{i=0}^{\infty} n^i = \frac{2-a-ca}{1-n} \\
Exp(GP,C) &= \sum_{i=0}^{\infty} \left[(1-a)^i CC^* + (1-(1-a)^i)DC^*\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(1-a)^i(2-a-ca) + (1-(1-a)^i)(3-2a-c)\right] n^i \\
&= \frac{2a^2n + ac - 4an + a + 2n - 2}{(n-1)(an-n+1)}
\end{aligned}
$$

[4]The more complicated of these infinite sums are calculated in Mathematica.

$$
\begin{aligned}
Exp(GP, D) &= \sum_{i=0}^{\infty} \left[(1-a^i)DD^* + a^i CD^*\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(1-a^i)(1+a-c(1-a)) + a^i(3a-ca)\right] n^i \\
&= \frac{-a^2cn - a^2n + 3acn - ac - 3an + 3a - cn + n}{(n-1)(an-1)} \\
Exp(F, D) &= Exp(F, F) = \sum_{i=0}^{\infty} DD^*(n^i) = (1+a-c(1-a)) \sum_{i=0}^{\infty} (n^i) \\
&= \frac{1+a-c(1-a)}{1-n} \\
Exp(F, C) &= \sum_{i=0}^{\infty} DC^*(n^i) = (3-2a-c) \sum_{i=0}^{\infty} n^i = \frac{3-2a-c}{1-n} \\
Exp(GT, D) &= Exp(GT, F) = \sum_{i=0}^{\infty} \left[(1-a^i)DD + a^i CD\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(1-a^i)(1+a) + 3a^{i+1}\right] n^i \\
&= \frac{-a^2n - 3an + 3a + n}{(n-1)(an-1)} \\
Exp(GT, C) &= \sum_{i=0}^{\infty} \left[(1-a)^i CC + (1-(1-a)^i)DC\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(1-a)^i(2-a) + (1-(1-a)^i)(3-3a)\right] n^i \\
&= \frac{3a^2n - 4an + a + 2n - 2}{(n-1)(an-n+1)} \\
Exp(F, GP) &= \sum_{i=0}^{\infty} \left[(p-ap+a)^i DC^* + (1-(p-ap+a)^i)DD^*\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(p-ap+a)^i(3-3a-c+ca) + (1-(p-ap+a)^i)(1+a-c+ca)\right] n^i \\
&= \frac{(a-1)(c(an(p-1) - np + 1) + n(a(p-1) + p + 2) - 3)}{(1-n)(an(p-1) - np + 1)} \\
Exp(GP, F) &= \sum_{i=0}^{\infty} \left[(p-ap+a)^i CD^* + (1-(p-ap+a)^i)DD^*\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(p-ap+a)^i(3a-ca) + (1-(p-ap+a)^i)(1+a-c+ca)\right] n^i \\
&= \frac{-a^2cnp + a^2cn - a^2np + a^2n + 2acnp - 3acn + ac + 3an - 3a - cnp + cn + np - n}{(n-1)(anp - an - np + 1)}
\end{aligned}
$$

$$
\begin{aligned}
Exp(C, GP) = Exp(C, GT) &= \sum_{i=0}^{\infty} \left[(1-a)^i CC + (1-(1-a)^i)CD\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(1-a)^i (2-a) + (1-(1-a)^i)(3a)\right] n^i \\
&= \frac{-3a^2 n - an + a + 2n - 2}{(n-1)(an-n+1)} \\
Exp(D, GP) = Exp(D, GT) &= \sum_{i=0}^{\infty} \left[(1-a^i)DD + a^i DC\right] n^i \\
&= \sum_{i=0}^{\infty} \left[(1-a^i)(1+a) + a^i(3-3a)\right] n^i \\
&= \frac{(a-1)((a-2)n+3)}{(1-n)(an-1)} \\
Exp(F, GT) &= \sum_{i=0}^{\infty} \left[a^i DC^* + (1-a^i)DD^*\right] n^i \\
&= \sum_{i=0}^{\infty} \left[a^i(3-2a-c) + (1-a^i)(1+a-c(1-a))\right] n^i \\
&= \frac{-a^2 cn - a^2 n + 2acn + 2an - 2a - c - 2n + 3}{(n-1)(an-1)}
\end{aligned}
$$

$Exp(GT, GT)$, $Exp(GT, GP)$ and $Exp(GP, GT)$ are a bit more complicated. Luckily, though, the first two are equal to one another and the third follows analogously. I computed $Exp(GT, GT)$ as follows:

Note that the expected return for round 1 in this case is:

$$E_1 = CC$$

Next, note that if we write the expectation for each round i as

$$E_i = x(i)CC + y(i)CD + z(i)DC + w(i)DD,$$

we get that

$$
\begin{aligned}
E_{i+1} = (1-a)^2 x(i)CC &+ (a(1-a)x(i) + ay(i))CD \\
&+ (a(1-a)x(i) + az(i))DC + (a^2 x(i) + (1-a)y(i) + (1-a)z(i) + w(i))DD.
\end{aligned}
$$

Considering x, y, z, and w as recursive functions with initial values $x(0) = 1$, $y(0) = z(0) = w(0) = 0$,

(and plugging into Mathematica) we get:

$$
\begin{aligned}
x(i) &= (1-a)^{2i-2} \\
y(i) &= z(i) = \frac{a(a^i-(1-a)^{2i}}{(a-1)((a-3)a+1)} \\
w(i) &= \frac{a(-2(a-1)a^i + a(1-a)^{2i} + (1-a)^{2i} + a - 3) - (1-a)^{2i} + 1}{(a-1)^2((a-3)a+1)}
\end{aligned}
$$

And at last we can express our final 3 expectations:

$$
\begin{aligned}
Exp(GT,GT) &= Exp(GT,GP) \\
&= \sum_{i=0}^{\infty} \Bigg[(1-a)^{2i-2} CC + \frac{a(a^i-(1-a)^{2i})}{(a-1)((a-3)a+1)}(CD+DC) \\
&\qquad + \frac{a(-2(a-1)a^i + a(1-a)^{2i} + (1-a)^{2i} + a - 3) - (1-a)^{2i} + 1}{(a-1)^2((a-3)a+1)} DD \Bigg] n^i \\
&= \sum_{i=0}^{\infty} \Bigg[(1-a)^{2i-2}(2-a) + \frac{a(a^i-(1-a)^{2i})}{(a-1)((a-3)a+1)}(3) \\
&\qquad + \frac{a(-2(a-1)a^i + a(1-a)^{2i} + (1-a)^{2i} + a - 3) - (1-a)^{2i} + 1}{(a-1)^2((a-3)a+1)}(1+a) \Bigg] n^i \\
&= \frac{-2+a-(-1+a)(2+a^2)n - a(1+(-3+a)a^2)n^2}{(-1+a)^2(-1+n)(-1+(-1+a)^2 n)(-1+an)}
\end{aligned}
$$

$$
\begin{aligned}
Exp(GP,GT) &= \sum_{i=0}^{\infty} \Bigg[(1-a)^{2i-2} CC^* + \frac{a(a^i-(1-a)^{2i})}{(a-1)((a-3)a+1)}(CD^*+DC^*) \\
&\qquad + \frac{a(-2(a-1)a^i + a(1-a)^{2i} + (1-a)^{2i} + a - 3) - (1-a)^{2i} + 1}{(a-1)^2((a-3)a+1)} DD^* \Bigg] n^i \\
&= \sum_{i=0}^{\infty} \Bigg[(1-a)^{2i-2}(2-a-ca) + \frac{a(a^i-(1-a)^{2i})}{(a-1)((a-3)a+1)}(3-c) \\
&\quad + \frac{a(-2(a-1)a^i + a(1-a)^{2i} + (1-a)^{2i} + a - 3) - (1-a)^{2i} + 1}{(a-1)^2((a-3)a+1)}(1+a-c(1-a)) \Bigg] n^i \\
&= \big[-(a^4 cn^2) + 5a^3 cn^2 - 4a^2 cn^2 + acn^2 - a^4 n^2 + 3a^3 n^2 - an^2 - a^3 cn - a^2 cn - a^3 n + a^2 n \\
&\qquad - 2an + 2n + ac + a - 2 \big] / \big[(-1+a)^2(-1+n)(-1+an)(-1+n-2an+a^2 n) \big]
\end{aligned}
$$