Kin-targeted Altruism With Noise*

Carmen Astorne-Figari †

This version: October 4, 2018

Abstract

Can pure altruism generate strategic altruism when kin recognition is noisy? This paper studies a prisoners' dilemma played between two individuals who exhibit altruistic preferences towards kin. The probability that a player's opponent is kin is common knowledge, but, instead of directly observing whether or not the other player is related, each player observes a noisy private signal. When the game is played once, players cooperate only with those identified as kin. However, when the prisoners' dilemma is played for two periods instead of one, uncertainty about relatedness brings strategic considerations into the game even if the odds of being related are small. There are Perfect Bayesian Equilibria in which players cooperate in the first round even after getting a negative kin signal. Since a player can make inferences about her opponent's signal based on first period actions, a non-relative mimics kin to induce cooperation in the second period.

Keywords: Altruism; noisy signaling; prisoners' dilemma

JEL Classification Numbers: C72, C73, D64

*The author has benefited from insightful comments from David Levine and John Nachbar. I also thank all participants in the WUSTL Monday Book and Discussion Group, and the Missouri Economics Conference.

†Department of Economics, University of Memphis, 427 Fogelman Admin Building, Memphis, Tennessee 38152; e-mail address: cmstrnfg@memphis.edu.

1 Introduction

Can repetition increase cooperation when players' preferences exhibit kin-targeted altruism and kin recognition is noisy? As suggested by kin selection models (Hamilton, 1964), players are altruistic towards relatives when such behavior increases inclusive fitness. Inclusive fitness, also known as "Hamilton's rule", is defined as an average of an individual's own survival probability and that of her kin's, weighted by the degree of relatedness between them. Favoring only kin and not those who are not related requires kin recognition mechanisms, which rely on using noisy signals to discriminate between kin and non-kin. We say that kin-targeted altruism occurs when players condition their strategies on these signals, choosing to cooperate only with those who exhibit the signal associated to kin. If, after observing the signals, the prisoners' dilemma is played for two periods instead of one, given that the signals are private, players' strategies may be used to reveal or to conceal information about the signal that they observed. Even if the probability of being related to one's opponent is not very high, repetition may generate an incentive to cooperate, either to conceal or to reveal such information.

Depending on the payoffs of the game and the degree of noise, we distinguish two cases. In both of them, a player who observes the relatedness signal acts like a behavioral type: either a grim trigger or a blind cooperator. In the first case, if the grim trigger punishment is strong enough, cooperation increases because players who observe the unrelatedness signal prefer to conceal their signal in order to avoid the grim trigger punishment (and defect against a cooperating opponent in the second period). However, if the grim trigger punishment is not strong enough, repetition generates the opposite result. In the second case, cooperation always

increases, because knowing that at least one player observed the relatedness signal is sufficient for both players to want to cooperate.

2 The model

Two players, 1 and 2, are randomly matched to play a Prisoner's dilemma, where players can choose to cooperate (C) or defect (D). When players meet, they may or may not be related to their opponent, represented by state space, $R = \{0, 1\}$, with $r \in R = 0$ meaning that players are not related, and $r = 1$ meaning that players are related.

First, we present an additively separable version of the prisoners' dilemma, shown below. To extend these payoffs to include relatedness, we will use Hamilton [1964]'s Rule from biology, also known as inclusive fitness.

Table 1: Payoffs to both players in a Prisoners Dilemma

		Player 2	
		C	D
Player 1	C	b, b	$-c, b+c$
	D	$b+c, -c$	$0, 0$

Both $b, c > 0$. This representation is convenient because all the information about the game that is of interest to us can be captured via the ratio $\frac{b}{c}$. A player whose opponent cooperates gets b. A player who defects on the other player gets c. If player 1 decides to cooperate with player 2 instead of defecting, she benefits player 2 in b dollars at a cost of c dollars that she could have kept for herself. Therefore, $\frac{b}{c}$ captures the cost of helping one's opponent in the original game.

Let $u_i(a_i, a_j)$ represent the payoffs denoted in Table 1. Using inclusive fitness,

we can calculate players' payoffs $U_i : A_i \times A_j \times R \to \mathbb{R}$ as follows.

$$U_i(a_i, a_j, r) = u_i(a_i, a_j) + r u_j(a_i, a_j) \tag{1}$$

That is, a player's individual payoff is augmented by her opponent's payoff weighted by the degree of relatedness between them. Thus, the payoffs shown in Table 1 represent inclusive fitness when players are not related ($r = 0$). That is, when players meet a non relative, the game they play is the original additively separable prisoners dilemma. When they meet a relative ($r = 1$), their payoffs are augmented by their opponent's payoff, using Hamilton's rule. For instance, suppose that players are related, player 2 cooperates and player 1 defects. Then $u_1(D, C) = b + c$, $u_2(C, D) = -c$, so player 1's payoffs are $U_1(D, C, 1) = (b + c) + (1)(-c) = b$ The payoffs to players who are related are shown in Table 2.

Table 2: Players' payoffs in state $r = 1$

		Player 2 C	Player 2 D
Player 1	C	$2b, 2b$	b, b
	D	b, b	$0, 0$

The probability that a player is related to his opponent, $P[r = 1] = z$, is common knowledge, which can also be interpreted as the probability of positive assortative matching. It is a known fact that the frequency of interaction with relatives favors cooperative behavior (See Bergstrom 1995 and 2003 for more references). In a population of cooperators and defectors, cooperation can be sustained when matching is assortative, since the cost of cooperating may be repaid by higher probabilities of playing against a cooperating opponent. We are not interested in the cases where matching is assortative, though. Instead, we will focus on cases

where being related is not that likely.

Given the payoffs described above, it is clear that players would want to cooperate if they knew they were facing a relative and to defect otherwise. However, in our model, players do not directly observe relatedness. Instead, they observe signals, which they use to attempt to discriminate between kin and non-kin. A large number of studies of animal populations as well as simulations performed by biologists provide evidence in favor of kin-targeted altruism through kin recognition mechanisms. These mechanisms operate through the recognition of an observable arbitrary trait that individuals have no control over, such as skin or eye color, which must also be heritable. It is often referred to as the "green beard effect" or the "armpit effect" (Dawkins, 1976). For instance, if an individual's family tends to have green beards, every time the individual meets someone with a green beard, he believes that he is very likely to be facing a relative.

In our model, kin recognition mechanisms rely on signals that are not perfectly informative, potentially leading to deception. In some of the cases, people will mistakenly identify individuals who are not kin as kin due to the presence of the trait. Conversely, players might dismiss a relative because the absence of the trait impeded his recognition. Ideally, the signals used for kin recognition should be traits that tend to be present in one's relatives and absent in non-relatives. However, given the multiplicity of characteristics that are present in human beings that people could check for, one cannot make sure that one's opponent is looking for the same trait that one is looking for, and, consequently, cannot be certain that one's own recognition is 100% correct. For example, player 1 could think that player 2 is a relative because he also has a green beard, but player 2 might not agree because their eye color is different. Therefore, we assume that the signal

observed by each player is private instead of public.

More formally, each player observes a signal $y_i \in Y_i = \{0, 1\}$. For simplicity, we assume that the distribution of the realized pair of signals $y \in Y_1 \times Y_2$ conditional on the state of the world is common knowledge. See Table 3.

Table 3: Conditional distribution of y

On state 0	On state 1
$\Pr[(0,0)\|r=0] = 1-2n$	$\Pr[(0,0)\|r=1] = 0$
$\Pr[(1,0)\|r=0] = n$	$\Pr[(1,0)\|r=1] = m$
$\Pr[(0,1)\|r=0] = n$	$\Pr[(0,1)\|r=1] = m$
$\Pr[(1,1)\|r=0] = 0$	$\Pr[(1,1)\|r=1] = 1-2m$

Both $m, n < 1/4$. This means that $y_i = 0$ is the unrelatedness signal, and $y_i = 1$ is the relatedness signal. Also, at least one of the two parameters m or n has to be strictly positive (otherwise there would be no noise). For simplicity, we assume that $\Pr[(0,0)|r=1] = 0$ and $\Pr[(1,1)|r=0] = 0$. That is, we assume that at least one of the players observes the "right" signal given the state of the world: if the state is unrelatedness, it will not be the case that both players observe the relatedness signal. The results of this paper are robust if these two probabilities are small enough. For example, if players knew that the signal realization is $(1,1)$, instead of knowing for sure that the state is relatedness, players would consider a very small probability that they might be unrelated.

2.1 The one-period game

Consider the following game: players meet, observe the private signal, and then play a one shot prisoners dilemma.

There are four pure strategies in this game:

1. Always cooperate.

2. Always defect.

3. Cooperate when observing $y_i = 1$ and defect otherwise.

4. Cooperate when observing $y_i = 0$ and defect otherwise.

We are interested in the case where players play kin-targeted altruism, namely, when players play according to strategy 3. We can calculate the distribution of relatedness conditional on the observed signal, y_i, shown in Table 4.

Table 4: Conditional distribution of r

On $y_i = 0$	On $y_i = 1$
$\Pr[r = 0 \mid y_i = 0] = \frac{(1-z)(1-n)}{(1-z)(1-n)+zm} = 1 - \alpha$	$\Pr[r = 1 \mid y_i = 1] = \frac{z(1-m)}{z(1-m)+(1-z)n} = \beta$
$\Pr[r = 1 \mid y_i = 0] = \frac{zm}{(1-z)(1-n)+zm} = \alpha$	$\Pr[r = 0 \mid y_i = 1] = \frac{(1-z)n}{z(1-m)+(1-z)n} = 1 - \beta$

Given these distributions, players can calculate expected payoffs to each pure strategy. See Table 5.

Table 5: Player i's expected payoffs conditional on y_i

Expected payoff given $y_i = 0$

		Player 2 C	Player 2 D
Player 1	C	$b + \alpha b$	$\alpha b - (1 - \alpha)c$
	D	$b + (1 - \alpha)c$	0

Expected payoff given $y_i = 1$

		Player 2 C	Player 2 D
Player 1	C	$b + \beta b$	$\beta b - (1 - \beta)c$
	D	$b + (1 - \beta)c$	0

Kin targeted altruism is the unique Nash equilibrium of this game if and only if the ratio $\frac{b}{c} \in [\frac{1-\beta}{\beta}, \frac{1-\alpha}{\alpha}]$.[1] When the relatedness signal is observed, $\frac{b}{c} \geq \frac{1-\beta}{\beta}$

[1]Notice that $\frac{1-\beta}{\beta} = \frac{(1-z)n}{z(1-m)} < \frac{1-\alpha}{\alpha} = \frac{(1-z)(1-n)}{zm}$ always, since $m, n < \frac{1}{4}$.

guarantees that the gains to cooperation are higher than the likelihood of mistakenly favoring a non-relative. On the other hand, when the unrelatedness signal is observed, $\frac{b}{c} \leq \frac{1-\alpha}{\alpha}$ guarantees that cooperating is too costly because non-relatives are more likely to be correctly identified.

Rearranging the terms, we can restate the conditions stated above as follows: $z \in [\frac{n}{(1-m)\frac{b}{c}+n}, \frac{(1-n)}{m\frac{b}{c}+(1-n)})]$. This can be interpreted in terms of assortativity: if positive assortative matching were too likely, players would blindly cooperate with any opponent, regardless of the observed signal, as explained earlier. On the other hand, if the likelihood of positive assortative matching were too low, players would prefer to defect all the time.

3 The two-period game

Suppose that players meet, observe their signal, and then play the prisoners dilemma for two periods. We want to know if repeating the game for one more period can induce more cooperation in games where kin-targeted altruism is an equilibrium.

Before introducing the two-period game, it is relevant to explore what would happen if players somehow knew the realized signal pair (y_1, y_2). They would prefer to cooperate if $(1, 1)$ takes place and to defect if $(0, 0)$ takes place, but what happens when the realization is $(1, 0)$ or $(0, 1)$ is not obvious. Since both cases are computationally equivalent, it is enough to analyze only one. We call the game where both players know that the signal realization is $y = (0, 1)$ "artificial" because it might never be played, but has important implications on the equilibrium of the two-times repeated game.

3.1 The artificial game

Suppose that both players know that the signal realization is $y = (0, 1)$ and they will play the game only once. In this case, both players know that one of them observed the wrong signal, but none of them knows who did. We can find the distribution of relatedness conditional on $y = (0, 1)$. Let $\gamma = \frac{zm}{zm+(1-z)n} = P[r = 1|y = (1, 0)]$. Then, we can write expected payoffs conditional on $y = (0, 1)$ as shown in Table 6.

Table 6: Player 1's expected payoffs in the artificial game

		Player 2	
		C	D
Player 1	C	$b + \gamma b$	$\gamma b - (1 - \gamma)c$
	D	$b + (1 - \gamma)c$	0

This game has a unique Nash equilibrium, which depends on the value of the ratio $\frac{b}{c}$. Defection is the unique Nash equilibrium of the artificial game if and only if $\frac{b}{c} \in [\frac{(1-\beta)}{\beta}, \frac{(1-\gamma)}{\gamma}]$. Conversely, cooperation is the unique Nash equilibrium if and only if $\frac{b}{c} \in [\frac{(1-\gamma)}{\gamma}, \frac{(1-\alpha)}{\alpha}]$. We will analyze each of these cases separately.

Proposition 1. *When players' preferences are altruistic towards kin and kin recognition is noisy, then there is strictly more cooperation in the twice repeated game if either:*

i. Cooperation is the unique Nash equilibrium of the artificial game, or

ii. The grim trigger punishment is strong enough.

The proof follows directly from Claims 1 and 2, which will be stated in the following subsections, and proved in the Appendix. The rest of the section is devoted to explain this result.

3.2 Games with defection in the artificial game

In this case, knowing the signal realization is $(0, 1)$, cooperating is too costly because favoring a non-relative is too likely. Given this, observing $y_i = 0$ is sufficient for wanting to defect in the twice repeated game. For instance, suppose that player 1 observed the unrelatedness signal 0. Learning which signal player 2 observed will not make a difference, since defecting is dominant in either case. Conversely, suppose that 1 observed $y_1 = 1$, so he would initially prefer to cooperate. Learning his opponent's signal would make a difference, since player 1 would rather defect whenever $y_2 = 0$. That is, learning the other player's signal can only induce more defection.

Thus, a separating equilibrium cannot increase cooperation in the twice repeated game. In the previous example, suppose that $y_2 = 1$. If player 1 (male) were given the choice to reveal or not reveal which signal he observed to player 2 (female), he would prefer not to reveal it. The reasoning is as follows. Given his signal, player 1's expected payoff is always higher when he defects instead of cooperating. If he keeps his information to himself, he knows that player 2 will cooperate, but if he chooses to reveal it, she will prefer to defect as well. Since defecting yields a higher payoff when his opponent cooperates, player 1 would be better off not revealing his signal. Hence, cooperation can only increase in the twice repeated game if players always cooperate in the first period. Note that a player who observes $y_i = 1$ acts like a grim trigger behavioral type: she cooperates as long as she doesn't find out that $y_j = 0$.

Let $\lambda = P[y_j = 0|y_i = 1] = \frac{zm+(1-z)n}{zm+(1-z)(1-n)}$. Given that this is a positive number, a player who observes the unrelatedness signal would only cooperate

during the first period to conceal her signal from her opponent (who observed the relatedness signal) and avoid the grim trigger punishment. That is, the grim trigger punishment must provide the right incentives for cooperation. We say that the grim trigger punishment is *strong enough* iff the following inequality holds.

$$\frac{b}{c} \geq \frac{1-\lambda}{\alpha+\lambda} \tag{2}$$

When a player observes the unrelatedness signal, she benefits from cooperating (i) when the opponent is related (with probability α), and (ii) when the opponent thinks that she is related (with probability λ). Her cost of cooperating is given by the times that she doesn't defect against a non-relative (with probability $1-\alpha$). The above condition implies that the benefit of cooperating in the first period and waiting until the second period to defect is greater than the benefit of defecting from the very beginning and being punished in the second period.

Claim 1. *When defection is the unique equilibrium of the artificial game and the grim trigger punishment is strong enough, the following strategy is a Perfect Bayesian Equilibrium of the original twice repeated game:*

- *Cooperate in the first period regardless of the observed signal.*
- *If $y_i = 1$ both players cooperated in the first period, cooperate in the second period. Defect in the second period otherwise.*

3.3 Games with cooperation in the artificial game

In this case, knowing the signal realization is $(0,1)$, cooperating is attractive because favoring a relative is more likely. Given this, observing $y_i = 1$ is sufficient

for wanting to cooperate, and player i acts like a blind cooperator in the twice repeated game. Thus, learning the other player's signal can never dissuade a cooperator from cooperating, but it can persuade a defector to cooperate.

Claim 2. *When cooperation is the unique equilibrium of the artificial game, the following strategy is a Perfect Bayesian Equilibrium of the original twice repeated game:*

- *If $y_i = 1$, cooperate in both periods.*
- *If $y_i = 0$, cooperate with probability ψ in the first period. In the second period, cooperate with probability ϕ if both players cooperated in the first period. Otherwise, defect in the second period.*

3.4 Can repetition generate the opposite result?

There is one case in which repetition strictly increases defection instead of cooperation. When there is defection in the artificial game and the grim trigger punishment is not strong enough, the effect of repetition is the exact opposite to the one described in the preceding subsections. In this case, players do not have incentives to conceal their signal from their opponents because the rewards of waiting one period to defect are too low. Therefore, in this case, private information is revealed at the end of the first period and players cooperate in the second period only if both of them observed the relatedness signal. In the rest of the cases, players defect in the second period. In other words, there is strictly more defection in the twice-repeated game compared to the case with no repetition.

Recall that a player who observes the relatedness signal always prefers to cooperate when he does not know which signal her opponent observed. When there

is defection in the artificial game, as seen in subsection 3.2, a player who observes the unrelatedness signal always prefers to defect in the second period. In the first period, she can cooperate, thus avoiding the grim trigger punishment at the cost of a lower payoff in that period, or she can defect, obtaining a higher payoff in the first period, but accepting the punishment in the second period. However, in this case, the grim trigger punishment is not strong enough, and thus, given her continuation strategy, the payoff of defecting is higher than the payoff of cooperating in the first period, as it can be seen in the proof of Claim 1. Then, the unique Perfect Bayesian Equilibrium of the twice repeated game is the following:

- If $y_i = 0$, play defect in both periods.
- If $y_i = 1$, cooperate in the first period. If cooperation is observed in the first period, cooperate in the second period. Otherwise, defect in the second period.

4 Conclusion

In games where players exhibit altruistic preferences towards relatives but kin recognition is noisy, playing a one shot prisoners' dilemma results in cooperation whenever players observe the relatedness signal. This kind of cooperation can be interpreted as pure altruism, since its only purpose is to favor kin. When the prisoners' dilemma is played for two periods instead of one, given that there is uncertainty about relatedness, other considerations are brought into the game, even if the odds of being related are small.

In the first case, players use their first period actions to dissuade their oppo-

nent from defecting. When there is defection in the artificial game and the grim trigger punishment is strong enough, learning that one's opponent observed the unrelatedness signal induces cooperators to defect. Cooperation increases in the first period because players who observe the unrelatedness signal want to conceal it from their opponent. In the second period, play is the same as in the one period game, and kin targeted altruism is played. However, when the grim trigger punishment is not strong enough, repetition generates the opposite effect and defection increases.

In the second case, players use their first period action to persuade their opponent to cooperate. When there is cooperation in the artificial game, a sufficient condition for wanting to cooperate is to know that at least one of the players observed the relatedness signal. Players who observe this signal cooperate to reveal their information to their opponents.

5 Appendix

5.1 Proof of Claim 1

Checking for perfection of the second period actions is immediate from the one-shot game, since no information is revealed at the end of the first period. We need to check for perfection of the first period actions. Assume that if a player observes defection in the first period, she attributes it to her opponent having observed the unrelatedness signal. Given this, and given that there is defection in the artificial game, it is easy to check incentives for the case when $y_i = 1$ is observed.

When $y_i = 0$ is observed, we need to verify that player i prefers to cooperate

in the first period. Calculate expected payoffs according to first period action:

	C	D
1^{st} period	$\lambda[b+\gamma b]+(1-\lambda)b$	$\lambda[b+(1-\gamma)c]+(1-\lambda)[b+c]$
2^{nd} period	$\lambda[b+(1-\gamma)c]$	0

Given that the grim trigger punishment is strong enough, the payoff of cooperating in the first period is higher than that of defecting.

5.2 Proof of Claim 2

Given that there is cooperation in the artificial game, incentives for the player observing $y_i = 1$ are easy to check because cooperation is always dominant.

Now, we check incentives when $y_i = 0$ is observed. In the second period, after defection is observed, player i knows for sure that the signal realization was $(0, 0)$ and, therefore, the state of the world is unrelatedness, so she prefers to defect. On the other hand, suppose that cooperation was observed in the first period. Let μ be the posterior probability that the opponent observed the relatedness signal. Player i's expected payoffs according to second period action are:

$$\begin{aligned} \text{Cooperating:} \quad & \mu[b+\gamma b]+(1-\mu)[\phi b+(1-\phi)c] \\ \text{Defecting:} \quad & \mu[b+(1-\gamma)c]+(1-\mu)\phi[b+c] \end{aligned}$$

Since player i is mixing, both payoffs must be equal. Solve for μ:

$$\mu = \frac{\lambda}{\lambda+\psi(1-\lambda)} = \frac{c}{\gamma(b+c)} \tag{3}$$

Given this equality, we can pin down the value of ψ:

$$\psi = \frac{\alpha(b+c)-\lambda c}{1-\lambda} \tag{4}$$

Now check for incentives in the first period. Calculate expected payoffs according to first period action:

	C
1^{st} period	$\lambda[b+\gamma b]+(1-\lambda)[\psi b-(1-\psi)c]$
2^{nd} period	$\lambda[b+\gamma b]+(1-\lambda)\psi[\phi b+(1-\phi)c]+(1-\lambda)(1-\psi)[\phi(b+c)]$

	D
1^{st} period	$\lambda[b+(1-\gamma)c]+(1-\lambda)[\psi(b+c)]$
2^{nd} period	$\lambda[b+\gamma b]-(1-\lambda)[\psi c]$

Given that player i mixes, set both payoffs equal to each other to solve for ϕ:

$$\phi=\frac{c-\alpha(b+c)}{(1-\lambda)(b+c)}$$

Given that $\frac{b}{c}\in[\frac{1-\gamma}{\gamma},\frac{1-\alpha}{\alpha}]$, both ϕ and ψ are positive numbers.

References

[1] Bergstrom, T. (2003). "The algebra of assortative encounters and the evolution of cooperation," *International Game Theory Review*. 5, No. 3, 1-18.

[2] Bergstrom, T. (1995). "On the evolution of ethical rules for siblings," *The American Economic Review* 85, Issue 1, 58-81.

[3] Binmore,K. (1998), *Just Playing. Game Theory and the Social Contract,* MIT Press.

[4] Dawkins, R. (1976). *The Selfish Gene*, Oxford University Press.

[5] Hamilton, W.D. (1964): "The Genetical Evolution of Social Behavior I," *Journal of Theoretical Biology* 7, 1-16.