Competitive Altruism, Mentalizing and Signalling

scottish institute for research in economics

SIRE DISCUSSION PAPER

SIRE-DP-2010-90

Competitive Altruism, Mentalizing and Signalling

Ed Hopkins

University of Edinburgh

www.sire.ac.uk

Competitive Altruism, Mentalizing and Signalling

Ed Hopkins*
School of Economics
University of Edinburgh
Edinburgh EH8 9JT, UK

October, 2010

Abstract

The human tendency to cooperate with nonkin even in short-run relationships remains a puzzle. Recently it has been hypothesized that altruism may be a byproduct of "mentalizing", the process of understanding and predicting the mental states of others. Another idea is based on sexual selection: altruism is a costly signal of good genes. The paper shows that these two arguments are stronger when combined in that altruists who can mentalize have a greater advantage over non-altruists when they can signal their type, even though these signals are costly. Further, once such an equilibrium is established, altruists will not be supplanted by mutants who have similar mentalizing abilities but who lack altruism.

Keywords: altruism, sexual selection, mentalizing, social preferences, signalling, tournaments, evolution.

JEL codes: C73, D64, D83.

*I thank Sevi Mora, Marco Faravelli and Tatiana Kornienko for helpful discussions. Errors remain my own. E.Hopkins@ed.ac.uk, http://homepages.ed.ac.uk/hopkinse

1 Introduction

One of the biggest puzzles in social science remains that of understanding cooperation in human society. Existing explanations have usually been based either on the theory of inclusive fitness or on the theory of repeated games. Yet, there is much evidence that people cooperate with unrelated individuals even in short run or one shot encounters. An alternative theory that sees prosocial activities as an attempt to signal desirability to potential mates has been proposed by Zahavi (1975) and Miller (2000). This sexual selection explanation of cooperation has been modelled formally by Gintis, Smith and Bowles (2001). They demonstrate that an equilibrium exists where a high quality individual can successfully signal that quality to potential partners by engaging in costly prosocial activity. This has been called "competitive altruism" (Roberts, 1998). Griskevicius et al. (2007) present supportive experimental evidence for the signalling role of prosocial behavior. They find that romantic thoughts can increase willingness in men and women to provide public service (see also Iredale et al. (2008)).

Another recent hypothesis is that altruism is a byproduct of a combination of empathy and a theory of mind. Perceptions of the emotional state of another leads to a representation of that state in the mind of an observer (de Waal, 2008). Building on this basic capacity for empathy, humans have the ability, which has been called "mentalizing", to reason about others' mental states. Possession of this ability allows prediction of others' actions, which is clearly advantageous. But this consideration of the others' emotional states may lead us to be other-regarding by default (Singer and Fehr, 2005).

There are problems with both theories. The signalling hypothesis does not explain why quality is signalled by doing good, when it could be equally well signalled by any costly activity (a problem noted by Gintis et al. (2001) themselves). After all, the leading example of sexual selection is the peacock's tail, where quality is signalled by the investment of resources into conspicuous waste. Or in a modern social context, why signal your wealth by giving to charity when you could also do so by conspicuous consumption or simply by burning money? Indeed, Griskevicius et al. (2007) also find that romantic thoughts increase men's willingness to engage in conspicuous consumption.

The explanation based on the theory of mind has a different question to answer. Even if empathy and the theory of mind evolved together, why have they remained linked? Specifically, since altruism is often costly, why are those individuals who have both altruism and a theory of mind not evolutionarily supplanted by those who are mentally sophisticated but not altruistic? There seems to be an unexploited opportunity to take the benefits without paying the costs.

This paper shows how it might be possible to solve both problems simultaneously. Suppose prosocial behavior is an equilibrium signal not of quality or wealth but of virtue or, more specifically, altruism. If the relevant signal is the level of contribution to a public

good, this solves the signal selection problem as if altruists wish to distinguish themselves from non-altruists, it is precisely in giving or contributing to a public good that they have a comparative advantage.[1] Further, since it would be necessary to make these visible contributions in order to attract favorable matching opportunities, those who did not undertake such public prosocial activities would not match as well. Thus, those who have a theory of mind but not altruism, would have to make the same contributions as altruists, and therefore would have no fitness advantage.

I assume that one group of individuals, contributors can be either altruists or non-altruists. Contributors have an opportunity to signal their type by their choice of contribution to a public good that will be seen by another group, observers. One possible interpretation is that the two groups represent the two genders. In any case, the observers, on the basis of the contributions they have witnessed, then choose with which contributor to match. Once matched in a pair, a contributor and an observer engage in a joint project, such as raising children, the success of which depends on the quality of the observer. A contributor's fitness depends on the total production of public goods, minus his own contribution, plus the outcome of this project. Altruists have additional subjective utility from contributions and thus will contribute more than non-altruists. However, they may gain more favorable matches, as observers are assumed to prefer to match with altruists.

Thus, altruists potentially have higher fitness if they can gain more in improved matching opportunities than they lose in additional costs of contribution. I find that the net effect is positive if and only if altruism is combined with superior ability in the post-match project. An example of this would be if altruists were superior at mentalizing and mentalizing was beneficial.[2] It is also shown that if altruists do not have superior ability, then the equilibrium cost of signaling will be higher than the benefits achieved, and that altruism will not be evolutionarily stable.

There are two apparent problems with the approach taken in this paper. First, if altruists have an advantage relative to non-altruists, for example, in mentalizing, is signalling needed? One might suppose that altruists will supplant non-altruists simply because they are better. Second, why are altruists not displaced evolutionarily by others that save unnecessary costs by not behaving altruistically?

In fact, signalling and altruism reinforce each other. First, I show that the advantage of altruists over non-altruists is larger when there is signalling than when contributions to the public good are not observed. This is the case even though with signalling altruists

[1] Millet and Dewitte (2007) find that giving in a public goods game and a separate measure of altruism are positively related with general intelligence.

[2] Dunbar and Shultz (2007) suggest that it is the prevalence of pair bonding in long term relationships in humans that has been important in developing human intelligence. That is, mentalizing is an important factor in successful pair bonding.

expend more effort on providing the public good. This is because the extra effort is more than compensated by the higher returns from the post-match project due to the better matching that follows once altruists identify themselves by signalling.

Second, signalling prevents altruists being supplanted by non-altruists who are equal at mentalizing. In a separating signalling equilibrium, any non-altruistic individuals would be forced to make the same level of prosocial contributions as altruists in order to gain favorable matches. Thus, they would have no advantage in fitness over altruists.

The approach used here is related to the indirect evolutionary approach that already has been used to explain human cooperation (Frank (1987), Güth (1995)). Under the indirect approach, individuals choose rationally given their preferences, but these preferences may not be identical with their objective self-interest or fitness. In particular, they may have altruistic preferences. But evolution will then select between preferences on the basis of actual fitness. Recent criticism in economics (Dekel et al. (2007)) of this approach has focussed on its assumption that agents' preferences are observable by other agents. Further, West and Gardner (2010) doubt whether cooperation based on cooperative or altruistic types being identifiable could be evolutionary stable: non-altruistic types with the same external appearance could invade. However, here I do not assume that individuals' preferences are observable. Rather, it is only if an individual's type is revealed by equilibrium behavior will observers know whether he is an altruist or not.

2 Signalling Altruism

There are n individuals which I will call contributors as all of them have to choose simultaneously and independently how much to contribute to the production a public good. Let the contribution of contributor i be x_i, then the total contributions will be $\sum_{j=1}^{n} x_j$ and the total amount of the public good be $G(\sum_{j=1}^{n} x_j)$, where G is a strictly increasing, smooth concave production function. Let us also assume that $G'(0) \geq 1$ and $\lim_{x\to\infty} G'(x) = 0$ (simple examples of suitable functions include $\log x$ and $\sqrt{x}$).

Following the contributors' choice of contribution, there will be an opportunity to match with another set of individuals, whom I will call observers. The observers see the contributors' choice of contribution before making their decision about which contributor to match with. Let the parameter s give the value of the match for the contributor.

As in the indirect evolutionary approach, an individual's utility may not coincide with her actual material payoff or fitness. Here, each agent's fitness is

$$\Phi_i = -x_i + G(\sum_{j=1}^{n} x_j) + \pi(\alpha_i, s_i) \tag{1}$$

which is increasing in the amount of the public good produced less an agent's contribution. The final term $\pi(\alpha_i, s_i)$ is the return in terms of matching opportunities. How this is determined will be described later.

In contrast to the material payoff which is the same for all contributors, some contributors have an altruistic preference for the welfare of others. Specifically, the utility of an individual i will be

$$U_i = -x_i + G(\sum_{j=1}^{n} x_j) + \frac{\alpha_i}{n-1} \sum_{j \neq i} (G(\sum_{j=1}^{n} x_j) - x_j) + \pi(\alpha_i, s_i) \tag{2}$$

where α_i is an altruism parameter. Importantly, let us assume there are $m \geq 1$ individuals with $\alpha_H > 0$ and $n - m$ with $\alpha_L = 0$. So non-altruists' ($\alpha = 0$) utility is identical to their material payoffs. However, altruists ($\alpha_H > 0$) care positively about the material payoffs of others, and thus, their preferences are different from their material payoff.

The proportion of altruists and non-altruists is known by all. However, in contrast to much of the literature using the indirect evolutionary approach, an agent's type is not known by the observers or other contributors.

I now turn to how the matching term $\pi(\alpha_i, s_i)$ is determined. The fundamental assumption is that each agent's choice of contribution x_i is observed by potential matches. These observers, possibly members of the opposite sex, cannot see an agent's type, only his choice of contribution. Contributors know that their choice of contribution are observed by potential matches.

I assume that observers prefer to match with altruists than with non-altruists. Thus, with complete information so that contributors' types were known, altruists would match better than non-altruists. Specifically, if a contributor's type was directly observable, an altruist would match with an observer of quality s_H and a non-altruist would have the worse outcome $s_L < s_H$.

Further, and this is crucial, the total return to a contributor of type α_i from matching with an observer of quality s_j is $\pi(\alpha_i, s_j)$, where π is a smooth function that is increasing in both arguments and $\pi_{\alpha s} = \partial^2 \pi / (\partial \alpha \partial s) > 0$. A simple example of such a function is $\pi(\alpha_i, s_i) = \alpha_i s_i$, the match return is the product of the contributor type and the observer type. It implies that an altruist α_H receives a higher payoff when matching with an observer of quality s_j than a non-altruist would. Further, an increase in match quality has a bigger effect on the return of an altruist than a non-altruist. This assumption corresponds with the idea that empathy and mentalizing are positively associated, so that the altruists are superior at mentalizing and that this gives them a higher return from matching than non-altruists.

However, as a useful benchmark, I first look at what contribution agents would choose in the absence of signalling considerations. That is, I look at the Nash equilibrium

of the public goods game assuming the additional term $\pi(\alpha_i, s_i)$ in (2) is independent of the choice of contribution. For example, it could be zero for both altruists and non-altruists. Let us call a Nash equilibrium where are altruists make the same choice, and all the non-altruists choose the same contribution (but not the same as the altruists), "quasi-symmetric". Then, there is the following preliminary result.

Proposition 1. *Suppose matching success is independent of one's choice of contribution, then there is a quasi-symmetric Nash equilibrium in which all m altruists choose the same contribution* $x_H^0 > 0$ *and all* $n - m$ *non-altruists choose the same contribution* $x_L^0 = 0$*. There is no other quasi-symmetric Nash equilibrium.*

Proof: Suppose that all the non-altruists choose zero. Then the altruists have an incentive to contribute as their marginal incentive to contribute $-1 + (1 + \alpha_H)G'(0) > 0$ is positive at zero total contribution. Further, as by assumption the marginal product of G falls to zero as contributions become large, one can increase the quantity chosen by the m altruists x_H up to a level x_H^0 such that

$$(1 + \alpha_H)G'(mx_H^0) = 1 \tag{3}$$

and thus the altruists have no incentive to raise their contribution further. But then it must be that $G'(mx_H^0) < 1$ so that the marginal incentive to contribute for the non-altruists is negative. So, they have no incentive to increase their contribution from zero and this strategy profile is an equilibrium. Given the concavity of G, if $x_L^0 = 0$, the contribution x_H^0 that satisfies the equilibrium condition $(1 + \alpha_H)G'(mx_H^0) = 1$ is unique. Lastly, clearly, there is no pair (x_L^0, x_H^0) with $x_L^0 > 0$ such that both types can be in equilibrium, as $(1 + \alpha_H)G'(mx_H^0 + (n - m)x_L^0) = 1 = G'(mx_H^0 + (n - m)x_L^0)$ is an impossibility. □

That is, even in the absence of signalling, altruists will contribute more than non-altruists. The point is this gives a quite natural story about how initial differences in behavior could arise. One would expect this would have made it easy for observers to learn how to distinguish types on the basis of their contributions, even before signalling behavior evolved.

The main results are, first, to show that there exists a separating equilibrium, where altruists choose a different level of contribution than non-altruists and, therefore, are identifiable by observers; second, to determine in such an equilibrium which type has a fitness advantage. For equilibrium, we need a contribution level for the high types x_H and a contribution level for the low types x_L where $x_H > x_L$ such that neither type wishes to deviate. Given the distinct choices of the two types, in equilibrium observers will correctly conclude that a contributor choosing x_H is an altruist and one choosing x_L is not. Thus, the matching return to the choice x_H will be s_H and the return to x_L will be s_L.[3]

[3]To determine the return to a choice of contribution that is neither x_H or x_L, one must specify appro-

Consequently, the only way for a low type to obtain the high matching return s_H will be to imitate the high types and choose x_H. Thus, the principal incentive compatibility (IC) condition for a separating equilibrium is that a low type must gain a higher utility from not imitating, or

$$U(\alpha_L, x_H, s_H) = -x_H + G(\bar{X}) + \pi(\alpha_L, s_H) \leq -x_L + G(X) + \pi(\alpha_L, s_L) = U(\alpha_L, x_L, s_L) \quad (4)$$

where X is the equilibrium total contribution $X = mx_H + (n-m)x_L$, and $\bar{X}$ is the total contribution if one low type deviates, or $\bar{X} = (m+1)x_H + (n-m-1)x_L$.

Equally, if a high type deviates to any contribution lower than x_H, she will only obtain s_L. Given this, the incentive compatibility constraint for a high type not to want to deviate to a lower contribution $x_L \in [0, x_H)$ will be

$$-x_H + (1+\alpha_H)G(X) + \pi(\alpha_H, s_H) \geq -x_L + (1+\alpha_H)G(\underline{X}) + \pi(\alpha_H, s_L) \quad (5)$$

where $\underline{X} = (m-1)x_H + (n-m+1)x_L$ or the total contribution if one high type deviates to x_L.[4]

In fact, it is easy to find contribution levels x_H, x_L that satisfy these IC conditions and, therefore, constitute a separating equilibrium. As in the original Spence signalling model, there will be a continuum of such separating equilibria.[5]

Proposition 2. *For any m such that $n > m \geq 1$, there exists an interval $[\underline{x}_H, \bar{x}_H]$ such that if $x_H \in [\underline{x}_H, \bar{x}_H]$ then the pair $\{x_H, x_L = 0\}$ satisfy the incentive compatibility conditions (4) and (5) and therefore constitute a pure strategy separating equilibrium.*

Proof: Again define $\underline{x}_H$ as the contribution x_H that solves the first IC condition (4) with equality and define $\bar{x}_H$ as the equivalent quantity from the second IC condition (5). We have $\underline{x}_H < \bar{x}_H$ if

$$G(\bar{X}) - G(X) + \pi(\alpha_L, s_H) - \pi(\alpha_L, s_L) < (1+\alpha_H)(G(X) - G(\underline{X})) + \pi(\alpha_H, s_H) - \pi(\alpha_H, s_L).$$

This holds as $\pi(\alpha_L, s_H) - \pi(\alpha_L, s_L) < \pi(\alpha_H, s_H) - \pi(\alpha_H, s_L)$ is true because $\pi_{\alpha s} > 0$ by assumption, and because as $\bar{X} - X = X - \underline{X} = x_H - x_L$, one has $G(X) - G(\underline{X}) \geq$

priate out-of-equilibrium beliefs. A sufficient condition for this form of separating equilibrium to hold is that the observers believe that any agent choosing a contribution $\hat{x}$ less than x_H must be a non-altruist. For simplicity, this is what I assume.

[4]There is third incentive compatibility condition that the separating contributions must be at least as large as would be chosen in the absence of signalling considerations, or $x_H \geq x_H^0$, $x_L \geq x_L^0 = 0$. This constraint would only be relevant if the parameter α_H is large relative to the size of possible improved matching $s_H - s_L$, but this case is neither plausible nor interesting. So, if $\underline{x}_H$ is the contribution that solves the IC condition (4), in what follows I assume that $\underline{x}_H > x_H^0$.

[5]And there will be a continuum of pooling equilibria too. I do not discuss pooling equilibria here, but the analysis would be similar to that found below in the section on non-observability.

$G(\bar{X}) - G(X)$ by the concavity of G. Combined with $\alpha_H > 0$, the above inequality clearly holds. So, the interval $[\underline{x}_H, \bar{x}_H]$ is non-empty and so both IC conditions can be satisfied simultaneously.

The non-altruists receive the same matching payoff $\pi(\alpha_L, s_L)$ for any choice of x in $[0, x_H)$ and do not wish to switch to any x in $[\underline{x}_H, \bar{x}_H]$ because of IC condition (4). By assumption the altruists' contributions are higher than x_H^0, the amount chosen in the absence of signalling considerations. Thus, for the non-altruists the marginal return to contribution is even lower and so the result in Proposition 1 is easily adapted to show that non-altruists' optimal choice is still to contribute zero. □

What is important is that in this separating equilibrium, altruists can have a higher material payoff than non-altruists. In such a separating equilibrium, we have material payoffs

$$\Phi_H = -x_H + G(X) + \pi(\alpha_H, s_H) \tag{6}$$

and

$$\Phi_L = -x_L + G(X) + \pi(\alpha_L, s_L). \tag{7}$$

Combining these, the material advantage of the high type is

$$\Phi_H - \Phi_L = \pi(\alpha_H, s_H) - \pi(\alpha_L, s_L) - (x_H - x_L) \tag{8}$$

This could be positive or negative depending on the relative size of $\pi(\alpha_H, s_H) - \pi(\alpha_L, s_L)$ (which is positive) and $x_H - x_L$. What I now show is that even in the separating equilibrium that is worst for altruists, altruists will have a higher material payoff than non-altruists, provided the number of altruists is sufficiently large.

Proposition 3. *Under the assumption that altruists gain a higher return to pair-bonding than non-altruists, if the number of altruists, m, is sufficiently large, then in any separating equilibrium the material payoff to altruists is higher than to non-altruists.*

Proof: If the second IC condition (5) holds with equality, so that we have the separating equilibrium that is worst for altruists, the difference in contributions will be:

$$x_H - x_L = \pi(\alpha_H, s_H) - \pi(\alpha_H, s_L) + (1 + \alpha_H)(G(X) - G(\underline{X})). \tag{9}$$

Then combining (9) with the equation (8), the advantage becomes

$$\Phi_H - \Phi_L = \pi(\alpha_H, s_L) - \pi(\alpha_L, s_L) - (1 + \alpha_H)(G(X) - G(\underline{X})) \tag{10}$$

In the equation (10), the term $A = \pi(\alpha_H, s_H) - \pi(\alpha_L, s_L)$ is a positive constant, while the term $B = -(1 + \alpha_H)(G(X) - G(\underline{X}))$ is negative and for a fixed x_H, by concavity of $G(\cdot)$, is decreasing in m the number of altruists. Further, by assumption $\lim_{x \to \infty} G'(x) = 0$. So if I can show that $X = mx_H$ goes to infinity as m becomes large, then B is less than A in

absolute size, and thus the high type has a material advantage, for m sufficiently large. The problem is that x_H depends on m. But one has

$$G(\bar{X}) - G(X) + \pi(\alpha_L, s_H) - \pi(\alpha_L, s_L) \leq x_H$$

so that x_H is bounded below as $\pi(\alpha_L, s_H) - \pi(\alpha_L, s_L) > 0$ by assumption. Thus, $\lim_{m\to\infty} mx_H = \infty$ and $\lim_{m\to\infty} G(X) - G(\underline{X}) = 0$. □

Thus, if the number of altruists is large, altruists certainly have a material payoff advantage. But note this result does not rule out that altruists will be advantaged even with small numbers. Indeed, altruists will do worse at very low numbers of altruists due to an implausible mechanism. The difference $G(X) - G(\underline{X})$ has to be so big that the level of contribution x_H by altruists is enormous.

However, notice that this result does depend on the assumption that the benefits to the match $\pi(\alpha, s)$ are increasing in the degree of altruism. If not, then it is still possible for altruists to distinguish themselves from non-altruists by signalling. However, in any separating equilibrium, the material payoff of altruists is lower than that of the non-altruists.

Proposition 4. *Under the alternative assumption that $\pi(\alpha_i, s_i) = s_i$, there is no benefit from altruism in pair-bonding, in any separating equilibrium altruists have strictly lower material payoffs than non-altruists.*

Proof: As $\bar{X} > X$, clearly

$$-x_H + G(\bar{X}) + s_H > -x_H + G(X) + s_H$$

Simply combining this with the first IC condition (4), we have that, in the separating equilibrium that is best for altruists, material payoffs must satisfy

$$\Phi_H = -x_H + G(X) + s_H < -x_L + G(X) + s_L = \Phi_L \tag{11}$$

That is, altruists have a lower material payoff. □

This results means that, in the absence of superior mentalizing ability, altruists would become extinct. Note the intuition for this result. The incentive compatibility condition is exactly that the non-altruists do not want to imitate the altruists. The difference between altruist and non-altruists is now only in preferences not in capabilities. Thus, because the preferences of non-altruists are identical to their material payoff, this means that necessarily they must earn a higher material payoff from the lower level of contribution if they prefer it to a higher level.

3 When Contributions are not Observable

In this section, I look at the case where altruists are assumed to be more productive, but where their contributions to the public good is not observed. The question is how this case compares to the signalling outcome of the previous section. The comparison would seem to be ambiguous: when not observed, altruists will have lower costs of contribution, but lower quality matching, as observers will not be able to distinguish altruists. This is, in fact, not the case. Instead, I show that altruists are always better off with signalling.

When not observed, altruists will still contribute more than non-altruists. Specifically, altruists will choose x_H^0 as specified in Proposition 1, the privately optimal contribution for the altruist type, and non-altruists will choose $x_L^0 = 0$. Since observers now cannot distinguish between altruists and non-altruists, both type of contribution obtain in expectation a match of intermediate value s_M where $s_L < s_M < s_H$. So, the material payoff to the altruists will be

$$\Phi_H^N = -x_H^0 + G(X) + \pi(\alpha_H, s_M) \tag{12}$$

and to the non-altruists.

$$\Phi_L^N = G(X) + \pi(\alpha_L, s_M) \tag{13}$$

with the N superscript indicating non-observability.

So the advantage to the altruists under non-observability is the difference,

$$A^N = \Phi_H^N - \Phi_L^N = \pi(\alpha_H, s_M) - \pi(\alpha_L, s_M) - x_H^0. \tag{14}$$

In contrast, the advantage to the altruists under the most advantageous separating equilibrium would be, using (4) and (8),

$$A^S = \Phi_H^S - \Phi_L^S = \pi(\alpha_H, s_H) - \pi(\alpha_L, s_H) - (G(\bar{X}) - G(X)) \tag{15}$$

where S is for separating.

It is easy to show that both A^N and A^S are increasing in m the number of altruists. But importantly, one can also show that the advantage to altruists with signalling is always greater than without observability. This is not obvious as, while with signalling there is more accurate sorting so that altruists match better, with signalling altruists also have to contribute more. What is crucial here is the assumption that $\pi_{\alpha s} > 0$, that is, increasing α increases the return to improving one's match.

Proposition 5. *The advantage to the altruists in a separating equilibrium A_S is greater than the advantage without observability A_N.*

Proof: In comparing A^N in (14) and A^S in (15), let us first consider the returns to the post-match project. Note that $\pi(\alpha_H, s_H) - \pi(\alpha_L, s_H) > \pi(\alpha_H, s_M) - \pi(\alpha_L, s_M)$ as $\pi_{\alpha s} > 0$.

Second, consider the cost of contributions. From (3), one has that $G'(mx_H^0) = 1/(1+\alpha_H) < 1$. Further, the slope of G is decreasing in contributions as G is concave. Thus, $G((m+1)x_H^0) - G(mx_H^0) < x_H^0$. Finally, as by assumption $x_H^0 < x_H$, and again because of the concavity of G, it holds that $G(\bar{X}) - G(X) = G((m+1)x_H) - G(mx_H) < x_H^0$ and the result follows. □

Crucially, what this result shows is possible that without signalling, altruism might not be able to establish itself. For example, it might be the case that $A^S > 0 > A^N$, when the number of altruists is small. If this is the case, then under signalling, altruism would spread within the population, but without signalling it would go extinct. Let us see an example of this.

Example 1. Let $\pi(\alpha, s) = (1+\alpha)s$ and $G(x) = \ln x$, and further $\alpha_H = 1/2$ and s_L, s_M, s_H be 1, 3/2, 2 respectively. Then, $x_H^0 = (1+\alpha_H)/m = 3/2m$ and, thus, $A^S = \Phi_H^S - \Phi_L^S = 1 - \log((m+1)/m) > 3/4 - 3/2m = \Phi_H^N - \Phi_L^N = A^N$. Indeed, in this example, the first altruist would fail to establish herself without observability, as non-altruists have an advantage when there is only one altruist. That is when $m = 1$, $A^N = -3/4 < 0$, whereas with signalling the advantage to the lone altruist is positive, $A^S = 1 - \log 2 > 0$.

4 If Some Non-Altruists Can Mentalize

Let us look at a further possibility: that there exists another type of contributor, who does not have altruistic preferences but is as capable of mentalizing as altruists. Thus, this type would be equally competent in the post-match project. This kind of intelligence without sympathy for others is sometimes called Machiavellian but, more neutrally, let us call this the P-type. We will see that the outcome is vastly different when there is signalling and when there is no observability.

Specifically, the P-type has preferences and fitness

$$U_P = \Phi_P = -x_i + G(\sum_{j=1}^{n} x_j) + \pi(\alpha_H, s_i). \tag{16}$$

That is, he has no altruism as his preferences match his fitness, but he has high productivity π in any match. Without observation, the P-type will choose $x_P^0 = 0$ but gain a product of $\pi(\alpha_H, s_M)$, where s_M is as in the previous section on non-observability. Thus, the fitness of the P-type will be

$$\Phi_P = G(X) + \pi(\alpha_H, s_M) \tag{17}$$

which is clearly greater than the fitness of the altruists Φ_H^N or of the non-altruists Φ_L^N, as defined in (12) and (13) in the previous section. Thus, without observation, the result will be a population consisting entirely of P-types.

In contrast, where observers do view the choice of contribution, the P-type would have a choice between the high contribution of the altruists and the low contribution of the non-altruists (remember that as part of the separating equilibrium, it must be that a choice of some intermediate level of contribution is interpreted as coming from a non-altruist). The high contribution gives a better match and the net fitness is higher than from the low choice, by Proposition 3. Thus, the P-types would choose the high contribution. But note that the P-type now does no better than the altruist. That is,

$$\Phi_P = -x_H + G(X) + \pi(\alpha_H, s_H) = \Phi_H, \tag{18}$$

where Φ_H is the material payoff to the altruist as given in (6).

Furthermore, there is no separating equilibrium where the P-types choose some intermediate level of contribution $\hat{x} \in (0, \underline{x}_H)$ and separate themselves both from the altruists and the low ability non-altruists. This is because, the contribution $\underline{x}_H$ is the minimum level of contribution that is high enough to deter the low types from also choosing to contribute.

This results suggests the following. Suppose mentalizing and empathy did arise together, and those with these joint characteristics started to signal to identify themselves. Then, it might seem that such types would be invadable by a type that could mentalize but avoided the costs of altruism. However, once signalling is in place, such a type would do no better than altruists as it would have to engage in just as much prosocial behavior in order to match well.

5 Conclusions

In this paper, I have shown the following. If having a theory of mind, "mentalizing", is positively associated with empathy, then those possessing these joint attributes can signal this otherwise hidden capability by prosocial behavior. Once such a signalling equilibrium is established, individuals who are able to mentalize but lack altruism would not be able to supplant those with both characteristics. However, in the absence of signalling, altruists would be driven extinct by those who mentalize but who do not empathize.

References

de Waal, Frans B. M. (2008) "Putting the Altruism Back into Altruism; the Evolution of Empathy", *Annual Review of Psychology*, 59, 279-300.

Dekel, Eddie, Jeffrey C. Ely and Okan Yilankaya (2007) "Evolution of Preferences", *Review of Economic Studies*, 74, 685-704.

Dunbar, R.I.M. and Susanne Shultz (2007) "Evolution in the social brain", *Science*, 317, 1344-1347.

Frank, Robert H. (1987) "If *homo economicus* could choose his own utility function, would he want one with a conscience?", *American Economic Review*, 77, 593-604.

Gintis, Herbert, Eric Alden Smith, Samuel Bowles (2001) "Costly Signaling and Cooperation", *Journal of Theoretical Biology*, 213, 103-119.

Griskevicius, V., J. M. Tybur, J. M. Sundie, R. B. Cialdini, G. F. Miller & D. T. Kenrick (2007) "Blatant benevolence and conspicuous consumption: When romantic motives elicit costly displays", *J. Personality and Social Psychology*, 93(1), 85-102.

Güth, W. (1995) "An evolutionary approach to explaining cooperative behavior by reciprocal incentives", *International Journal of Game Theory*, 24, 323-344.

Iredale,W., Van Vugt, M. and Dunbar, R. (2008) "Showing off in humans: Male generosity as a mating signal", *Evolutionary Psychology*, 6(3), 386-392.

Miller, Geoffrey (2000) *The Mating Mind*, London: Heinemann.

Millet, Kobe and Siegfried Dewitte (2007) "Altruistic behavior as a costly signal of general intelligence", *Journal of Research in Personality*, 41, 316-326.

Roberts, Gilbert (1998) "Competitive altruism: from reciprocity to the handicap principle", *Proceedings of the Royal Society B*, 265, 427-431.

Singer, Tania and Ernst Fehr (2005) "The Neuroeconomics of Mind Reading and Empathy", *American Economic Review*, 95, 340-345.

West, Stuart A. and Andy Gardner (2010) "Altruism, Spite and Greenbeards", *Science*, 327: 1341-1344.

Zahavi, A. (1975) "Mate selection- a selection for a handicap", *Journal of Theoretical Biology*, 53, 205-214.