American Economic Journal: Microeconomics 2014, 6(4): 272–292
http://dx.doi.org/10.1257/mic.6.4.272

Competitive Altruism, Mentalizing, and Signaling[†]

By Ed Hopkins*

One explanation of altruism is that it arises from "mentalizing," the process of understanding the mental states of others. Another is based on sexual selection: altruism is a costly signal of good genes. This paper shows that these two arguments are stronger together in that altruists who can mentalize have a greater advantage over nonaltruists when they can signal their type, even though these signals are costly, when such signaling allows better matching opportunities. Finally, it shows how mentalizing leads to higher payoffs for both partners in a long-term relationship, modeled as a repeated game with private monitoring. (*JEL* C73, D64, D82)

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 kin selection 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 pro-social 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 modeled 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 pro-social activity. This has been called "competitive altruism'' (Roberts 1998). Griskevicius et al. (2007) present supportive experimental evidence for the signaling role of pro-social behavior. They find that romantic thoughts can increase willingness in men and women to provide public service (see also Iredale, Van Vugt, and Dunbar 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'' or having a "theory of mind,'' 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).

* School of Economics, University of Edinburgh, Edinburgh EH8 9JT, UK (e-mail: E.Hopkins@ed.ac.uk). I thank Ingela Alger, Marco Faravelli, Tatiana Kornienko, Sevi Mora, József Sákovics, Larry Samuelson, Stuart West, seminar audiences at UBC and UC Irvine and two referees for helpful comments. Errors remain my own.

[†] Go to http://dx.doi.org/10.1257/mic.6.4.272 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

There are problems with both theories. The signaling hypothesis does not explain why quality is signaled by doing good, when it could be equally well signaled by any costly activity (a problem noted by Gintis, Alden Smith, and Bowles 2001 themselves and by Mohr 2011). After all, the leading example of sexual selection is the peacock's tail, where quality is signaled 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.[1]

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 pro-social 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 nonaltruists, it is precisely in giving or contributing to a public good that they have a comparative advantage.[2] 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 pro-social 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 nonaltruists. Nonaltruists' preferences are identical to their actual fitness. Altruists care both about their own fitness but also the fitness of others. That is, similar to the "indirect" evolutionary approach (Frank 1987; Güth 1995), individuals' preferences may differ from their actual fitness. However, in contrast to the indirect approach, here it is assumed that these preferences, and thus a contributor's type, are not observable. Rather, 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, because of their intrinsic preferences, contribute more than

[1] Cole, Mailath, and Postlewaite (1995); Hoppe, Moldovanu, and Sela (2009); and Hopkins (2012) model wasteful signaling and matching.

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

nonaltruists. Precisely because of these preferences, which do not correspond to their true fitness, we would expect altruists to have lower fitness than nonaltruists. However, altruists may gain more favorable matches, if observers 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. It also gives a material reason for observers to prefer to match with altruists. It is further 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. Therefore it is important to specify the mechanism by which mentalizing might give a relevant advantage. Dunbar and Shultz (2007) suggest that long-term social relationships are particularly mentally demanding, and that consequently these relationships have been important in developing human intelligence. I thus conclude by modeling the post-match project as a particular form of long-term relationship—a repeated game with private monitoring based on the recent work by Compte and Postlewaite (2012)—where mentalizing ability gives a material advantage. Further, the quality of the observer with whom one matches is shown to be a strict complement to one's own mentalizing ability.

There are three apparent problems with the approach taken in this paper. First, if altruists have an advantage relative to nonaltruists, for example, in mentalizing, is signaling needed? One might suppose that altruists will supplant nonaltruists simply because they are better. Second, why are altruists not displaced evolutionarily by others that save unnecessary costs by not behaving altruistically? Third, would not some correlation between altruism and some other form of productivity besides mentalizing work just as well?

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

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

Third, it is important that the productive characteristic has a direct physiological link to altruism, as has been proposed for mentalizing, such that it would be difficult to separate them. In particular, the tie in fitness between altruists and nonaltruists just mentioned would turn into a strict disadvantage for nonaltruists if there were any costs associated with separating altruism from mentalizing. However, without these separation costs, the tie could easily go the other way through small costs to altruism.

The approach in this paper is also novel. It builds upon the indirect evolutionary approach that already has been used to explain human cooperation (Frank 1987; Güth 1995). Under the indirect approach as in the current model, 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. Here, however, there is a crucial difference. Recent criticism in economics (Dekel, Ely, and Yilankaya 2007) of this approach has focused on its assumption that agents' preferences are observable by other agents, which seems difficult to defend. 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 that observers will know whether he is an altruist or not. Thus, as with the indirect approach, this paper shows that preferences that are not identical with objective fitness can be evolutionarily stable, but it does so without assuming these preferences are naturally observable to others.[3]

The approach here also relates to the proposed "Greenbeard" mechanism for cooperation: a predisposition for cooperation and a visible external sign (the "green beard") are both encoded in the same gene. However, West and Gardner (2010) doubt whether cooperation based on cooperative or altruistic types being identifiable could be evolutionarily stable in humans, because of the implausibility of altruistic behavior and external signs, such as smiles or promises, being governed by the same small number of genes. Thus, it would be relatively easy for others to develop the external appearance of cooperators without in fact being cooperative. In contrast, here copying is not easy because altruists are identified by costly signals rather than their external appearance. Further, the idea behind the link between altruism and mentalizing is that these two propensities are not separable, the first follows directly from the second. Thus, it would be biologically costly to "rewire" humans to be nonaltruistic while maintaining mentalizing.

I. Signaling 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 of 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 will be $G\left(\sum_{j=1}^{n} x_j\right)$, 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.

[3] Recently, Alger and Weibull (2013) have also proposed a model of indirect evolutionary selection without observability of types. However, they do not allow for signaling and matching between individuals is exogenous.

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

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

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

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

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 nonaltruists' $(\alpha = 0)$ 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.[4]

Importantly, 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. One can assume that each contributor's type is determined by an independent random draw, where the (strictly positive) probability of being an altruist is common knowledge but the total number of altruists is unknown.

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 nonaltruists. Thus, with complete information so that contributors' types were known, altruists would match better than nonaltruists. Specifically, if a contributor's type was directly observable, an altruist would match with a high quality observer with probability s_H and a nonaltruist would have the worse probability $s_L < s_H$. See Section IVA below.

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

[4] It has been suggested that the major motivation for giving to charity or public goods is a "warm glow" rather than altruism, where the subjective utility from donating depends on the donation x_i, not the outcome G. Here, one could assume a warm glow alternative specification, where $U_i = -x_i + \beta x_i + G(\cdot) + \pi(\alpha_i, s_i)$ for some $\beta \in (0, 1)$. All results would be qualitatively unchanged.

contributor type and the observer type. This assumption on the cross-derivative implies the property of increasing differences so that not only does an altruist α_H receive a higher payoff when matching with an observer of quality s_j than a nonaltruist would, but also an increase in match quality has a bigger effect on the return of an altruist than of a nonaltruist. 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 nonaltruists. It will be important in Proposition 5 below. An explicit model that justifies the assumptions on payoffs to contributors and the preferences of observers is given in Section IV.

However, as a useful benchmark, I first look at what contribution agents would choose in the absence of signaling 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 nonaltruists. Let us call a Nash equilibrium, where all altruists make the same choice and all the nonaltruists 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* $n - m$ *nonaltruists choose the same contribution* $x_L^0 = 0$. *There is no other quasi-symmetric Nash equilibrium.*

PROOF:

Suppose that all the nonaltruists 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

$$(3) \qquad (1 + \alpha_H)G'(mx_H^0) = 1,$$

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 nonaltruists 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 signaling, altruists will contribute more than nonaltruists. 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 signaling behavior evolved.

The main results are, first, to show that there exists a separating equilibrium, where altruists choose a different level of contribution than nonaltruists 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.[5]

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

$$\begin{aligned}(4) \qquad U(\alpha_L, x_H, s_H) &= -x_H + G(\overline{X}) + \pi(\alpha_L, s_H) \\ &\leq -x_L + G(X) + \pi(\alpha_L, s_L) = U(\alpha_L, x_L, s_L),\end{aligned}$$

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

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

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

where $\underline{X} = (m - 1)x_H + (n - m + 1)x_L$ or the total contribution if one high type deviates to x_L.[6]

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 signaling model, there will be a continuum of such separating equilibria.[7]

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

[5] To determine the return to a choice of contribution that is neither x_H or x_L, one must specify appropriate 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 nonaltruist. For simplicity, this is what I assume.

[6] There is third incentive compatibility condition that the separating contributions must be at least as large as would be chosen in the absence of signaling 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 $\underline{x}_H > x_H^0$.

[7] 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 nonobservability.

PROOF:

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

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

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 $\overline{X} - X = X - \underline{X} = x_H - x_L$, one has $G(X) - G(\underline{X}) \geq G(\overline{X}) - G(X)$ by the concavity of G. Combined with $\alpha_H > 0$, the above inequality clearly holds. So, the interval $[\underline{x}_H, \overline{x}_H]$ is nonempty and so both IC conditions can be satisfied simultaneously.

The nonaltruists 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 $[\underline{x}_H, \overline{x}_H]$ because of IC condition (4). By assumption, the altruists' contributions are higher than x_H^0, the amount chosen in the absence of signaling considerations. Thus, for the nonaltruists the marginal return to contribution is even lower, and so the result in Proposition 1 is easily adapted to show that nonaltruists' optimal choice is still to contribute zero.

What is important is that in this separating equilibrium, altruists can have a higher material payoff than nonaltruists. 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 nonaltruists, provided the number of altruists is sufficiently large.

PROPOSITION 3: *Under the assumption that altruists gain a higher return to the post-match project than nonaltruists (π is strictly increasing in α), if the number of altruists, m, is sufficiently large, then, in any separating equilibrium, the material payoff to altruists is higher than to nonaltruists.*

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

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

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

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

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

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

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 nonaltruists by signaling. However, in any separating equilibrium, the material payoff of altruists is lower than that of the nonaltruists.

PROPOSITION 4: *Under the alternative assumption that $\pi(\alpha_i, s_i) = s_i$, there is no benefit from altruism in the post-match project, in any separating equilibrium altruists have strictly lower material payoffs than nonaltruists.*

PROOF:

As $\overline{X} > X$, clearly

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

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

$$(11) \qquad \Phi_H = -x_H + G(X) + s_H < -x_L + G(X) + s_L = \Phi_L.$$

That is, altruists have a lower material payoff.

This result 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 nonaltruists do not want to imitate the altruists. The difference between altruist and nonaltruists is now only in preferences not in capabilities. Thus, because the preferences of nonaltruists 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. This also shows the limits of the indirect evolutionary approach. When individuals have to pay to reveal their preferences (and it is a pure preference, that is not correlated with a superior ability), preferences that differ from material payoffs do not survive.

II. When Contributions Are Not Observable

It might be argued that if altruists have an advantage in mentalizing there is no need for them to signal to succeed. To investigate this hypothesis, I now 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 signaling 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 signaling.

When not observed, altruists will still contribute more than nonaltruists. Specifically, altruists will choose x_H^0 as specified in Proposition 1, the privately optimal contribution for the altruist type, and nonaltruists will choose $x_L^0 = 0$. Since observers now cannot distinguish between altruists and nonaltruists, both types of contribution obtain in expectation of 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 nonaltruists,

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

with the N superscript indicating nonobservability.

So the advantage to the altruists under nonobservability 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 signaling is always greater than without observability. This is not obvious as, while with signaling there is more accurate sorting so that altruists match better, with signaling 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 postmatch 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 it is possible that without signaling, 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 signaling, altruism would spread within the population, but without signaling 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, $A^S = \Phi_H^S - \Phi_L^S = 1 - \ln((m + 1)/m) > 3/4 - 3/2m = \Phi_H^N - \Phi_L^N = A^N$. Indeed, in this example, the first altruist would fail to establish herself without observability, as nonaltruists have an advantage when there is only one altruist. That is when $m = 1$, $A^N = -3/4 < 0$, whereas with signaling the advantage to the lone altruist is positive, $A^S = 1 - \ln 2 > 0$.*

III. If Some Nonaltruists Can Mentalize

An important argument in favor of the current approach is that altruism and mentalizing are directly linked. Nonetheless, suppose that it might be possible, perhaps at some cost in fitness, to "rewire" this proposed hard-wired connection. Would altruists survive? To test this, suppose 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 signaling and when there is no observability.

Specifically, the P-type has preferences and fitness

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

That is, he has no altruism as his preferences match his fitness, but he has high productivity π in any match. The cost in fitness of separating altruism from mentalizing is $c \geq 0$. 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 nonobservability. Thus, the fitness of the P-type will be

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

which is clearly greater than the fitness of the altruists Φ_H^N or of the nonaltruists Φ_L^N, as defined in (12) and (13) in the previous section, if c is not too large. 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 nonaltruists (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 nonaltruist). 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 and is likely to do worse. Specifically,

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

where Φ_H is the material payoff to the altruist as given in (6). That is, with a strictly positive rewiring cost c, the P-type will do strictly worse.

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 nonaltruists. 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.

Thus, one can observe the following. Suppose that altruism and mentalizing are hard-wired together such that there would be "rewiring costs" in terms of lost fitness to separate them. Then, when there is no observability, the nonaltruists will dominate if these costs are smaller than the payoff advantage derived above. However, with observability and signaling, the nonaltruists will die out as their payoffs, before rewiring costs, are no better, so their total fitness is lower. Further, this also illustrates why one has to consider two attributes that are strongly linked. For example, suppose altruism was by chance correlated with some form of productivity other than mentalizing, then the signaling analysis of the previous sections would go through. But as the link was coincidental there would be no substantial rewiring costs. Then, a P-type could easily arise that would strictly dominate without observability and would tie under signaling. But it is easy to think of circumstances in which the tie would be broken in favor of the P-types, for example, if only some of public goods games were observed. In summary, both observability/signaling and rewiring costs are important for the survival of altruists.

IV. Matching and the Post-Match Project

In this section, I present a model that would generate payoffs to matching consistent with the assumptions in the main part of the paper. It specifies both the matching process between contributors and observers and gives an example of a post-match project in which mentalizing has an advantage. The crucial assumption is that a post-match project is a repeated pair-based relationship. The most important result shown here is that the payoff obtained in this relationship is increasing in the sophistication of each participant. Thus, both sides would prefer to match with the most sophisticated partner available. Further, the payoff of a contributor will satisfy increasing differences in his type and the quality of the observer with whom he matches.

A. *Matching*

Starting with matching, the simplest assumption is that the n contributors wish to match in pairs with n observers, after the observers have seen their contributions to the public good. If indeed there is a separating outcome in terms of contributions, then observers will be able to deduce precisely the type of each contributor. Suppose the observers also differ in quality or fitness with k having quality v_H and $n - k$ having quality v_L with $v_H > v_L$, but their quality is perfectly observable.[8] Contributors prefer to match with high quality observers and observers prefer to match with contributors with high mentalizing ability (α). Reasons for these preferences are made explicit below.

Suppose $k \leq m$, where m is the number of high type contributors, then a matching will be stable if and only if the k available high-type observers are paired with any k high-type contributors, with the $m - k$ unlucky high-type contributors being matched with v_L observers, as are all the low-type contributors.[9] Further, let us assume that one of these stable matchings is chosen at random. Then the probability of a high-type contributor matching with a high-quality observer is k/m . Low-quality contributors match with certainty with a low-quality observer. So now we can define s in terms of the probability of matching with a high-value observer, so that $s_H = k/m$ and $s_L = 0$.

If instead $k > m$, then some low-type contributors can match with high-quality observers. Thus, in this case, one would have $s_H = 1$ and $s_L = (k - m)/(n - m)$. Note that what is important is simply that $s_H > s_L$ and this will hold in all cases as long as $n > k \geq 1$.

Finally, suppose, as in Section II, contributions are not observable and so all match at random. Then, it is easy to define $s_M = k/n < k/m = s_H$. All these

[8] It would be possible to assume that observers, like contributors, have to signal in order for their type to be known without changing much.

[9] I use stable in the well-known sense of Gale and Shapley (1962). Matching a high-type observer with a low-type contributor is not stable as she could form a blocking pair to this proposed matching by matching instead with any high-quality contributor provisionally matched with a low-quality observer.

results could be easily generalized to the case where the number of observers is not equal to the number of contributors.

B. *Post-Match Project as a Repeated Game*

I now turn from the matching process to the postmatch project. In surveying research on the evolution of the human brain, Dunbar (2008) argues that long-term pair bonding is particularly cognitively demanding. Issues of coordination and monitoring are possible causes of this complexity. He writes that it is this "specific need that may have provided the trigger for the evolution of those social cognitive skills associated with theory of mind in humans" (Dunbar 2008, 18). Thus, it seems reasonable to consider a long-term relationship as the postmatch project, a relationship in which cognitive skills are important.

Specifically, I use the model of Compte and Postlewaite (2012) which analyzes a repeated prisoner's dilemma with private monitoring.[10] The interpretation here is the following. Each day a couple work independently (for example, one hunts, the other gathers) and each may share the food obtained (cooperate) or consume it all without sharing (defect). Since the success of foraging varies, even when no food is shared, an individual does not know for certain whether her partner has defected or was just unlucky. Compte and Postlewaite (CP) assume that each period, both partners each receive a private informative signal about the behavior of the other partner. The novel interpretation proposed here is that the accuracy of the signal is increasing in an individual's ability to mentalize.[11] The interpretation is quite natural in that CP themselves give a psychological interpretation to their model. In the equilibrium they propose, an individual only defects when she is "upset." Thus, an individual who can infer with relative accuracy when her partner is upset can also infer with relative accuracy when her partner defects.

I now outline the model of CP with a simple extension to allow for signal accuracy to differ for the two partners. This permits the analysis of the heterogeneity in mentalizing ability that is the focus of the current study. Two players play an indefinitely repeated game. Each period, each player i chooses an action a_i from the action set $\{C, D\}$. The expected payoffs from the players' action choices in each period are given by

$$\begin{array}{l|c|c} & C & D \\ \hline \text{Cooperate} & 1, 1 & -L, 1+L \\ \hline \text{Defect} & 1+L, -L & 0, 0 \\ \hline \end{array} \tag{19}$$

[10] Private monitoring implies that players cannot observe the actions of other players with certainty, but in each period each player receives a private signal about the other player's action in the previous period. See, for example, Kandori (2002) for an introduction.

[11] Mohlin (2012) and Monte, Robalino, and Robson (2012) have recently proposed quite different models of theory of mind applied to strategic situations.

for some $L > 0$. Actions taken by the other player are not observed, but each period after actions are taken, a player i receives a signal y_i that can be good $(y_i = 1)$ or bad $(y_i = 0)$. The assumption is that

$$p_i = \Pr\{y_i = 0 \mid a_j = D\} = \Pr\{y_i = 1 \mid a_j = C\},$$

with $p_i > \frac{1}{2}$. Further, at the start of each period, players receive a public signal $z \in \{0, 1\}$ with $q = \Pr\{z = 1\}$ with $q \in \left(0, \frac{1}{2}\right)$. The public signal is used to coordinate a return to cooperation after a period of punishment.

CP demonstrate that there can exist an equilibrium, hereafter the "CP equilibrium," of the repeated game of the following form. Each player is in one of two states, *N* or *U*. When she is in *N*, she plays *C* and when in *U*, she plays *D*. She moves from *N* to *U* if and only if $y_i = 0$ and $z = 0$, that is both private and public signals are bad. She then stays in state *U* until the public signal is good, i.e. if $z = 1$, when she returns to *N*. CP show that such an equilibrium exists for a range of parameter values q and L but under the assumption that both players have the same level of accuracy p. In the Appendix, I show that the CP equilibrium still exists if p_1 and p_2 are not too dissimilar and if they are close to one.

Further, it is shown here that, in the CP equilibrium, fitness is increasing in mentalizing ability. There is a potential complication in that mentalizing individuals are assumed to have altruistic preferences. Crucially, however, these altruistic preferences do not change the actions taken or the material payoffs earned in equilibrium. The problem rather is that the CP equilibrium may not exist, as completely altruistic agents would switch to always playing *C*, and thus eroding the advantage that mentalizing would give in the CP equilibrium. It is shown formally in the Appendix that the CP equilibrium continues to exist as long as the level of altruism is not too high.[12] Further, some numeric work (also in the Appendix) in fact indicates that small or intermediate levels of altruism actually make playing this equilibrium easier.

Embedding this in the signaling and matching model considered here, let the accuracy of the signals of the two players be $p_1(\alpha)$ and $p_2(v)$, where v is the observer quality as introduced in the previous subsection on matching. The accuracy of these two signals are assumed to be strictly increasing in α and v, respectively. That is, a high-type contributor has a more accurate signal, just as the signal accuracy of an observer is increasing in her quality.

Just as do CP, I consider the long-run payoffs of the repeated game, that is, those consistent with the probability distribution over the four possible states (*NN*, *NU*, *UN*, and *UU*) generated by equilibrium play. Let ϕ_{ij} be the long-run probability that player 1 is in state i and player 2 is in state j. The expected long-run payoffs to player 1 are

$$(20) \qquad \gamma(p_1, p_2) = \phi_{NN}(p_1, p_2) - L\phi_{NU}(p_1, p_2) + (1 + L)\phi_{UN}(p_1, p_2),$$

[12] Even if small amounts of altruism mean that the advantage through mentalizing is small, it will still be strictly positive. Over many generations even small payoff differences imply big differences in relative frequency.

where the payoffs are derived from (19). These long-run probabilities ϕ_{ij} all depend on p_1, p_2, and q as detailed in the proof to the next result. However, it is relatively easy to calculate that

$$\phi_{NN} = \frac{q}{1 - p_1 p_2(1-q)}. \tag{21}$$

That is the probability that both players cooperate in a given period is increasing in the accuracy p_1, p_2 of their signals. This would justify both observers preferring to match with high-type contributors and contributors preferring high-quality observers.

Furthermore, this probability satisfies $\partial^2 \phi_{NN}/\partial p_1 \partial p_2 > 0$ so that it has increasing differences in p_1 and p_2. The point is that clearly a more accurate signal avoids mistakes where a player incorrectly perceives the other to have defected and, thus, allows for more frequent successful cooperation. But more than that, to stay in the good state *NN*, both players have to simultaneously avoid mistakes. Thus, having a partner whose perception is more accurate increases the value of one's own accuracy.

Finally, let

$$\pi(\alpha, s) = s\gamma(p_1(\alpha), p_2(v_H)) + (1-s)\gamma(p_1(\alpha), p_2(v_L)), \tag{22}$$

where s can take the values $\{s_H, s_M, s_L\}$ as detailed in the subsection above on matching. That is, the overall project payoff $\pi(\alpha, s)$ is defined as the expected payoff taken over both the matching process and play in the repeated game.

PROPOSITION 6: *Given the CP equilibrium, where both players play C in N, play D in U, the payoff $\gamma(p_1(\alpha), p_2(v))$, for $p_1 p_2$ close enough to* 1 *and for* $0 < L < 1/q - 1$, *is increasing in both α and v and, further,* $\gamma_{12} > 0$. *Thus,* $\pi(\alpha, s)$ *as defined as* (22), *is increasing in both arguments and has* $\pi_{\alpha s} > 0$.

PROOF:

The probability of staying in state *NN* is $q + (1-q)p_1 p_2$, whilst the probability of transiting back to *NN* from any other state is q. Thus, this implies that $\phi_{NN} = q + (1-q)p_1 p_2 \phi_{NN}$, which is easily solvable to give (21). It is also possible to calculate

$$\phi_{NU} = \frac{p_1(1-p_2)(1-q)}{q + p_1(1-q)} \phi_{NN}, \quad \phi_{UN} = \frac{(1-p_1)p_2(1-q)}{q + p_2(1-q)} \phi_{NN}.$$

One can check that these probabilities are well-defined and continuous for all values of p_1, p_2, and q in $[0, 1]$. Further, one can calculate directly if tediously from the above and (20) that

$$\left.\frac{\partial \gamma(p_1, p_2)}{\partial p_1}\right|_{p_1 = p_2 = 1} = -2 - L(1-q) + \frac{1}{q} + q;$$

$$\left.\frac{\partial\gamma(p_1,p_2)}{\partial p_2}\right|_{p_1=p_2=1} = L(1-q) + \tfrac{1}{q} - 1 .$$

The first is strictly positive if $L < 1/q - 1$ as assumed. The second is clearly positive for $q < 1$. Further,

$$\left.\frac{\partial^2\gamma(p_1,p_2)}{\partial p_1 \partial p_2}\right|_{p_1=p_2=1} = \gamma_{12}(1,\,1) = \frac{(1-q)^2(2+q^2)}{q^2} > 0.$$

Thus, we have $\gamma_1 > 0$ and $\gamma_2 > 0$ and $\gamma_{12} > 0$ in the neighborhood of $p_1 = p_2 = 1$.

Since $p_1(\alpha)$ is strictly increasing in α, both γ and π, as defined in (22), are increasing in α. Further, we have $\pi_s = \gamma(p_1(\alpha),\, p_2(v_H)) - \gamma(p_1(\alpha),\, p_2(v_L)) > 0$. Finally, $\pi_{\alpha s} = p_1'(\alpha)(\gamma_1(p_1(\alpha),p_2(v_H)) - \gamma_1(p_1(\alpha),p_2(v_L))) > 0$ as $\gamma_{12} > 0$.

Note that this proposed equilibrium payoff structure is robust to deviations at the first stage. Suppose a low-type contributor deviates and chooses x_H. He would be mistaken for a high type and could match appropriately with a high-type observer. But as his signal accuracy in the repeated relationship would be only $p(\alpha_L)$, his payoff in the project would be as required $s_H\gamma(p_1(\alpha_L),p_2(v_H)) + (1-s_H)\gamma(p_1(\alpha_L),p_2(v_L)) = \pi(\alpha_L,s_H) < \pi(\alpha_H,s_H) = s_H\gamma(p_1(\alpha_H),p_2(q_H)) + (1-s_H)\gamma(p_1(\alpha_H),p_2(v_L))$. Thus, the proposed deviation would not be profitable.

V. 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 pro-social behavior. It is shown that mentalizing contributes positively to the return from long-run relationships and, thus, is a desirable attribute in a partner. Successfully signaling one's type therefore increases the quality of partner.

Thus, it has been shown that other-regarding preferences would survive even though these preferences differ from those that would maximize (short-run) fitness. However, one should recognize the limits of this result. Strictly speaking, the model considered here only allows for pro-social behavior in the particular context of demonstrating one's fitness for matching into a long-term relationship. Such behavior may or may not extend into other contexts.[13] It is also important to emphasize that this proposed explanation for cooperative behavior does not exclude other explanations, particularly those based on reciprocity and repeated games. Indeed, one aspect of the current model is the interplay between a short-term interaction in a matching market and a longer term relationship after matching. Finally, sustaining a cooperative equilibrium in repeated relationships is not possible if altruism is too high. This suggests why altruism, although providing a fitness advantage at low but positive levels, does not increase without limit.

[13] How much people discriminate between different contexts in such behavior is unclear. Indeed, one prominent alternative explanation for cooperation in short-run encounters is simply that it represents the mistaken use of reciprocal behavior in an inappropriate context (West, El Mouden, and Gardner 2011).

Mathematical Appendix

In CP, it is shown that an equilibrium for both players to play C in N and D in U, for a nonempty set of the parameter space (L, p, q). In this Appendix, building on this result, I briefly show that the CP equilibrium can still exist when the accuracy probability of private signals p is not equal across the two players.

PROPOSITION 7: *Fix L and q at values such that if $p_2 = p_1 = p$, the CP equilibrium exists for p on the interval $[p_0, 1]$ for some $p_0 < 1$. Then, for p_1 close enough to one, there exists $\underline{p}$ and $\overline{p}$, such that $\overline{p} > p_1 > \underline{p}$ and such that if $p_2 \in [\underline{p}, \overline{p}]$, the CP equilibrium exists.*

PROOF:

The upper bound $\overline{p}$ for p_2 is set by the condition that player 1 should not have an incentive to switch to playing C always, and $\underline{p}$ is fixed by the incentive for player 1 not to deviate to all D . The first condition is

$$\text{(A1)} \qquad \gamma(p_1, p_2) \geq \phi_N^C - L(1 - \phi_N^C) = (1 + L)\frac{q}{q + (1 - q)(1 - p_2)} - L,$$

where ϕ_N^C is the proportion of time player 2 spends in state N when player 1 deviates to all C. Fix p_1 at some value in $(p_0, 1)$. Then, let $\overline{p}$ be the value of p_2 that solves (A1) with equality. The second is

$$\text{(A2)} \qquad \gamma(p_1, p_2) \geq (1 + L)\phi_N^D = (1 + L)\frac{q}{q + (1 - q)p_2},$$

where ϕ_N^D is the proportion of time player 2 spends in state N when player 1 deviates to all D. Similarly define $\underline{p}$ as the value that solves (A2) with equality. It is possible to verify that $\gamma(p_1, p_2)$ as given in (20) is strictly increasing in p_2 . Further, one can verify that, for p_2 close to 1, the right-hand side of (A1) is greater than that of (A2), provided again that $L < 1/q - 1$. Thus, $\overline{p} > \underline{p}$. Suppose $p_1 \notin (\underline{p}, \overline{p})$, then this would contradict the existence of equilibrium, for which the incentive not to defect is strict, for $p_1 = p_2 = p \in (p_0, 1)$, given the continuity of payoffs in p_1 and p_2.

For example, take $L = 1$, $q = 0.3$, then from CP, p_0 is approximately 0.851, so that the symmetric equilibrium exists for $p \in [0.851, 1]$. Keeping these values for L and q, but moving to differing accuracy levels, let the accuracy level of player one be $p_1 = 0.9$. Then, one can calculate that $\underline{p} = 0.852$ and $\overline{p} = 0.967$. That is, an asymmetric equilibrium exists if player 2's accuracy is between these levels.

A. *Altruism in the Post-match Project*

I briefly look at the implications of altruistic preferences for the repeated interaction of the postmatch project. I show that the CP equilibrium exists as long as the altruism parameter is not too large.

Suppose players 1 and 2 have altruism levels α_1 and α_2, respectively. Then subjectively they are playing the following game:

		C	D
(A3)	Cooperate	$1 + \alpha_1, 1 + \alpha_2$	$-L + \alpha_1(1 + L), 1 + L - \alpha_2 L$
	Defect	$1 + L - \alpha_1 L, -L + \alpha_2(1 + L)$	$0, 0$

The subjective payoff in the CP equilibrium, rather than (20), will be

$$\gamma(p_1, p_2; \alpha) = (1 + \alpha)\phi_{NN} + (-L + \alpha(1 + L))\phi_{NU} + (1 + L - \alpha L)\phi_{UN}. \tag{A4}$$

Importantly, this change in subjective payoffs has no effect on the findings of Proposition 6, which considers the effect of play on fitness, provided both partners continue to play the CP equilibrium.

The problem is, as Bernheim and Stark (1988) pointed out, altruism can hinder cooperation in repeated relationships, as it makes individuals unwilling to punish those who have deviated. Here, this effect manifests itself in the question of existence of the CP equilibrium. For sufficiently high levels of altruism, the CP equilibrium ceases to exist as individuals will prefer to switch to all C.[14]

I show that the CP equilibrium still exists if altruism α is less than α^*, defined as

$$\alpha^* = \frac{L(1 - \phi_N^C) + \phi_N - \phi_N^C}{L(1 - \phi_N^C) + 1 - \phi_N}, \tag{A5}$$

which is strictly positive if $L > \underline{L}$, where $\underline{L}$ is the lower bound that CP derive for the existence of their equilibrium ($\underline{L} = (\phi_N^C - \phi_N)/(1 - \phi_N^C) > 0$, where $\phi_N = \phi_{NN} + \phi_{UN}$). That is, if and only if the parameters are such that the CP equilibrium exists, the CP equilibrium still exists under some degree of altruism.

PROPOSITION 8: *Fix L and q at values such that if $p_2 = p_1 = p$, the CP equilibrium exists for p on the interval $[p_0, 1]$, for some $p_0 < 1$. Fix $p \in (p_0, 1)$. Then, for altruistic preferences (A3), the symmetric CP equilibrium still exists if $\alpha < \alpha^*$. Further, suppose that $p_1 \neq p_2$. Then, for p_1 close enough to one, there exists $\underline{p}$ and $\overline{p}$, such that $\overline{p} > p_1 > \underline{p}$ and such that if $p_2 \in [\underline{p}, \overline{p}]$, the CP equilibrium exists.*

PROOF:

Take the incentive compatibility condition (A1), impose the modified payoffs (A3) to obtain,

$$\gamma(p_1, p_2; \alpha) \geq \phi_N^C(1 + \alpha) + (-L + \alpha(1 + L))(1 - \phi_N^C). \tag{A6}$$

[14] This is a problem as if altruists play C instead of the CP equilibrium, fitness in the post-match project will no longer be increasing in mentalizing ability. Thus, a population of altruists could be invaded by individuals who do not mentalize but who, in contrast to the situation in the CP equilibrium, would suffer no fitness disadvantage.

Note that with the CP assumption that $p_1 = p_2 = p$, $\gamma(p_1, p_2, \alpha) = (1 + \alpha)\gamma(p)$. Thus, the above solves under equality to obtain α^* as given in (A5). The other condition (A2) becomes

$$\gamma(p_1, p_2; \alpha) \geq (1 + L - \alpha L)\phi_N^D. \tag{A7}$$

Again assuming $p_1 = p_2$ so that $\gamma(p_1, p_2, ; \alpha) = (1 + \alpha)\gamma(p)$, it can be seen that if the above inequality is satisfied for $\alpha = 0$, then it is clearly satisfied for any $\alpha > 0$.

Now let $p_1 \neq p_2$. One finds $\underline{p}$ from (A6) and $\overline{p}$ from (A7). Again, $\gamma(p_1, p_2; \alpha)$ is increasing in p_2 and one can verify that, for p_2 close to 1, the right-hand side of (A6) is greater than that of (A7), provided, again, that $L < 1/q - 1$. Thus, $\overline{p} > \underline{p}$.

For example, consider $L = 1$, $q = 0.3$, and $p_1 = p_2 = 0.9$, then for altruism above $\alpha^* = 0.26$, individuals earn a higher subjective payoff from playing all C. Nonetheless, it seems that *low* values of altruism aid cooperation relatively to no altruism in that the CP equilibrium exists for a wider part of the parameter space. Again take $L = 1$, $q = 0.3$, and $p_1 = 0.9$. Above it was calculated that $\underline{p} = 0.852$ and $\overline{p} = 0.967$, so that the permissable range for player 2's accuracy is $\overline{p} - \underline{p} = 0.967 - 0.852 = 0.115$. Suppose now $\alpha_2 = 0.2$. Then, $\underline{p} = 0.728$ and $\overline{p} = 0.923$, so that $\overline{p} - \underline{p} = 0.923 - 0.728 = 0.195$. Altruism changes both constraints, making it both less attractive to switch to all D and more attractive to switch to all C. This numeric example indicates that for low levels of altruism, the first effect is stronger.

REFERENCES

Alger, Ingela, and Jörgen W. Weibull. 2013. "Homo Moralis – Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica* 81 (6): 2269–2302.

Bernheim, B. Douglas, and Oded Stark. 1988. "Altruism within the Family Reconsidered: Do Nice Guys Finish Last?" *American Economic Review* 78 (5): 1034–45.

Cole, Harold L., George J. Mailath, and Andrew Postlewaite. 1995. "Incorporating Concern for Relative Wealth into Economic Models." *Federal Reserve Bank of Minneapolis Quarterly Review* 19 (3): 12–21.

Compte, Olivier, and Andrew Postlewaite. 2012. "Plausible Cooperation." Unpublished.

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 (3): 685–704.

Dunbar, R. I. M., and Susanne Shultz. 2007. "Evolution in the Social Brain." *Science* 317 (5843): 1344–47.

Dunbar, Robin I. M. 2008. "Why Humans Aren't Just Great Apes." *Issues in Ethnology and Anthropology* 3 (3): 15–33.

Frank, Robert H. 1987. "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review* 77 (4): 593–604.

Gale D., and L. S. Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69 (1): 9–15.

Gintis, Herbert, Eric Alden Smith, and Samuel Bowles. 2001. "Costly Signaling and Cooperation." *Journal of Theoretical Biology* 213 (1): 103–19.

Griskevicius, V., J. M. Tybur, J. M. Sundie, R. B. Cialdini, G. F. Miller, and D. T. Kenrick. 2007. "Blatant Benevolence and Conspicuous Consumption: When Romantic Motives Elicit Strategic Costly Signals." *Journal of Personality and Social Psychology* 93 (1): 85–102.

Güth, Werner. 1995. "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives." *International Journal of Game Theory* 24 (4): 323–44.

Hopkins, Ed. 2012. "Job Market Signalling of Relative Position or Becker Married to Spence." *Journal of the European Economic Association* 10 (2): 290–322.

Hoppe, Heidrun C., Benny Moldavanu, and Aner Sela. 2009. "The Theory of Assortative Matching Based on Costly Signals." *Review of Economic Studies* 76 (1): 253–81.

Iredale, W., M. Van Vugt, and R. Dunbar. 2008. "Showing off in Humans: Male Generosity as a Mating Signal." *Evolutionary Psychology* 6 (3): 386–92.

Kandori, Michihiro. 2002. "Introduction to Repeated Games with Private Monitoring." *Journal of Economic Theory* 102 (1): 1–15.

Miller, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. London: Heinemann.

Millet, Kobe, and Siegfried Dewitte. 2007. "Altruistic Behavior as a Costly Signal of General Intelligence." *Journal of Research in Personality* 41 (2): 316–26.

Mohlin, Erik. 2012. "Evolution of Theories of Mind." *Games and Economic Behavior* 75 (1): 299–318.

Mohr, Sascha J. 2011. "Generous Behavior: An Unconventional Application of the Theory of Costly Signaling." http://www.researchgate.net/publication/228921306_Generous_Behavior_An_Unconventional_Application_of_the_Theory_of_Costly_Signaling.

Monte, Daniel, Nikolaus Robalino, and Arthur Robson. 2012. "The Evolution of 'Theory of Mind.'" https://econresearch.uchicago.edu/sites/econresearch.uchicago.edu/files/ToM.pdf/.

Roberts, Gilbert. 1998. "Competitive Altruism: From Reciprocity to the Handicap Principle." *Proceedings of the Royal Society B* 265 (1394): 427–31.

Singer, Tania, and Ernst Fehr. 2005. "The Neuroeconomics of Mind Reading and Empathy." *American Economic Review* 95 (2): 340–45.

West, Stuart A., Claire El Mouden, and Andy Gardner. 2011. "Sixteen Common Misconceptions about the Evolution of Cooperation in Humans." *Evolution and Human Behavior* 32 (4): 231–62.

West, Stuart A., and Andy Gardner. 2010. "Altruism, Spite and Greenbeards." *Science* 327 (5971): 1341–44.

Zahavi, A. 1975. "Mate Selection- A Selection for a Handicap." *Journal of Theoretical Biology* 53 (1): 205–14.