arXiv:1410.3363v2 [cs.GT] 4 Nov 2014

Translucent Players: Explaining Cooperative Behavior in Social Dilemmas

Valerio Capraro
Centre for Mathematics and Computer Science (CWI)
Amsterdam, 1098 XG, The Netherlands
V.Capraro@cwi.nl

Joseph Y. Halpern
Cornell University
Computer Science Department
Ithaca, NY 14853
halpern@cs.cornell.edu

November 5, 2014

Abstract

In the last few decades, numerous experiments have shown that humans do not always behave so as to maximize their material payoff. Cooperative behavior when non-cooperation is a dominant strategy (with respect to the material payoffs) is particularly puzzling. Here we propose a novel approach to explain cooperation, assuming what Halpern and Pass (2013) call *translucent players*. Typically, players are assumed to be *opaque*, in the sense that a deviation by one player does not affect the strategies used by other players. But a player may believe that if he switches from one strategy to another, the fact that he chooses to switch may be visible to the other players. For example, if he chooses to defect in Prisoner's Dilemma, the other player may sense his guilt. We show that by assuming translucent players, we can recover many of the regularities observed in human behavior in well-studied games such as Prisoner's Dilemma, Traveler's Dilemma, Bertrand Competition, and the Public Goods game.

1 Introduction

In the last few decades, numerous experiments have shown that humans do not always behave so as to maximize their material payoff. Many alternative models have consequently been proposed to explain deviations from the money-maximization paradigm. Some of them assume that players are boundedly rational and/or make mistakes in the computation of the expected utility of a strategy (Camerer, Ho, and Chong 2004; Costa-Gomes, Crawford, and Broseta 2001; Halpern and Pass 2014; McKelvey and Palfrey 1995; Stahl and Wilson 1994); yet others assume that players have other-regarding preferences (Bolton and Ockenfels 2000; Charness and Rabin 2002; Fehr and Schmidt 1999); others define radically different solution concepts, assuming that players do not try to maximize their payoff, but rather try to minimize their regret (Halpern and Pass 2012; Renou and Schlag 2010), or maximize the forecasts associated to coalition structures (Capraro 2013; Capraro, Venanzi, Polukarov, and Jennings 2013), or maximize the total welfare (Apt and Schäfer 2014; Rong and Halpern 2013). (These references only scratch the surface; a complete bibliography would be longer than this paper!)

Cooperative behaviour in one-shot anonymous games is particularly puzzling, especially in games where non-cooperation is a dominant strategy (with respect to the material payoffs): why should you pay a cost to help a stranger, when no clear direct or indirect reward seems to be at stake? Nevertheless, the secret of success of our societies is largely due to our ability to cooperate. We do not cooperate only with family members, friends, and co-workers. A great deal of cooperation can be observed also in one-shot anonymous interactions (Camerer 2003), where none of the five rules of cooperation proposed by Nowak (2006) seems to be at play.

Here we propose a novel approach to explain cooperation, based on work of Halpern and Pass (2013) and Salcedo (2013), assuming what Halpern and Pass call *translucent players*. Typically, players are assumed to be *opaque*, in the sense that a deviation by one player does not affect the strategies used by other players. But a player may believe that if he switches from one strategy to another, the fact that he chooses to switch may be visible to the other players. For example, if he chooses to defect in Prisoner's Dilemma, the other player may sense his guilt. (Indeed, it is well known that there are facial and bodily clues, such as increased pupil size, associated with deception; see, e.g., (Ekman and Friesen 1969). Professional poker players are also very sensitive to *tells*—betting patterns and physical demeanor that reveal something about a player's hand and strategy.)

We use the idea of translucency to explain cooperation. This may at first seem somewhat strange. Typical lab experiments of social dilemmas consider anonymous players, who play each other over computers. In this setting, there are no tells. However, as Rand and his colleagues have argued (see, e.g., (Rand et al. 2012; Rand et al. 2014)), behavior of subjects in lab experiments is strongly influenced by their experience in everyday interactions. People internalize strategies that are more successful in everyday interactions and use them as default strategies in the lab. We would argue that people do not just internalize strategies; they also internalize *beliefs*. In everyday interactions, changing strategies certainly affects how other players react in the future. Through tells and leaks, it also may affect how other players react in current play. Thus, we would argue that in everyday interactions, people assume a certain amount of translucency, both because it is a way of taking the future into account in real-world situations that are repeated and because it is a realistic assumption in one-shot games that are played in settings where players have a great deal of social interaction. We claim that players then apply these beliefs in lab settings where they are arguably inappropriate.

There is some additional experimental evidence that can be viewed as supporting translucency. There is growing evidence that showing people simple images of watching eyes has a marked effect on behavior, ranging from giving more in Public Goods games to littering less (see (Bateson et al. 2013) for a discussion of some of this work and an extensive list of references). One way of understanding these results is that the eyes are making people feel more translucent.

We apply the idea of translucency to a particular class of games that we call *social dilemmas* (cf. (Dawes 1980)). A social dilemma is a normal-form game with two properties:

1. there is a unique Nash equilibrium s^N, which is a pure strategy profile;

2. there is a unique welfare-maximizing profile s^W, again a pure strategy profile, such that each player's utility if s^W is played is higher than his utility if s^N is played.

Although social dilemmas are clearly a restricted class of games, they contain some of the best-studied games in the game theory literature, including Prisoner's Dilemma, Traveler's Dilemma (Basu 1994), Bertrand Competition, and the Public Goods game. (See Section 3 for more discussion of these games.)

There are (at least) two reasons why an agent may be concerned about translucency in a social dilemma: (1) his opponents may discover that he is planning to defect and punish him by defecting as well, (2) many other people in his social group (which may or may not include his opponent) may discover that he is planning to defect (or has defected, despite the fact that the game is anonymous) and think worse of him.

For definiteness, we focus here on the first point and assume that, in social dilemmas, players have a degree of belief α that they are translucent, so that if they intend to cooperate (by playing their component of the welfare-maximizing strategy) and decide to deviate, there is a probability α that another player will detect this, and play her component of the Nash equilibrium strategy. (These detections are independent, so that the probability of, for example, exactly two players other than i detecting a deviation by i is $\alpha^2(1-\alpha)^{N-3}$, where N is the total number of players.) Of course, if $\alpha = 0$, then we are back at the standard game-theoretic framework. We show that, with this assumption, we can already explain a number of experimental regularities observed in social dilemmas (see Section 3). We can model the second point regarding concerns about translucency in much the same way, and would get qualitatively similar results (see Section 5).

The rest of the paper is as follows. In Section 2, we formalize the notion of translucency in a game-theoretic setting. In Section 3, we define the social dilemmas that we focus on in this paper; in Section 4, we show that by assuming translucency, we can obtain as predictions of the framework a number of regularities that have been observed in the experimental literature. In Section 5, we show that most of the other approaches proposed for explaining human behavior in social dilemmas do not predict all these regularities.

In the appendix, we discuss a solution concept that we call *translucent equilibrium*, based on translucency, closely related to the notion of *individual rationality* discussed by Halpern and Pass (2013), and show how it can be applied in social dilemmas.

2 Rationality with translucent players

In this section, we briefly define rationality in the presence of translucency, motivated by the ideas in Halpern and Pass (2013).

Formally, a (finite) normal-form game $\mathcal{G}$ is a tuple $(P, S_1, \ldots, S_N, u_1, \ldots, u_N)$, where $P = \{1, \ldots, N\}$ is the set of players, S_i is the set of strategies for player i, and u_i is player i's utility function. Let $S = S_1 \times \cdots \times S_N$ and $S_{-i} = \prod_{j \neq i} S_j$. We assume that S is finite and that $N \geq 2$.

In standard game theory, it is assumed that a player i has beliefs about the strategies being used by other players; i is rational if his strategy is a best response to these beliefs. The standard definition of best response is the following.

Definition 2.1. *A strategy $s_i \in S_i$ is a best response to a probability μ_i on S_{-i} if, for all strategies s_i' for player i, we have*

$$\sum_{s'_{-i} \in S_{-i}} \mu_i(s'_{-i}) u_i(s_i, s'_{-i}) \geq \sum_{s'_{-i} \in S_{-i}} \mu_i(s'_{-i}) u_i(s'_i, s'_{-i}).$$

Definition 2.1 implicitly assumes that i's beliefs about what other agents are doing do not change if i switches from s_i, the strategy he was intending to play, to a different strategy. (In general, we assume that i always has an *intended strategy*, for otherwise it does not make sense to talk about i switching to a different strategy.) So what we really have are beliefs $\mu_i^{s_i, s_i'}$ for i indexed by a pair of strategies s_i and

s_i'; we interpret $\mu_i^{s_i,s_i'}$ as i's beliefs if he intends to play s_i but instead deviates to s_i'. Thus, $\mu_i^{s_i,s_i}$ represents i's beliefs if he plays s_i and does not deviate.

We can now define a best response for i with respect to a family of beliefs $\mu_i^{s_i,s_i'}$.

Definition 2.2. *Strategy $s_i \in S_i$ is a* best response for i with respect to the beliefs $\{\mu_i^{s_i,s_i'} : s_i' \in S_i\}$ *if, for all strategies $s_i' \in S_i$, we have*

$$\sum_{s'_{-i} \in S_{-i}} \mu_i^{s_i,s_i}(s'_{-i}) u_i(s_i, s'_{-i}) \geq \sum_{s'_{-i} \in S_{-i}} \mu_i^{s_i,s_i'}(s'_{-i}) u_i(s_i', s'_{-i}).$$

We are interested in players who are making best responses to their beliefs, but we define best response in terms of Definition 2.2, not Definition 2.1. Of course, the standard notion of best response is just the special case of the notion above where $\mu_i^{s_i,s_i'} = \mu_i^{s_i,s_i}$ for all s_i': a player's beliefs about what other players are doing does not change if he switches strategies.

Definition 2.3. *We say that a player is* translucently rational *if he best responds to his beliefs in the sense of Definition 2.2.*

Our assumptions about translucency will be used to determine $\mu_i^{s_i,s_i'}$. For example, suppose that Γ is a 2-player game, player 1 believes that, if he were to switch from s_i to s_i', this would be detected by player 2 with probability α, and if player 2 did detect the switch, then player 2 would switch to s_j'. Then $\mu_i^{s_i,s_i'}$ is $(1-\alpha)\mu^{s_i,s_i} + \alpha\mu'$, where μ' assigns probability 1 to s_j'; that is, player 1 believes that with probability $1-\alpha$, player 2 continues to do what he would have done all along (as described by μ^{s_i,s_i}) and, with probability α, player 2 switches to s_j'.

3 Social dilemmas

Social dilemmas are situations in which there is a tension between the collective interest and individual interests: every individual has an incentive to deviate from the common good and act selfishly, but if everyone deviates, then they are all worse off. Personal and professional relationships, depletion of natural resources, climate protection, security of energy supply, and price competition in markets are all instances of social dilemmas.

As we said in the introduction, we formally define a social dilemma as a normal-form game with a unique Nash equilibrium and a unique welfare-maximizing profile, both pure strategy profiles, such that each player's utility if s^W is played is higher than his utility if s^N is played. While this is a quite restricted set of games, it includes many that have been quite well studied. Here, we focus on the following games:

Prisoner's Dilemma. Two players can either cooperate (C) or defect (D). To relate our results to experimental results on Prisoner's Dilemma, we think of cooperation as meaning that a player pays a cost $c > 0$ to give a benefit $b > c$ to the other player. If a player defects, he pays nothing and gives nothing. Thus, the payoff of (D, D) is $(0, 0)$, the payoff of (C, C) is $(b - c, b - c)$, and the payoffs of (D, C) and (C, D) are $(b, -c)$ and $(-c, b)$, respectively. Condition $b > c$ implies that (D, D) is the unique Nash equilibrium and (C, C) is the unique welfare-maximizing profile.

Public Goods game. $N \geq 2$ contributors are endowed with 1 dollar each; they must simultaneously decide how much, if anything, to contribute to a public pool. (The contributions must be in whole cent amounts.) The total amount in the pot is then multiplied by a constant strictly between 1 and N, and then evenly redistributed among all players. So the payoff of player i is $u_i(x_1, \ldots, x_N) = 1 - x_i + \rho(x_1 + \ldots + x_N)$, where x_i denotes i's contribution, and $\rho \in \left(\frac{1}{N}, 1\right)$ is the *marginal return*. (Thus, the pool is multiplied by ρN before being split evenly among all players.) Everyone contributing nothing to the pool is the unique Nash equilibrium, and everyone contributing their whole endowment to the pool is the unique welfare-maximizing profile.

Bertrand Competition. $N \geq 2$ firms compete to sell their identical product at a price between the "price floor" $L \geq 2$ and the "reservation value" H. (Again, we assume that H and L are integers, and all prices must be integers.) The firm that chooses the lowest price, say s, sells the product at that price, getting a payoff of s, while all other firms get a payoff of 0. If there are ties, then the sales are split equally among all firms that choose the lowest price. Now everyone choosing L is the unique Nash equilibrium, and everyone choosing H is the unique welfare-maximizing profile.[1]

Traveler's Dilemma. Two travelers have identical luggage, which is damaged (in an identical way) by an airline. The airline offers to recompense them for their luggage. They may ask for any dollar amount between L and H (where L and H are both positive integers). There is only one catch. If they ask for the same amount, then that is what they will both receive. However, if they

[1] We require that $L \geq 2$ for otherwise we would not have a unique Nash equilibrium, a condition we imposed on Social Dilemmas. If $L = 1$ and $N = 2$, we get two Nash equilibria: $(2, 2)$ and $(1, 1)$; similarly, for $L = 0$, we also get multiple Nash equilibria, for all values of $N \geq 2$.

ask for different amounts—say one asks for m and the other for m', with $m < m'$—then whoever asks for m (the lower amount) will get $m + b$ (m and a bonus of b), while the other player gets $m - b$: the lower amount and a penalty of b. It is easy to see that (L, L) is the unique Nash equilibrium, while (H, H) maximizes social welfare, independent of b.

From here on, we say that a player *cooperates* if he plays his part of the socially-welfare maximizing strategy profile and *defects* if he plays his part of the Nash equilibrium strategy profile.

While Nash equilibrium predicts that people should always defect in social dilemmas, in practice, we see a great deal of cooperative behavior; that is, people often play (their part of) the welfare-maximizing profile rather than (their part of) the Nash equilibrium profile. Of course, there have been many attempts to explain this. Evolutionary theories may explain cooperative behavior among genetically related individuals (Hamilton 1964) or when future interactions among the same subjects are likely (Nowak and Sigmund 1998; Trivers 1971); see (Nowak 2006) for a review of the five rules of cooperation. However, we often observe cooperation even in one-shot anonymous experiments among unrelated players (Rapoport 1965).

Although we do see a great deal of cooperation in these games, we do not always see it. Here are some of the regularities that have been observed:

- The degree of cooperation in the Prisoner's dilemma depends positively on the benefit of mutual cooperation and negatively on the cost of cooperation (Capraro, Jordan, and Rand 2014; Engel and Zhurakhovska 2012; Rapoport 1965).

- The degree of cooperation in the Traveler's Dilemma depends negatively on the bonus/penalty (Capra, Goeree, Gomez, and Holt 1999).

- The degree of cooperation in the Public Goods game depends positively on the constant marginal return (Gunnthorsdottir, Houser, and McCabe 2007; Isaac, Walker, and Thomas 1984).

- The degree of cooperation in the Public Goods game depends positively on the number of players (Barcelo and Capraro 2014; Isaac, Walker, and Williams 1994; Zelmer 2003).

- The degree of cooperation in the Bertrand Competition depends negatively on the number of players (Dufwenberg and Gneezy 2002).

- The degree of cooperation in the Bertrand Competition depends negatively on the price floor (Dufwenberg, Gneezy, Goeree, and Nagel 2007).

4 Explaining social dilemmas using translucency

As we suggested in the introduction, we hope to use translucency to explain cooperation in social dilemmas. To do this, we have to make assumptions about an agent's beliefs. Say that an agent i has *type* (α, β, C) if i intends to cooperate (the parameter C stands for *cooperate*) and believes that (a) if he deviates from that, then each other agent will independently realize this with probability α; (b) if an agent j realizes that i is not going to cooperate, then j will defect; and (c) all other players will either cooperate or defect, and they will cooperate with probability β.

The standard assumption, of course, is that $\alpha = 0$. Our results are only of interest if $\alpha > 0$. The assumption that i believes that agent j will defect if she realizes that i is going to deviate from cooperation seems reasonable; defection is the "safe" strategy. We stress that, for our results, it does not matter what j actually does. All that matters are i's beliefs about what j will do. The assumption that players will either cooperate or defect is trivially true in Prisoner's Dilemma, but is a highly nontrivial assumption in the other games we consider. While cooperation and defection are arguably the most salient strategies, we do in practice see players using other strategies. For instance, the distribution of strategies in the Public Goods game is typically tri-modal, concentrated on contributing nothing, contributing everything, and contributing half (Capraro, Jordan, and Rand 2014). We made this assumption mainly for technical convenience; it makes the calculations much easier. We believe that results qualitatively similar to ours will hold under a much weaker assumption, namely, that a type (α, β, C) player believes that other players will cooperate with probability β (without assuming that they will defect with probability $1 - \beta$).

Similarly, the assumptions that a social dilemma has a unique Nash equilibrium and a unique social-welfare maximizing strategy were made largely for technical reasons. We can drop these assumptions, although that would require more complicated assumptions about players' beliefs.

The key feature of our current assumptions is that the type of player i determines the distributions $\mu_i^{s_i,s_i'}$. In a social dilemma with N agents, the distribution $\mu_i^{s_i,s_i}$ assigns probability $\beta^r(1-\beta)^{N-1-r}$ to a strategy profile s_{-i} for the players other than i if exactly r players cooperate in s_{-i} and the remaining $N-1-r$ players defect; it assigns probability 0 to all other strategy profiles. The distributions $\mu_i^{s_i,s_i'}$ for $s_i' \neq s_i$ all have the form $\sum_{J\subseteq\{1,\ldots,i-1,i+1,\ldots,N\}} \alpha^{|J|}(1-\alpha)^{N-1-|J|}\mu_i^J$, where μ_i^J is the distribution that assigns probability $\beta^k(1-\beta)^{N-|J|-k}$ to a profile where $k \leq N-1-|J|$ players not in J cooperate, and the remaining players (which includes all the players in J) defect. Thus, μ_i^J is the distribution that describes what player i's beliefs would be if he knew that exactly the players in J

had noticed his deviation (which happens with probability $\alpha^{|J|}(1-\alpha)^{N-1-|J|}$). In the remainder of this section, when we talk about best response, it is with respect to these beliefs.

For our purposes, it does not matter where the beliefs α and β that make up a player's type come from. We do not assume, for example, that other players are (translucently) rational. For example, i may believe that some players cooperate because they are altruistic, while others may cooperate because they have mistaken beliefs. We can think of β as summarizing i's previous experience of cooperation when playing social dilemmas. Here we are interested in the impact of the parameters of the game on the reasonableness of cooperation, given a player's type.

The following four propositions analyze the four social dilemmas in turn. We start with Prisoners Dilemma. Recall that b is the benefit of cooperation and c is its cost.

Proposition 4.1. *In Prisoner's Dilemma, it is translucently rational for a player of type (α, β, C) to cooperate if and only if $\alpha\beta b \geq c$.*

Proof. If player i has type (α, β, C) and cooperates in Prisoner's Dilemma, then his expected payoff is $\beta(b-c)-(1-\beta)c$, since player i believes that $j \neq i$ will cooperate with probability β. However, if i deviates from his intended strategy of cooperation, then j will catch him with probability α and also defect. Thus, if i deviates, then i's belief that j will cooperate goes down from β to $(1-\alpha)\beta$. (We remark that this is the case in all social dilemmas; this fact will be used in all our arguments.) This means that i's expected payoff if he deviates by defecting is $(1-\alpha)\beta b$. So cooperating is a best response if $\beta(b-c)-(1-\beta)c \geq (1-\alpha)\beta b$. A little algebra shows that this reduces to $\alpha\beta b \geq c$. □

As we would expect, if $\alpha = 0$, then cooperation is not a best response in Prisoner's Dilemma; this is just the standard argument that defection dominates cooperation. But if $\alpha > 0$, then cooperation can be rational. Moreover, if we fix α, the greater the benefit of cooperation and the smaller the cost, then the smaller the value of β that still allows cooperation to be a best response.

We next consider Traveler's Dilemma. Recall that b is the reward/punishment, and H and L are the high and low payoffs, respectively.

Proposition 4.2. *In Traveler's Dilemma, it is translucently rational for a player of (α, β, C) to cooperate if and only if* $b \leq \begin{cases} \frac{(H-L)\beta}{1-\alpha\beta} & \textit{if } \alpha \geq \frac{1}{2} \\ \min\left(\frac{(H-L)\beta}{1-\alpha\beta}, \frac{H-L-1}{1-2\alpha}\right) & \textit{if } \alpha < \frac{1}{2}; \end{cases}$

Proof. If player i has type (α, β, C) and cooperates in Traveler's Dilemma, then his expected payoff is $\beta H + (1-\beta)(L-b)$, since player i believes that $j \neq i$

will cooperate with probability β. If i deviates and plays $x \neq H$, then j will catch him with probability α and play L. Recall from the proof of Proposition 4.1 that, if i deviates, i's belief that j cooperates is $(1-\alpha)\beta$. This means that i's expected payoff if he deviates to $x < H$ is $(1-\alpha)\beta(x+b)+(1-\beta+\alpha\beta)(L-b)$ if $x > L$, and $(1-\alpha)\beta(L+b)+(1-\beta+\alpha\beta)L = L+(\beta-\alpha\beta)b$ if $x = L$. It is easy to see that i maximizes his expected payoff either if $x = H-1$ or $x = L$. Thus, cooperation is a best response if $\beta H+(1-\beta)(L-b) \geq \max((1-\alpha)\beta(H+b-1)+(1-\beta+\alpha\beta)(L-b), L+(\beta-\alpha\beta)b)$. Again, straightforward algebra shows that this condition is equivalent to the one stated, as desired. (It is easy to check that if $\alpha \geq 1/2$, then the condition $\beta H+(1-\beta)(L-b) \geq (1-\alpha)\beta(H+b-1)+(1-\beta+\alpha\beta)(L-b)$ is guaranteed to hold, which is why we get the two cases depending on whether $\alpha \geq 1/2$.) □

Proposition 4.2 shows that as b, the punishment/reward, increases, a player must have greater belief that his opponent is cooperative and/or a greater belief that the opponent will learn about his deviation and/or a greater difference between the high and low payoffs in order to make cooperation a best response. (The fact that increasing β increases $\frac{(H-L)\beta}{1-\alpha\beta}$ follows from straightforward calculus.)

We next consider the Public Goods game. Recall the ρ is the marginal return of cooperating.

Proposition 4.3. *In the Public Goods game with N players, it is translucently rational for a player of type (α, β, C) to cooperate if and only if $\alpha\beta\rho(N-1) \geq 1-\rho$.*

Proof. Suppose player i, of type (α, β, C), cooperates. Since he expects a player to cooperate with probability β, the expected number of cooperators among the other players is $\beta(N-1)$. Since he himself will cooperate, the total expected number of cooperators is $1+\beta(N-1)$. Since i's payoff is ρm if m players including him cooperate, and thus is linear in the number of cooperators, his expected payoff is exactly his payoff if the expected number of players cooperate. Since his expected payoff with $1+\beta(N-1)$ cooperators is $\rho(1+\beta(N-1))$, this is his expected payoff if he cooperates.

On the other hand, if i deviates by contributing $x < 1$, his expected payoff if m other players cooperate is $(1-x)+\rho(m+x)$. Again, if i deviates, his expected belief that j will cooperate is $(1-\alpha)\beta$. Thus, the expected number of cooperators is $(1-\alpha)\beta(N-1)$, and his expected payoff is $1-x+\rho((1-\alpha)\beta(N-1)+x)$. Since $\rho < 1$, he gets the highest expected payoff by defecting (i.e., taking $x = 0$).

Thus, cooperation is a best response if $\rho(1+\beta(N-1)) \geq 1+\rho(1-\alpha)\beta(N-1)$. Simple algebra shows that this condition holds iff $\alpha\beta\rho(N-1) \geq 1-\rho$. □

Proposition 4.3 shows that if $\rho = 1$, then cooperation is certainly a best response (you always get out at least as much as you contribute). For fixed α and β, there is guaranteed to be an N_0 such that cooperation is a best response for all $N \geq N_0$; moreover, for fixed α, as N gets larger, smaller and smaller βs are needed for cooperation to be a best response.

Finally we consider the Bertrand competition. Recall that H is the reservation value and L is the price floor.

Proposition 4.4. *In Bertrand Competition, it is translucently rational for a player of type* (α, β, C) *to cooperate iff* $\beta^{N-1} \geq f(\gamma, N)LN/H$*, where* $f(\gamma, N) = \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k \gamma^{N-k-1}/(k+1)$ *and* $\gamma = (1-\alpha)\beta$.

Proof. Clearly, if player i cooperates, then his expected payoff is $\beta^{N-1}H/N$, since he gets H/N if everyone else cooperates (which happens with probability β^{N-1}), and otherwise gets 0.

Let $\gamma = (1-\alpha)\beta$. Again, this is the probability that i ascribes to another player playing H if he deviates. If i deviates, then it is easy to see (given his beliefs) that the optimal choices for deviation are $H - 1$ and L. In the former case,

i's expected payoff is $\gamma^{N-1}(H-1)$. In the latter case, i's expected payoff is $\sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k \gamma^{N-k-1} L/(k+1)$: with probability $(1-\gamma)^k \gamma^{N-k-1}$, exactly k other players will play L, and i's payoff will be $L/(k+1)$. Moreover, each possible subset of k defectors, has to be count $\binom{N-1}{k}$ times. Let $f(\gamma, N) = \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k \gamma^{N-k-1}/(k+1)$. Note that, as the notation suggests, this expression depends only on γ and N (and not any of the other parameters of the game). Thus, i's expected payoff in this case is $f(\gamma, N)L$, so cooperation is a best response iff $\beta^{N-1}H/N \geq \max(\gamma^{N-1}(H-1), f(\gamma, N)L)$. While it seems difficult to find a closed-form expression for $f(\gamma, N)$, this does not matter for our purposes.[2] Since we clearly have $\beta^{N-1}H/N \geq \gamma^{N-1}(H-1)$, cooperation is a best response iff $\beta^{N-1}H/N \geq f(\gamma, N)L$, or, equivalently, $\beta^{N-1} \geq f(\gamma, N)LN/H$, □

Note that $f(\gamma, N) = \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k \gamma^{N-k-1}/(k+1) \geq \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k \gamma^{N-k}/N = 1/N$, so Proposition 4.4 shows cooperation is irrational if $\beta^{N-1} < L/H$. Thus, while cooperation may be achieved for reasonable values of α and β if N is small, a player must be more and more certain of cooperation in order to cooperate in Bertrand Competition as the number of players increases. Indeed, for a fixed type (α, β, C), there exists N_0 such that cooperation is not a best response for all $N \geq N_0$. Moreover, if we fix the number N of players, more values of α

[2] Note that the expected value of $L/(k+1)$ cannot be computed by plugging in the expected value of k, in the spirit of our earlier calculations, since $L/(k+1)$ is not linear in k.

and β allow cooperation as L/H gets smaller. In particular, if we fix H and raise the floor L, fewer values of α and β allow cooperation.

While Propositions 4.1–4.4 are suggestive, we need to make extra assumptions to use these propositions to make predictions. A simple assumption that suffices is that there are a substantial number of translucently rational players whose types have the form (α, β, C), and for each pair (u, v) and (u', v') of open intervals in $[0, 1]$, there is a positive probability of finding someone of type (α, β, C) with $\alpha \in (u, v)$ and $\beta \in (u', v')$. With this assumption, it is easy to see that all the regularities discussed in Section 3 hold.

5 Discussion

We have presented an approach that explains a number of well-known observations regarding the extent of cooperation in social dilemmas. In addition, our approach can also be applied to explain the apparent contradiction that people cooperate more in a one-shot Prisoner's dilemma when they do not know the other player's choice than when they do. In the latter case, Shafir and Tversky (1992) found that most people (90%) defect, while in the former case, only 63% of people defect. Our model of translucent players predicts this behavior: if player 1 knows player 2 choices, then there is no translucency and thus our model predicts that player 1 defects for sure. On the other hand, if player 1 does not know player 2's choice and believes that he is to some extent translucent, then, as shown in Proposition 4.1, he may be willing to cooperate. Seen in this light, our model can also be interpreted as an attempt to formalize *quasi-magical thinking* (Shafir and Tversky 1992), the kind of reasoning that is supposed to motivate those people who believe that the others' reasoning is somehow influenced by their own thinking, even though they know that there is no causal relation between the two. Quasi-magical thinking has also been formalized by Masel (2007) in the context of the Public Goods gam and by Daley and Sadowski (2014) in the context of symmetric 2×2 games. The notion of translucency goes beyond these models, since it may applied to a much larger set of games.

Besides a retrospective explanation, our model makes new predictions for social dilemmas which, to the best of our knowledge, have never been tested in the lab. In particular, it predicts that

- the degree of cooperation in Traveler's dilemma increases as the difference $H - L$ increases;
- for fixed L and N, the degree of cooperation in Bertrand Competition increases as H increases, and what really matters is the ratio L/H.

Clearly much more experimental work needs to be done to validate the approach. For one thing, it is important to understand the predictions it makes for other social dilemmas and for games that are not social dilemmas. Perhaps even more important would be to see if we can experimentally verify that people believe that they are to some extent translucent, and, if so, to get a sense of what the value of α is. In light of the work on watching eyes mentioned in the introduction, it would also be interesting to know what could be done to manipulate the value of α.

As we mentioned, there have been many other attempts to explain cooperation in social dilemmas, especially recently. Most of other approaches that we are aware of are not able to obtain all the regularities that we have mentioned.

- The Fehr and Schmidt (1999) inequity-aversion model assumes that subjects play a Nash equilibrium of a modified game, in which players do not only care about their monetary payoff, but also they care about equity. Specifically, player i's utility when strategy s is played is assumed to be $U_i(s) = u_i(s) - \frac{a_i^{FS}}{N-1}\sum_{j\neq i}\max(u_j(s)-u_i(s),0) - \frac{b_i^{FS}}{N-1}\sum_{j\neq i}\max(u_i(s)-u_j(s),0)$, where $u_i(s)$ is the material payoff of player i, and $0 \leq b_i^{FS} \leq a_i^{FS}$ are individual parameters, where a_i^{FS} represents the extent to which player i is averse to inequity in favor of others, and b_i^{FS} represents his aversion to inequity in his favor. Consider the Public Goods game with N players. The strategy profile $(x,\ldots,x)$, where all players contribute x gives player i a utility of $(1-x)+\rho Nx$. If $x > 0$ and player i contributes $x' < x$, then his payoff is $(1-x')+\rho((N-1)x+x') - b_i^{FS}\rho(x-x')$. Thus, $(x,\ldots,x)$ is an equilibrium if $b_i^{FS}\rho(x-x') \geq (1-\rho)(x-x')$, that is, if $b_i^{FS} \geq (1-\rho)/\rho$. Thus, if $b_i^{FS} \geq (1-\rho)/\rho$ for all players i, then $(x,\ldots,x)$ is an equilibrium for all choices of x and all values of N. While there may be other pure and mixed strategy equilibria, it is not hard to show that if $b_i^{FS} < (1-\rho)/\rho$, then player i will play 0 in every equilibrium (i.e., not contribute anything). As a consequence, assuming, as in our model, that players believe that there is a probability β that other agents will cooperate and that the other agents either cooperate or defect, Fehr and Schmidt (1999) model does not make any clear prediction of a group-size effect on cooperation in the public goods game.

- McKelvey and Palfrey's (1995) *quantal response equilibrium (QRE)* is defined as follows.[3] Taking $\sigma_i(s)$ to be the probability that mixed strategy σ_i assigns to the pure strategy s, given $\lambda > 0$, a mixed strategy profile σ is a

[3] We actually define here a particular instance of QRE called the *logit QRE*; λ is a free parameter of this model.

QRE if, for each player i, $\sigma_i(s) = \frac{e^{\lambda EU_i(s,\sigma_{-i})}}{\sum_{s'_i \in S_i} e^{\lambda EU_i(s'_i,\sigma_{-i})}}$.

To see that QRE does not describe human behaviour well in social dilemmas, observe that in the Prisoner's Dilemma, for all choices of parameters b and c in the game, all choices of the parameter λ, all players i, and all (mixed) strategies s_{-i} of player $-i$, we have $EU_i(C, s_{-i}) < EU_i(D, s_{-i})$. Consequently, whatever the QRE σ is, we must have $\sigma_i(C) < \frac{1}{2} < \sigma_i(D)$, that is, QRE predicts that the degree of cooperation can never be larger than 50%. However, experiments show that we can increase the benefit-to-cost ratio so as to reach arbitrarily large degrees of cooperation (close to 80% in (Capraro, Jordan, and Rand 2014) with $b/c = 10$).

- *Iterated regret minimization* (Halpern and Pass 2012) does not make appropriate predictions in Prisoner's Dilemma and the Public Goods game, because it predicts that if there is a dominant strategy then it will be played, and in these two games, playing the Nash equilibrium is the unique dominant strategy.

- Capraro's (2013) notion of *cooperative equilibrium*, while correctly predicting the effects of the size of the group on cooperation in the Bertrand Competition and the Public Goods game (Barcelo and Capraro 2014), fails to predict the negative effect of the price floor on cooperation in the Bertrand Competition.

- Rong and Halpern's (2013) notion of *cooperative equilibrium* (which is different from that of Capraro (2013)) focuses on 2-player games. However, the definition for games with greater than 2 players does not predict the decrease in cooperation as N increases in Bertrand Competition, nor the increase as N increases in the Public Goods Game.

The one approach besides ours that we are aware of that obtains all the regularities discussed above is that of Charness and Rabin (2002). Charness and Rabin, like Fehr and Schmidt (1999), assume that agents play a Nash equilibrium of a modified game, where players care not only about their personal material payoff, but also about the social welfare and the outcome of the least fortunate person. Specifically, player i's utility is assumed to be $(1 - a_i^{CR})u_i(s) + a_i^{CR}(b_i^{CR} \min_{j=1,\ldots,N} u_j(s) + (1 - b_i^{CR}) \sum_{j=1}^{N} u_j(s))$. Assuming, as in our model, that agents believe that other players either cooperate or defect and that they cooperate with probability β, then it is not hard to see that Charness and Rabin (2002) also predict all the regularities that we have been considering.

Although it seems difficult to distinguish our model from that of Charness and Rabin (2002) if we consider only social dilemmas, they are distinguishable if we look at other settings and take into account the other reason we mentioned for translucency: that other people in their social group might discover how they acted. We can easily capture this in the framework we have been considering by doubling the number of agents; for each player i, we add another player i^* that represent's i's social network. Player i^* can play only two actions: n (for "did *n*ot observe player i's action) and o (for "*o*bserved player i's action").[4] The payoffs of these new players are irrelevant. Player i's payoff depends on the action of player i^*, but not on the actions of player j^* for $j^* \neq i^*$. Now player i must have a prior probability γ_i about whether his action will be observed; in a social dilemma, this probability might increase to $\gamma_i' \geq \gamma_i$ if he intends to cooperate but instead deviates and defects. It should be clear that, even if $\gamma_i' = \gamma_i$, if we assume that player i's utilities are significantly lower if his non-cooperative action is observed, with this framework we would get qualitatively similar results for social dilemmas to the ones that we have already obtained.

The advantage of taking into account what your social group thinks is that it can be applied even to single-player games like the Dictator Game (Kahneman, Knetsch, and Thaler 1986). To do so, we would need to think about what a player's utility would be if his social group knew the extent to which he shared the pot. But it should be clear that reasonable assumptions here would lead to some degree of sharing.

While this would still not distinguish our predictions from those of the Charness-Rabin model, there is a variant of the Dictator Game considered by Capraro (2014) that does allow us to distinguish between the two. In this game, there are only two possible allocations of money: either the agent gets x and the other players gets $-x$, or the other player gets x and the agent gets $-x$. In this game, the Charness-Rabin approach would predict that the agent will choose to keep the x. But the translucency approach would allow that there would be types of agents who would think that their social group would approve of them giving away x, so, if the action was observed by their social group, they would get high utility by giving away x. And, indeed, Capraro's results show that a significant fraction (25%) of people do choose to give away x.

Of course, we do not have to assume $\alpha > 0$ to get cooperation in social dilemmas such as Traveler's Dilemma or Bertrand Competition. But we do if we want to consider what we believe is the appropriate equilibrium notion. Suppose that rational players are chosen at random from a population and play a social dilemma. Players will, of course, then update their beliefs about the likelihood of seeing cooperation, and perhaps change their strategy as a consequence. Will these beliefs stabilize and the strategies played stabilize? By *stability* here, we mean that (1)

[4] Alternatively, we could take player i's payoff to depend on the state of the world, where the state would model whether or not player i's action was observed.

players are all best responding to their beliefs, and (2) players' beliefs about the strategies played by others are correct: if player i ascribes probability p to player j playing a strategy s_j, then in fact a proportion p of players in the population play s_j. We have deliberately been fuzzy here about whether we mean best response in the sense of Definition 2.1 or Definition 2.2. If we use Definition 2.1 (or, equivalently use Definition 2.2 and take $\alpha = 0$), then it is easy to see (and well known) that the only way that this can happen is if the distribution of strategies played by the players represents a mixed strategy Nash equilibrium. On the other hand, if $\alpha > 0$ and we use Definition 2.2, then we can have stable beliefs that accurately reflect the strategies used and have cooperation (in all the other social dilemmas that we have studied). We make this precise in the appendix using the framework of Halpern and Pass (2013), by defining a notion of *translucent equilibrium*. Roughly speaking, we construct a model where, at all states, players are translucently rational (so we have common belief of translucent rationality), the strategies used are common knowledge, and we nevertheless have cooperation at some states. Propositions 4.1–4.4 play a key role in this construction; indeed, as long as the strategies used satisfy the constraints imposed by these results, we get a translucent equilibrium.

We have not focused on translucent equilibrium here in the main text because it makes strong assumptions about players' rationality and beliefs (e.g., it implicitly assumes common belief of translucent rationality). We do not need such strong assumptions for our results.

A Translucent equilibrium

In the main text of this paper we have described how cooperation can be rational if players are translucent, that is, if they believe that if they switch from one strategy to another, the fact that they choose to switch may be visible to the other players. In this appendix, we show how to use counterfactual structures to define a notion of equilibrium with translucent players and we observe that rationality of cooperation shown in the main text corresponds to having a mixed strategy translucent equilibrium, where cooperation is played with non-zero probability. We start by reviewing the relevant definitions from (Halpern and Pass 2013).

A.1 Game theory with translucent players

Let $\mathcal{G} = \mathcal{G}(P, S, u)$ be a (finite) normal form game, where $P = \{1, \ldots, N\}$ is the set of players, each of which has finite pure strategy set S_i and utility function u_i.

Definition A.1. A finite counterfactual structure appropriate for the game $\mathcal{G}$ is a tuple $M = (\Omega, \mathbf{s}, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_N)$, where:

- Ω is a finite space of states;
- $\mathbf{s} : \Omega \to S$ is the function that associates to each state ω the strategy profile that is supposed to be played at ω;
- f is the closest-state function, which describes what would happen if player i switched strategy to s'_i at state ω. Thus, $f : \Omega \times P \times S_i \to \Omega$ has to verify the following properties:

  CS1. $\mathbf{s}_i(f(\omega, i, s')) = s'_i$;

  CS2. $f(\omega, i, \mathbf{s}_i(\omega)) = \omega$.

  Property CS1 assures that, at state $f(\omega, i, s'_i)$, player i plays s'_i, and Property CS2 assures that the state does not change if player i does not change strategy.

- $\mathcal{PR}_i$ are player i's beliefs, which depends on the state i is reasoning about. Specifically, for each $\omega \in \Omega$, $\mathcal{PR}_i(\omega)$ is a probability measure on Ω satisfying the following properties:

  PR1. $\mathcal{PR}_i(\omega)(\{\omega' \in \Omega : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}) = 1$ (where $\mathbf{s}_i(\omega)$ denotes player i's strategy in $\mathbf{s}(\omega)$);

  PR2. $\mathcal{PR}_i(\omega)(\{\omega' \in \Omega : \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)\}) = 1$.

  These assumptions guarantee that player i assigns probability 1 to his actual strategy and beliefs. □

We can now define i's beliefs at ω if he were to switch to strategy s'. Intuitively, if he were to switch to strategy s' at ω, the probability that i would assign to state ω' is the sum of the probabilities that he assigns to all the states ω'' such that he believes that he would move from ω'' to ω' if he used strategy s'. Thus we define

$$\mathcal{PR}_{i,s'}(\omega)(\omega') := \sum_{\{\omega'' : f(\omega'', i, s') = \omega'\}} \mathcal{PR}_i(\omega)(\omega'').$$

We define the expected utility of player i at state ω in the usual way as the sum of the product of his expected utility of the strategy profile played at each state ω' and the probability of ω': $EU_i(\omega) = \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') u_i(\mathbf{s}_i(\omega), \mathbf{s}_{-i}(\omega'))$.[5]

[5] Given a profile $t = (t_1, \ldots, t_N)$, as usual, we define $t_{-i} = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_N)$. We extend this notation in the obvious way to functions like $\mathbf{s}$, so that, for example, $\mathbf{s}_{-i}(\omega) = (\mathbf{s}_1(\omega), \ldots, \mathbf{s}_{i-1}(\omega), \mathbf{s}_{i+1}(\omega), \ldots, \mathbf{s}_n(\omega))$.

Now we define i's expected utility at ω if he were to switch to s'. The usual way to do so is to simply replace i's actual strategy at ω by s' at all states, keeping the strategies of the other players the same; that is,

$$\sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') u_i(s', \mathbf{s}_{-i}(\omega')).$$

In this definition, player i's beliefs about the strategies that the other players are using do not change when he switches from $\mathbf{s}_i(\omega)$ to s'. The key point of counterfactual structures is that these beliefs may well change. Thus, we define i's expected utility at ω if he switches to s' as

$$EU_i(\omega, s') = \sum_{\omega' \in \Omega} \mathcal{PR}_{i,s'}(\omega)(\omega') u_i(s', \mathbf{s}_{-i}(\omega')).$$

Finally, we can define rationality in counterfactual structures using these notions:

Definition A.2. Player i is *rational* at state ω if, for all $s' \in S_i$,

$$EU_i(\omega) \geq EU_i(\omega, s').$$

□

A.2 Translucent equilibrium

In this section, we define translucent equilibrium and we observe that the results reported in the main text imply that social dilemmas have a counterfactual structure according to which each player plays, in equilibrium, his part of the welfare maximizing strategy with non-zero probability.

We start with some preliminary notation. Given a probability measure τ on a finite set T, let supp(τ) denote the support of τ, that is, supp$(\tau) = \{t \in T : \tau(t) \neq 0\}$. Given a mixed strategy profile σ, note that σ_{-i} can can be viewed as a probability on S_{-i}, where $\sigma_{-i}(s_{-i}) = \prod_{j \neq i} \sigma_j(s_j)$. Similarly σ can be viewed as a probability measure on S. In the sequel, we view σ_{-i} and σ as probability measures without further comment (and so talk about their support).

Definition A.3. A strategy profile σ in a game $\mathcal{G}$ is *translucent equilibrium* in a counterfactual structure $M = (\Omega, \mathbf{s}, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_N)$ appropriate for $\mathcal{G}$ if there exists a subset $\Omega' \subseteq \Omega$ such that, for each state ω in Ω', the following properties hold:

TE1. $\mathbf{s}(\omega) \in$ supp(σ);

TE2. $\mathrm{supp}(\mathcal{PR}_i(\omega)) \subseteq \Omega'$;

TE3. $\mathrm{s}_{-i}(\mathcal{PR}_i(\omega)) = \sigma_{-i}$ (i.e., for each strategy profile $s_{-i} \in S_{-i}$, we have $\sigma_{-i}(s_{-i}) = \mathcal{PR}_i(\omega)(\{\omega' : \mathrm{s}_{-i}(\omega') = s_{-i}\})$).

TE4. each player is rational at ω.

The mixed strategy profile σ is a translucent equilibrium of $\mathcal{G}$ if there exists a counterfactual structure M appropriate for $\mathcal{G}$ such that σ is a translucent equilibrium in M. □

Intuitively, σ is a translucent equilibrium in M if, for each strategy s_i in the support of σ_i, the expected utility of playing s_i given that other players are playing according to σ_{-i} is at least as good as switching to some other strategy s_i', given what i would believe about what strategies the other players are playing if he were to switch to s_i'.

This notion of translucent equilibrium is closely related to a condition called *IR* (for *individually rational*) by Halpern and Pass (2013). The main difference is that Halpern and Pass considered only pure strategy profiles; we allow mixed-strategy profiles here. We discuss the relationship between the notions at greater length in Section A.3.

A.3 Characterization of translucent equilibria

While it is easy to see that every Nash equilibrium is a translucent equilibrium (see Proposition A.4), the converse is far from true. As we show, for example, cooperation can be an equilibrium in social dilemmas (see below and Section A.4). In this section, we provide a characterization of translucent equilibria that will prove useful when discussing social dilemmas.

Proposition A.4. *Every Nash equilibrium of $\mathcal{G}$ is a translucent equilibrium.*

Proof. Given a Nash equilibrium $\sigma = (\sigma_1, \ldots, \sigma_n)$, consider the following counterfactual structure $M_\sigma = (\Omega, \mathrm{s}, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_N)$:

- Ω is the set of strategy profiles in the support of σ;
- $\mathrm{s}(s) = s$;
- $\mathcal{PR}_i(s_i, s_{-i})(s_i', s_{-i}') = \begin{cases} 0 & \text{if } s_i' \neq s_i \\ \sigma_{-i}(s_{-i}') & \text{if } s_i' = s_i; \end{cases}$
- $f((s_i, s_{-i}), i, s') = (s', s_{-i})$.

It is easy to check that σ is a translucent equilibrium in M_σ; we simply take $\Omega' = \Omega$. The fact that f is an "opaque" closest-state function, which is not affected by the strategy used by players, means that rationality in M reduces to the standard definition of rationality. We leave details to the reader. □

Although the fact that we can consider arbitrary counterfactual structures (appropriate for $\mathcal{G}$) means that many strategy profiles are translucent equilibria, the notion of translucent equilibrium has some bite. For example, the strategy profile (C, D), where player 1 cooperates and player 2 defects, is not a translucent equilibrium in Prisoner's Dilemma: if player 1 believes that player 2 is playing defecting with probability 1, there are no beliefs that 1 could have that would justify cooperation. However, as we shall see, both (C, C) and (D, D) are translucent equilibria. This follows from the characterization of translucent equilibrium that we now give.

Definition A.5. A mixed-strategy profile σ in $\mathcal{G}$ is *coherent* if for all players $i \in P$, all $s_i \in \text{supp}(\sigma_i)$, and all $s'_i \in S_i$, there is $s'_{-i} \in S_{-i}$ such that

$$u_i(s_i, \sigma_{-i}) \geq u_i(s')$$

(where, of course, $u_i(s_i, \sigma_{-i}) = \sum_{s''_{-i} \in S'_{-i}} \sigma_{-i}(s''_{-i}) u_i(s_i, s''_{-i})$). □

That is, σ is coherent if, for all pure strategies for player i in the support of σ_i, if i's belief about the strategies being played by the other players is given by σ_{-i}, there is no obviously better strategy that i can switch to in the weak sense that, if i contemplates switching to s'_i, there are beliefs that i could have about the other players (namely, that they would definitely play s'_{-i} in this case) that would make switching to s'_i better than sticking with s_i.

It is easy to see that (C, C) and (D, D) in Prisoner's Dilemma are both coherent; on the other hand, (C, D) is not.

Halpern and Pass (2013) define a pure strategy profile to be *individually rational* if it is coherent. Definition A.5 extends individual rationality to mixed strategies. Halpern and Pass prove that a pure strategy profile is individually rational if there is a model where it is commonly known that σ is played and there is common belief of rationality. The definition of translucent equilibrium can be seen as the generalization of this characterization of IR to mixed strategies. As the following theorem shows, we get an analogous representation.

Theorem A.6. *The mixed strategy profile σ of game $\mathcal{G}$ is coherent iff σ is a translucent equilibrium of $\mathcal{G}$.*

Proof. Let σ be a coherent strategy profile in $\mathcal{G}$. We construct a counterfactual structure $M = (\Omega, \mathrm{s}, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_N)$ as follows:

- $\Omega = S$;
- $\mathbf{s}(s) = s$;
- $\mathcal{PR}_i(\omega)(\omega') = \begin{cases} 1 & \text{if } \omega \notin \text{supp}(\sigma_i), \omega = \omega' \\ 0 & \text{if } \omega \notin \text{supp}(\sigma_i), \omega \neq \omega' \\ \sigma_{-i}(\mathbf{s}_{-i}(\omega')) & \text{if } \omega \in \text{supp}(\sigma_i), \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega) \\ 0 & \text{if } \omega \in \text{supp}(\sigma_i), \mathbf{s}_i(\omega') \neq \mathbf{s}_i(\omega); \end{cases}$
- $f(\omega, i, s_i') = \begin{cases} (s_i', \mathbf{s}_{-i}(\omega)) & \text{if } \omega \notin \text{supp}(\sigma_i) \\ \omega & \text{if } \omega \in \text{supp}(\sigma_i), s_i' = \mathbf{s}_i(\omega) \\ (s_i', s_{-i}') & \text{if } \omega \in \text{supp}(\sigma_i), s_i' \neq \mathbf{s}_i(\omega)\text{, where } s_i' \text{ is a} \\ & \text{strategy such that } u_i(\mathbf{s}_i(\omega), \sigma_{-i}) \geq u_i(s'); \\ & \text{such a strategy is guaranteed to exist since} \\ & \sigma \text{ is coherent.} \end{cases}$

We first show that M is a finite counterfactual structure appropriate for $\mathcal{G}$; in particular, $\mathcal{PR}_i$ satisfies PR1 and PR2 and f satisfies CS1 and CS2. For PR1 and PR2, there are two cases. If $\omega \notin \text{supp}(\sigma)$, then $\mathcal{PR}_i(\omega)(\omega) = 1$, so PR1 and PR2 clearly hold. If $\omega \notin \text{supp}(\omega)$, then $\mathcal{PR}_i(\omega)(\omega) > 0$ iff $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$. Moreover, if $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$, then it is immediate from the definition that $\mathcal{PR}_i(\omega) = \mathcal{PR}_i(\omega')$, so PR2. holds. That CS1 and CS2 hold is immediate from the definition of f.

To show that σ is a translucent equilibrium in M, let $\Omega' = \text{supp}(\sigma)$. For each state $\omega \in \Omega'$, TE1 clearly holds. Note that if $\omega \in \text{supp}(\sigma)$, then $\mathcal{PR}_i(\omega) = (\mathbf{s}_i(\omega), \sigma_{-i}(\omega))$ (identifying the strategy profile with a probability measure), so TE2 and TE3 clearly hold. It remains to show that TE4 holds, that is, that every player is rational at every state $\omega \in \Omega'$.

Thus, we must show that $EU_i(\omega) \geq EU(\omega, s_i^*)$ for all $s_i^* \in S_i$. Note that

$$\begin{array}{rcl} EU_i(\omega) & = & \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') u_i(\mathbf{s}_i(\omega), \mathbf{s}_{-i}(\omega')) \\ & = & \sum_{\{\omega' \in \Omega : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}} \sigma_{-i}(\mathbf{s}_{-i}(\omega')) u_i(\mathbf{s}_i(\omega), \mathbf{s}_{-i}(\omega')) \\ & = & \sum_{s''_{-i} \in S_{-i}} u_i(\mathbf{s}_i(\omega), s''_{-i}) \\ & = & u_i(\mathbf{s}_i(\omega), \sigma_{-i}). \end{array}$$

By definition,

$$EU_i(\omega, s_i^*) = \sum_{\omega' \in \Omega} \mathcal{PR}_{i, s_i^*}(\omega)(\omega') u_i(s_i^*, \mathbf{s}_{-i}(\omega'))$$

and

$$\mathcal{PR}_{i,s'}(\omega)(\omega') = \sum_{\{\omega'' : f(\omega'', i, s') = \omega'\}} \mathcal{PR}_i(\omega)(\omega'').$$

Now if $s_i^* = \mathbf{s}_i(\omega)$, then $f(\omega, i, s_i^*)$. In this case, it is easy to check that $\mathcal{PR}_{i,s_i^*}(\omega) = \mathcal{PR}_i(\omega)$, so $EU_i(\omega, s_i^*) = EU_i(\omega) = EU_i(s_i, \sigma_{-i})$, and TE4 clearly holds. On the other hand, if $s_i^* \neq \mathbf{s}_i(\omega)$, then

$$\begin{array}{rcl} EU_i(\omega, s_i^*) & = & \sum_{\omega' \in \Omega} \sum_{\{\omega'' : f(\omega'', i, s_i^*) = \omega'\}} \mathcal{PR}_i(\omega)(\omega'') u_i(s_i^*, \mathbf{s}_{-i}(\omega')) \\ & = & \sum_{\{\omega' \in \Omega : \mathbf{s}_i(\omega') = s_i^*\}} \sum_{\{\omega'' : f(\omega'', i, s_i^*) = \omega',\, \mathbf{s}_i(\omega'') = \mathbf{s}_i(\omega)\}} \sigma_{-i}(\omega'') u_i(s_i^*, \mathbf{s}_{-i}(\omega')) \\ & = & \sum_{\{\omega' \in \Omega : \mathbf{s}_i(\omega') = s_i^*\}} \sum_{\{\omega'' : f(\omega'', i, s_i^*) = \omega',\, \mathbf{s}_i(\omega'') = \mathbf{s}_i(\omega)\}} \sigma_{-i}(\omega'') u_i(f(\omega'', i, s_i^*)). \end{array}$$

By definition, $u_i(f(\omega'', i, s_i^*)) \leq u_i(\mathbf{s}_i(\omega''), \sigma_{-i}) = u_i(\mathbf{s}_i(\omega), \sigma_{-i})$. Thus,

$$\begin{array}{rcl} EU_i(\omega, s_i^*) & \leq & \sum_{\{\omega' \in \Omega : \mathbf{s}_i(\omega') = s_i^*\}} \sum_{\{\omega'' : f(\omega'', i, s_i^*) = \omega',\, \mathbf{s}_i(\omega'') = \mathbf{s}_i(\omega)\}} \sigma_{-i}(\omega'') u_i(\mathbf{s}_i(\omega), \sigma_{-i}) \\ & = & u_i(\mathbf{s}_i(\omega), \sigma_{-i}) \sum_{\{\omega' \in \Omega : \mathbf{s}_i(\omega') = s_i^*\}} \sum_{\{\omega'' : f(\omega'', i, s_i^*) = \omega',\, \mathbf{s}_i(\omega'') = \mathbf{s}_i(\omega)\}} \sigma_{-i}(\omega'') \\ & = & u_i(\mathbf{s}_i(\omega), \sigma_{-i}). \end{array}$$

This completes the proof that TE4 holds, and the proof of the "only if" direction of the argument

The "if" is actually much simpler. Suppose, by way of contradiction, that σ is not coherent. Then there is a player i and a strategy $s_i \in \text{supp}(\sigma_i)$ such that for all $s'_{-i} \in S_i$, we have $u_i(s_i, \sigma_{-i}) < u_i(s')$. It follows that, for all counterfactual structures M, no matter what the beliefs and the closest-state functions are in M, it is always strictly profitable for player i to switch strategy from s_i to s'_i. Consequently, i is not rational at a state ω such that $s_i(\omega) = s_i$, contradicting TE4. □

A.4 Translucent equilibrium in social dilemmas

As we now show, our characterizations of Propositions 4.1–4.4 can be used to provide conditions on when translucent equilibrium exists in these social dilemmas.

We start our analysis with Prisoner's Dilemma. We capture the assumption that β is the probability of cooperation, and that players either cooperate or defect, by assuming that players follow a mixed strategy where they cooperate with probability β and defect with probability $1 - \beta$.

Proposition A.7. *$(\beta_1 C + (1-\beta_1)D, \beta_2 C + (1-\beta_2)D)$ is a translucent equilibrium of Prisoner's dilemma iff $\beta_i b \geq c$, for $i = 1, 2$, or $\beta_1 = \beta_2 = 0$.*

Proof. Suppose that $(\beta_1 C + (1-\beta_1)D, \beta_2 C + (1-\beta_2)D)$. If $\beta_1 > 0$, then by Theorem A.6, it easily follows we must have $u_1(C, \beta_2 C + (1-\beta_2)D) \geq u_1(D, D)$. Thus, we must have $\beta_2(b - c) + (1 - \beta_2)(-c) \geq 0$; equivalently, $\beta_2 b \geq c$. Note that since $c > 0$, this means that we must have $\beta_2 > 0$. Similarly, if $\beta_2 > 0$, then $\beta_1 b \geq c$. By Theorem A.4, (D, D) is a translucent equilibrium, since it is a Nash equilibrium. Thus, either $\beta_i b \geq c$ for $i = 1, 2$ or $\beta_1 = \beta_2 = 0$.

Conversely, if $\beta_i b \geq c$ for $i = 1, 2$, then it again easily follows from Theorem A.6 that $(\beta_1 C + (1 - \beta_1)D, \beta_2 C + (1 - \beta_2)D)$ is a translucent equilibrium. As we have observed, (D, D) (the case that $\beta_1 = \beta_2 = 0$) is also a translucent equilibrium. □

Proposition A.7 is not all that interesting, since it does not take into account a player's beliefs regarding translucency. The following definition is a step towards doing this. Suppose that M is counterfactual structure appropriate for a social dilemma Γ. *Player i has type α_i in M* if, at each state ω in M, player i believes that if he intends to cooperate in ω and deviates from that, then each other agent will independently realize this with probability α_i and will defect. Formally, this means that, at each state ω in M, we have

- if $\mathbf{s}_i(\omega) = s_i^W$ (i.e., i is cooperating in ω by playing his component of the social-welfare maximing strategy profile), then, for each $J \subseteq P \setminus \{i\}$, we have $\mathcal{PR}_i(\omega)(\{\omega' : f(\omega', i, s_i^N) = \omega'', s_j(\omega') = s_j^C, s_j(\omega'') = s_j^N, \forall j \in J\}) = \alpha_i^{|J|}\mathcal{PR}_i(\omega)\{\omega' : s_j(\omega') = s_j^C, \forall j \in J\})$.

Proposition A.8. *$(\beta_1 C + (1-\beta_1)D, \beta_2 C + (1-\beta_2)D)$ is a translucent equilibrium of the Prisoner's dilemma in a structure where player i has type α_i if and only if $\beta_1 = \beta_2 = 0$ or $\alpha_i \beta_{3-i} b \geq c$ for $i \geq 1, 2$.*

Proof. Suppose that $\alpha_i \beta_{3-i} b \geq c$ for $i = 1, 2$ or $\beta_1 = \beta_2 = 0$. We show that $(\beta_1 C + (1 - \beta_1)D, \beta_2 C + (1 - \beta_2)D)$ is a translucent equilibrium in a structure where player i has type α_i. Consider the counterfactual structure $M(\alpha_1, \alpha_2)$ defined as follows

- $\Omega = \{C, D\} \times \{0, 1\}^2$. (The second component of the state, which is an element of $\{0, 1\}^2$, is used to determine the closest-state function. Roughly speaking, if $v_j = 1$, then player j learns about a deviation if there is one; if $v_j = 0$, he does not.)
- $\mathbf{s}((s, v)) = s$.
- $f((s, v), i, s_i^*) = \begin{cases} (s, v) & \text{if } s_i = s_i^*, \\ (s', v) & \text{if } s_i \neq s_i^*, \text{where } s_i' = s_i^* \text{ and for } j \neq i, \\ & s_j' = s_j \text{ if } v_j = 0 \text{ and } s_j' = s_j^N \text{ if } v_j = 1. \end{cases}$

  Thus, if player i changes strategy from s_i to s_i', $s_i' \neq s_i$, then each other player j either deviates to his component of the Nash equilibrium or continues with his current strategy, depending on whether v_j is 0 or 1. Roughly speaking, he switches to his component of the Nash equilibrium if he learns about a deviation (i.e., if $v_j = 1$).

- $\mathcal{PR}_i(s,v)(s',v') = \begin{cases} 0 & \text{if } s_i \neq s'_i, \text{or } v_i \neq v'_i, \\ \sigma_{3-i}(s_{3-i})\pi_i(v_{3-i}) & \text{if } s = s', \end{cases}$ where σ_{3-i} is the distribution on strategies that puts probability β_{3-i} on C and probability $1-\beta_{3-i}$ on D, while π_i is the distribution that puts probability α_i on 1 and probability $1-\alpha_i$ on 0. Thus, if $s = s'$, then the probability of the v' component is determined by assuming that the other player ($3-i$) independently learns about a deviation by i with probability α_i.

Clearly, $M(\alpha_1,\alpha_2)$ is a structure where player i has type α_i, for $i = 1,2$. We claim that $(\beta_1 C + (1-\beta_1)D, \beta_2 C + (1-\beta_2)D)$ is a translucent equilibrium in the counterfactual structure $M(\alpha_1,\alpha_2)$.

There are two cases. If $\beta_1 = \beta_2 = 0$, then let Ω' consist of all states of the form $((D,D),v)$. It is easy to check that TE1–4 hold. If $\alpha_i\beta_{3-i}b \geq c$ for $i \geq 1,2$, let $\Omega' = \Omega$. It is immediate that TE1, TE2, and TE3 hold. Since $\alpha_i\beta_{3-i}b \geq c$, it follows from Proposition 4.1 that player i is rational at each state in Ω; thus, TE4 holds.

For the converse, suppose that M is a structure where player i has type α_i, for $i = 1,2$, and $(\beta_1 C + (1-\beta_1)D, \beta_2 C + (1-\beta_2)D)$ is a translucent equilibrium in M. If it is not the case that either $\beta_1 = \beta_2 = 0$ or $\alpha_i\beta_{3-i}b \geq c$ for $i = 1,2$, without loss of generality we can assume that $\beta_1 > 0$ and that $\alpha_1\beta_2 b < c$. Let ω be a state in the set Ω' where player 1 cooperates. Since player 1 must be rational at Ω', we must have $u_1(C, \beta_2 C + (1-\beta_2)D) \geq ((1-\beta_2)+\alpha_1\beta_2)u_1(D,D) + (1-\alpha_1)\beta_2 u_1(D,C)$. Simple calculations show that this inequality holds iff $\beta_2(b-c) + (1-\beta_2)(-c) \geq (1-\alpha_1)\beta_2 b$ or, equivalently, $\alpha_1\beta_2 b \geq c$. This gives the desired contradiction. □

The following propositions can be proved in a similar fashion. We leave details to the reader.

Proposition A.9. *$(\beta_1 H + (1-\beta_1)L, \beta_2 H + (1-\beta_2)L)$ is a translucent equilibrium of the Traveler's dilemma if and only if $b \leq \frac{(H-L)\beta_i}{1-\beta_i}$, for $i = 1,2$, or $\beta_1 = \beta_2 = 0$.* □

Proposition A.10. *$(\beta_1 H + (1-\beta_1)L, \beta_2 H + (1-\beta_2)L)$ is a translucent equilibrium of the Traveler's dilemma in a structure where player i has type α_i if and only if $\beta_1 = \beta_2 = 0$ or*

$$b \leq \begin{cases} \frac{(H-L)\beta_{3-i}}{1-\alpha_i\beta_{3-i}} & \textit{if } \alpha_i \geq \frac{1}{2} \\ \min\left(\frac{(H-L)\beta_{3-i}}{1-\alpha_i\beta_{3-i}}, \frac{H-L-1}{1-2\alpha_i}\right) & \textit{if } \alpha_i < \frac{1}{2}. \end{cases}$$

□

In the following propositions, let C and D denote, respectively, the full contribution and the null contribution in the Public Goods game. Given an N-tuple $(r_1, \ldots, r_N)$ of real numbers, $\bar{r}_{-i}$ denotes the average of the numbers r_j, with $j \neq i$.

Proposition A.11. *$(\beta_1 C + (1-\beta_1)D, \ldots, \beta_N C + (1-\beta_N)D)$ is a translucent equilibrium of the Public Goods game if and only if $\rho\bar{\beta}_{-i}(N-1) \geq 1-\rho$ for all i, or $\beta_i = 0$ for all i.* □

Proposition A.12. *$(\beta_1 C + (1-\beta_1)D, \ldots, \beta_N C + (1-\beta_N)D)$ is a translucent equilibrium of the Public Goods game in a structure where player i has type α_i if and only if $\beta_i = 0$ for all i or $\alpha_i\rho\bar{\beta}_{-i} \geq 1-\rho$ for all i.* □

Proposition A.13. *$(\beta_1 H + (1-\beta_1)L, \ldots, \beta_N H + (1-\beta_N)L)$ is a translucent equilibrium of the Bertrand competition if and only if $\beta_i = 0$ for all i, or $\prod_{j\neq i}\beta_j \geq \frac{L}{H}$ for all i.* □

Proposition A.14. *$(\beta_1 H + (1-\beta_1)L, \ldots, \beta_N H + (1-\beta_N)L)$ is a translucent equilibrium of the Bertrand competition in a structure where player i has type α_i if and only if $\beta_i = 0$ for all i, or $\prod_{j\neq i}\beta_j \geq f(\gamma_{i,j}, N)LN/H$ for all i, where $f(\gamma_{i,j}, N) = \sum_{J\subseteq P-\{i\}}(\prod_{j\in P-(J\cup\{i\})}\gamma_{i,j}\prod_{j\in J}(1-\gamma_{i,j}))/(|J|+1)$ and $\gamma_{i,j} = (1-\alpha_i)\beta_j$.* □

Acknowledgments: We thank Krzysztof Apt for useful comments on an earlier draft of this paper. Joseph Halpern was supported in part by NSF grants IIS-0911036 and CCF-1214844, AFOSR grant FA9550-08-1-0438, ARO grant W911NF-14-1-0017, and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by AFOSR under grant FA9550-12-1-0040. Valerio Capraro was funded by the Dutch Research Organization (NWO) grant 612.001.352.

References

Apt, K. and G. Schäfer (2014). Selfishness level of strategic games. *Journal of Artificial Intelligence Research 49*, 207–240.

Barcelo, H. and V. Capraro (2014). Group size effect on cooperation in social dilemmas. Working paper, available.

Basu, K. (1994). The traveler's dilemma: paradoxes of rationality in game theory. *American Economic Review 84*(2), 391–395.

Bateson, M., L. Callow, J. R. Holmes, M. L. Redmond Roche, and D. Nettle (2013). Do images of "watching eyes" induce behaviour that is more prosocial or more normative? A field experiment on littering. *PLoS ONE 8*(12).

Bolton, G. E. and A. Ockenfels (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review 90*(1), 166–193.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, N. J.: Princeton University Press.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics 119*, 861–897.

Capra, M., J. K. Goeree, R. Gomez, and C. A. Holt (1999). Anomalous behavior in a traveler's dilemma. *American Economic Review 89*(3), 678–690.

Capraro, V. (2013). A model of human cooperation in social dilemmas. *PLoS ONE 8*(8), e72427.

Capraro, V. (2014). The emergence of altruistic behaviour in conflictual situations. Working Paper.

Capraro, V., J. J. Jordan, and D. G. Rand (2014). Heuristics guide the implementation of social preferences in one-shot prisoner's dilemma experiments. *Scientific Reports*. In press.

Capraro, V., M. Venanzi, M. Polukarov, and N. R. Jennings (2013). Cooperative equilibria in iterated social dilemmas. In *Proc. Sixth International Symposium on Algorithmic Game Theory (SAGT '13)*, pp. 146–158.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

Costa-Gomes, M., V. Crawford, and B. Broseta (2001). Cognition and behavior in normal form games: An experimental study. *Econometrica 69*(5), 1193–1235.

Daley, B. and P. Sadowski (2014). A strategic model of magical thinking: Axioms and analysis. Available at http://www.princeton.edu/economics/seminar-schedule-by-prog/behavioralf14/Daley_Sadowski_MT.pdf.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology 31*, 169–193.

Dufwenberg, M. and U. Gneezy (2002). Information disclosure in auctions: an experiment. *Journal of Economic Behavior and Organization 48*, 431–444.

Dufwenberg, M., U. Gneezy, J. K. Goeree, and R. Nagel (2007). Price floors and competition. *Special Issue of Economic Theory 33*, 211–224.

Ekman, P. and W. Friesen (1969). Nonverbal leakage and clues to deception. *Psychiatry 32*, 88–105.

Engel, C. and L. Zhurakhovska (2012). When is the risk of cooperation with taking? The Prisoner's Dilemma as a game of multiple motives. Working Paper.

Fehr, E. and K. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

Gunnthorsdottir, A., D. Houser, and K. McCabe (2007). Dispositions, history and contributions in public goods experiments. *Journal of Economic Behavior and Organization 62*(2), 304–315.

Halpern, J. Y. and R. Pass (2012). Iterated regret minimization: a new solution concept. *Games and Economic Behavior 74*(1), 194–207.

Halpern, J. Y. and R. Pass (2013). Game theory with translucent players. In *Theoretical Aspects of Rationality and Knowledge: Proc. Fourteenth Conference (TARK 2013)*, pp. 216–221.

Halpern, J. Y. and R. Pass (2014). Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory*. To appear. A preliminary version entitled "Game theory with costly computation" appears in *Proc. First Symposium on Innovations in Computer Science*, 2010.

Hamilton, W. D. (1964). The genetical evolution of social behavior. i. *Journal of Theoretical Biology 7*, 1–16.

Isaac, M. R., J. M. Walker, and S. Thomas (1984). Divergent evidence on free riding: an experimental examination of possible explanations. *Public Choice 43*(1), 113–149.

Isaac, M. R., J. M. Walker, and A. W. Williams (1994). Group size and the voluntary provision of public goods. *Journal of Public Economics 54*, 1–36.

Kahneman, D., J. Knetsch, and R. H. Thaler (1986). Fairness and the assumptions of economics. *Journal of Business 59*(4), S285–300.

Masel, J. (2007). A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. *Journal of Economic Behavior and Organization 64*(1), 216–231.

McKelvey, R. and T. Palfrey (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior 10*(1), 6–38.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science 314*(5805), 1560–1563.

Nowak, M. A. and K. Sigmund (1998). Evolution of indirect reciprocity by image scoring. *Nature 393*, 573–577.

Rand, D. G., J. D. Green, and M. A. Nowak (2012). Spontaneous giving and calculated greed. *Nature 489*, 427–430.

Rand, D. G., A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene (2014). Social heuristics shape intuitive cooperation. *Nature Communications 5*, 3677.

Rapoport, A. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan Press.

Renou, L. and K. H. Schlag (2010). Minimax regret and strategic uncertainty. *Journal of Economic Theory 145*, 264–286.

Rong, N. and J. Y. Halpern (2013). Towards a deeper understanding of cooperative equilibrium: characterization and complexity. In *Proc. Twelfth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 319–326.

Salcedo, B. (2013). Implementation without commitment in moral hazard environments. Working paper.

Shafir, E. and A. Tversky (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology 24*, 449–474.

Stahl, D. and P. Wilson (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization 25*(3), 309–327.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology 46*, 35–57.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics 6*, 299–310.