# 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

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 [27] call *translucent players*. Typically, players are assumed to be *opaque*, in the sense that a deviation by one player in a normal-form game 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 [7, 15, 28, 35, 46]; yet others assume that players have other-regarding preferences [5, 14, 23]; others define radically different solution concepts, assuming that players do not try to maximize their payoff, but rather try to minimize their regret [26, 41], or maximize the forecasts associated to coalition structures [9, 13], or maximize the total welfare [1, 42]. (These references only scratch the surface; a complete bibliography would be longer than this paper!)

Cooperative behavior 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 [6], where none of the five rules of cooperation proposed by Nowak [36] seems to be at play.

Here we propose a novel approach to explain cooperation, based on work of Halpern and Pass [27] and Salcedo [43], assuming what Halpern and Pass call *translucent players*. Typically, players are assumed to be *opaque*, in the sense that a deviation by one player in a normal-form game 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

Submitted to:
TARK 2015

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., [21]. Professional poker players are also very sensitive to *tells*—betting patterns and physical demeanor that reveal something about a player's hand and strategy.)[1]

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., [38, 39]), 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 possible leaks about changes in plans, 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 transparency, 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 experimental evidence that can be viewed as providing support for players believing that they are transparent. Gilovich et al. [24] show that people tend to overestimate the extent to which others can discern their internal states. For instance, they showed that liars overestimate the detectability of their lies and that people believe that their feelings of disgust are more apparent than they actually are. There is also 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 [4] 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 transparent.

We apply the idea of translucency to a particular class of games that we call *social dilemmas* (cf. [18]). A social dilemma is a normal-form game with two aproperties:

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.

These uniqueness assumptions are not necessary, but they make definitions and computations easier. Although these restrictions are nontrivial, many of the best-studied games in the game-theory literature satisfy them, including Prisoner's Dilemma, Traveler's Dilemma [3], 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.

[1]The idea of translucency is motivated by some of the same concerns as Solan and Yariv's [45] *games with espionage*, but the technical details are quite different. A game with espionage is a two-player extensive-form game that extends an underlying normal-form game by adding a step where player 1 can purchase some noisy information about player 2's planned move. Here, the information is free and all players may be translucent. Moreover, the effect of the translucency is modeled by the players' counterfactual beliefs rather than by adding a move to the game.

For definiteness, we focus here on the first point and assume that, in social dilemmas, players have a degree of belief $\alpha$ that they are transparent, so that if they intend to cooperate (by playing their component of the welfare-maximizing strategy) and decide to deviate, there is a probability $\alpha$ that another player will detect this, and play her component of the Nash equilibrium strategy. (The assumption that cooperation acts as a default strategy is supported by experiments showing that people forced to make a decision under time pressure are, on average, more cooperative than those forced to made a decision under time delay [38, 39]. Rand and his colleagues suggest that this is due to the internalization of strategies that are successful in everyday interactions.) We assume that 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 transparency in much the same way, and would get qualitatively similar results (see Section 6).

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. We discuss related work in Section 5. Section 6 concludes. Proofs are deferred to the full paper, where we also 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 [27], 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 [27].

Formally, a (finite) normal-form game $\mathscr{G}$ 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 $\mu_i$ on $S_{-i}$ if, for all strategies $s_i'$ for player $i$,

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

□

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 modify the standard definition of best response by defining best response 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 $s_i' \in S_i$,

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

□

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.** 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 $\Gamma$ 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 $\alpha$, 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 $\mu'$ 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 $\mu^{s_i,s_i}$) and, with probability $\alpha$, 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. Many personal and professional relationships, the depletion of natural resources, climate protection, the security of energy supply, and price competition in markets can all be viewed as 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. The condition $b > c$ implies that $(D,D)$ is the unique Nash equilibrium and $(C,C)$ is the unique welfare-maximizing profile.

**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 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 $b$.

**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 contribution pot is then multiplied by a constant strictly between 1 and $N$, and then evenly redistributed among all players.[2] 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 $\rho N$ 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.[3]

From here on, we say that a player *cooperates* if he plays his part of the 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 [30] or when future interactions among the same subjects are likely [37, 47]; see [36] for a review of the five rules of cooperation. However, we often observe cooperation even in one-shot anonymous experiments among unrelated players [40].

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 [11, 22, 40].
- The degree of cooperation in the Traveler's Dilemma depends negatively on the bonus/penalty [8].
- The degree of cooperation in the Public Goods game depends positively on the constant marginal return [25, 31].
- The degree of cooperation in the Public Goods game depends positively on the number of players [2, 32, 48].
- The degree of cooperation in the Bertrand Competition depends negatively on the number of players [19].
- The degree of cooperation in the Bertrand Competition depends negatively on the price floor [20].

[2] We thus consider only *linear* Public Goods games. This choice is motivated by the fact that our purpose is to compare the predictions of our model with experimental data. Most experiments have adopted linear Public Goods games, since they have much easier instructions and thus they minimize noise due to participants not understanding the rules of the game.

[3] 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 $N \geq 2$.

## 4 Explaining social dilemmas using translucency

As we suggested in the introduction, we hope to use translucency to explain cooperation in social dilemmas even when players cannot see each other. We expect that people get so used to assuming some degree of transparency in their everyday interactions, which are typically face-to-face, that they bring these strategies and beliefs in the lab setting, even though they are arguably inappropriate.

To do this, we have to make assumptions about an agent's beliefs. Say that an agent $i$ has *type* $(\alpha,\beta,C)$ 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 $\alpha$; (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 $\beta$.

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 [11]. 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 $(\alpha,\beta,C)$ player believes that other players will cooperate with probability $\beta$ (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.

Our assumptions ensure 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 $\mu_i^J$ 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, $\mu_i^J$ 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 $\alpha$ and $\beta$ 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 $\beta$ 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; the proofs can be found in the full paper. We start with Prisoner's 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* $(\alpha,\beta,C)$ *to cooperate if and only if* $\alpha\beta b \geq c$. □

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 $\alpha$, the greater the benefit of cooperation and the smaller the cost, then the smaller the value of $\beta$ that still allows cooperation to be a best response.

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

**Proposition 4.2.** *In Traveler's Dilemma, it is translucently rational for a player of* $(\alpha,\beta,C)$ *to cooperate if and only if*

$$b \leq \begin{cases} \frac{(H-L)\beta}{1-\alpha\beta} & \text{if } \alpha \geq \frac{1}{2} \\ \min\left(\frac{(H-L)\beta}{1-\alpha\beta}, \frac{H-L-1}{1-2\alpha}\right) & \text{if } \alpha < \frac{1}{2}. \end{cases}$$

□

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 $\beta$ increases $\frac{(H-L)\beta}{1-\alpha\beta}$ follows from straightforward calculus.)

We next consider the Public Goods game. Recall the $\rho$ 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* $(\alpha,\beta,C)$ *to cooperate if and only if* $\alpha\beta\rho(N-1) \geq 1-\rho$. □

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 $\alpha$ and $\beta$, there is guaranteed to be an $N_0$ such that cooperation is a best response for all $N \geq N_0$; moreover, for fixed $\alpha$, as $N$ gets larger, smaller and smaller $\beta$s 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* $(\alpha,\beta,C)$ *to cooperate iff* $\beta^{N-1} \geq \max(\gamma^{N-1}N(H-1)/H, f(\gamma,N)LN/H)$, *where* $\gamma = (1-\alpha)\beta$ *and* $f(\gamma,N) = \sum_{k=0}^{N-1}\binom{N-1}{k}(1-\gamma)^k\gamma^{N-k-1}/(k+1)$. □

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 $\alpha$ and $\beta$ 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 $(\alpha,\beta,C)$, 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 $\alpha$ and $\beta$ allow cooperation as $L/H$ gets smaller. In particular, if we fix $H$ and raise the floor $L$, fewer values of $\alpha$ and $\beta$ 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 $(\alpha,\beta,C)$. Formally, assume that for each pair $(u,v)$ and $(u',v')$ of open intervals in $[0,1]$, there is a positive probability of finding someone of type $(\alpha,\beta,C)$ 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 Comparison to other approaches

Here we show that approaches (that we are aware of) other than that of Charness and Rabin and possibly that of Bolton and Ockenfels are not able to obtain all the regularities that we mentioned in Section 3. We consider a number of approaches in turn.

- The Fehr and Schmidt [23] *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 $\beta$ that other agents will cooperate and that the other agents either cooperate or defect, Fehr and Schmidt [23] model does not make any clear prediction of a group-size effect on cooperation in the public goods game.
- McKelvey and Palfrey's [35] *quantal response equilibrium (QRE)* is defined as follows.[4] Taking $\sigma_i(s)$ to be the probability that mixed strategy $\sigma_i$ assigns to the pure strategy $s$, given $\lambda > 0$, a mixed strategy profile $\sigma$ is a 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 $\lambda$, 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 $\sigma$ 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 [11] with $b/c = 10$).
- *Iterated regret minimization* [26] 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 [9] 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 [2], fails to predict the negative effect of the price floor on cooperation in the Bertrand Competition.
- Rong and Halpern's [29, 42] notion of *cooperative equilibrium* (which is different from that of Capraro [9]) 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.

[4] We actually define here a particular instance of QRE called the *logit QRE*; $\lambda$ is a free parameter of this model.

- Bolton and Ockenfels' [5] *inequity-aversion model* assumes that a player $i$ aims at maximizing his or her *motivational function* $v_i = v_i(x_i, \sigma_i)$, where $x_i$ is $i$'s monetary payoff and $\sigma_i = \sigma_i(x_1, \sum_{j=1,\ldots,N} x_j) = x_i / \sum_{j=1,\ldots,N} x_j$. The motivational function is assumed to be twice differentiable, weakly increasing in the first argument, and concave in the second argument with a maximum at $\sigma_i = \frac{1}{N}$, but otherwise is unconstrained. For each of the social dilemmas that we have considered, it is not hard to define a motivational function that will obtain the regularities observed. However, we have not been able to find a single motivational function that gives the observed regularities for all four social dilemmas that we have considered. In any case, just as with the Charness and Rabin model, once we consider the interaction between social groups and translucency, we can distinguish our approach from this inequity-aversion model. Specifically, consider a situation where people are given a choice between giving \$1 to an anonymous stranger, rather than burning it. In such a situation, inequity aversion would predict that people would burn the dollar to maintain equity (i.e., a situation where no one gets \$1). However, perhaps not surprisingly, Capraro et al. [12] found that over 90% people prefer giving away the dollar to burning it. Of course, translucency (and a number of other approaches) would have no difficulty in explaining this phenomenon.

The one approach besides ours that we are aware of that obtains all the regularities discussed above is that of Charness and Rabin [14]. Charness and Rabin, like Fehr and Schmidt [23], 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 $\beta$, then it is not hard to see that Charness and Rabin [14] also predict all the regularities that we have been considering.

Although it seems difficult to distinguish our model from that of Charness and Rabin [14] if we consider only social dilemmas, the models 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").[5] 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 $\gamma_i$ 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. Again, a player has beliefs about the extent to which he is transparent, and we can set the payoffs so that the effects of transparency are the same if a player's social network learns about his actions and if other players learn about his action.

The advantage of taking into account what your social group thinks is that it allows us to apply ideas of translucency even to single-player games like the Dictator Game [33]. To do so, we need to make assumptions 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.

[5] 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.

While this would still not distinguish our predictions from those of the Charness-Rabin model, there is a variant of the Dictator Game that has recently been considered to show existence of hyper-altruism in conflict situations [16, 10]. In the simplest version of this game, there are only two possible allocations of money: either the agent gets $x$ and the other player 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 either keep $x$ or be indifferent between keeping $x$ and giving it away. But assuming translucency allows for the possibility that some types of agents would think that their social group would approve of them giving away $x$, so if the action were observed by their social group, they would get high utility by giving away $x$. However, recent results by Capraro [10] show that a significant fraction (1/6) of people are *hyper-altruistic*: they strictly prefer giving away $x$ to keeping it [10].

Just to be clear, we do not mean to imply that translucency is the unique "right" explanation for cooperation in social dilemmas and all the other explanations that we discussed above are "wrong". There are probably a number of factors that contribute to cooperation. We hope in future work to tease these apart.

# 6 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 [44] 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's choices then there is no translucency, so 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* [44], 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 [34] in the context of the Public Goods game and by Daley and Sadowski [17] in the context of symmetric $2 \times 2$ games. The notion of translucency goes beyond these models, since it may be 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 $\alpha$ 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 $\alpha$.

One feature of our approach is that, at least if we take the concern with translucency to be due to an opponent discovering what you are going to do (rather than other members of your social group

discovering what you are going to do), then, unlike many other approaches to explaining social dilemmas, our approach does not involve modifying the utility function; that is, we can apply translucency while still identifying utility with the material payoff. While this make it an arguably simpler explanation, that does not necessarily make it "right", of course. We do not in fact believe that there is a unique "right" explanation for cooperation in social dilemmas and all the other explanations that we discussed above are "wrong". There are probably a number of factors that contribute to cooperation. We hope in future work to tease these apart.

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) 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 full paper, using the framework of Halpern and Pass [27], 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.

In the full paper, we also characterize those profiles of strategies that can be translucent equilibria, using ideas similar in spirit to those of Halpern and Pass [27]. While allowing people to believe that they are to a certain extent transparent means that the set of translucent equilibria is a superset of the set of Nash equilibria, not all strategy profiles can be translucent equilibria. For example, (C,D) is not a translucent equilibrium in Prisoner's dilemma. We have not focused on translucent equilibrium 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.

## References

[1] K. Apt & G. Schäfer (2014): *Selfishness level of strategic games*. Journal of Artificial Intelligence Research 49, pp. 207–240.

[2] H. Barcelo & V. Capraro (2015): *Group size effect on cooperation in one-shot social dilemmas*. Scientific Reports 5.

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

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

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

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

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

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

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

[10] V. Capraro (2014): *The emergence of altruistic behaviour in conflictual situations*. Working Paper.

[11] V. Capraro, J. J. Jordan & D. G. Rand (2014): *Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments*. Scientific Reports 4.

[12] V. Capraro, C. Smyth, K. Mylona & G. A. Niblo (2014): *Benevolent characteristics promote cooperative behaviour among humans*. PloS ONE 9(8), p. e102881.

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

[14] G. Charness & M. Rabin (2002): *Understanding social preferences with simple tests*. Quarterly Journal of Economics 117(3), pp. 817–869.

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

[16] M. J. Crockett, Z. Kurth-Nelson, J. Z. Siegel, P. Dayan & R. J. Dolan (2014): *Harm to others outweighs harm to self in moral decision making*. Proceedings of the National Academy of Sciences 111(48), pp. 17320–17325.

[17] B. Daley & 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.

[18] R. M. Dawes (1980): *Social dilemmas*. Annual Review of Psychology 31, pp. 169–193.

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

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

[21] P. Ekman & W.V. Friesen (1969): *Nonverbal leakage and clues to deception*. Psychiatry 32, pp. 88–105.

[22] C. Engel & L. Zhurakhovska (2012): *When is the risk of cooperation worth taking? The Prisoner's Dilemma as a game of multiple motives*. Working Paper.

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

[24] T. Gilovich, K. Savitsky & V. H. Medvec (1998): *The illusion of transparency: biased assessments of others' ability to read one's emotional states.* Journal of Personality and Social Psychology 75(2), p. 332.

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

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

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

[28] J. Y. Halpern & R. Pass (2015): *Algorithmic rationality: Game theory with costly computation*. Journal of Economic Theory 156, pp. 246–268.

[29] J. Y. Halpern & N. Rong (2010): *Cooperative equilibrium*. In: *Proc. Ninth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 1465–1466.

[30] W. D. Hamilton (1964): *The genetical evolution of social behavior. I.* *Journal of Theoretical Biology* 7, pp. 1–16.

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

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

[33] D. Kahneman, J.L. Knetsch & R. H. Thaler (1986): *Fairness and the assumptions of economics*. *Journal of Business* 59(4), pp. S285–300.

[34] J. Masel (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), pp. 216–231.

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

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

[37] M. A. Nowak & K. Sigmund (1998): *Evolution of indirect reciprocity by image scoring*. *Nature* 393, pp. 573–577.

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

[39] D. G. Rand, A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak & J. D. Greene (2014): *Social Heuristics shape intuitive cooperation*. *Nature Communications* 5, p. 3677.

[40] A. Rapoport (1965): *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press.

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

[42] N. Rong & 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.

[43] B. Salcedo (2013): *Implementation without commitment in moral hazard environments*. Working paper.

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

[45] E. Solan & L. Yariv (2004): *Games with espionage*. *Games and Economic Behavior* 47, pp. 172–199.

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

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

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