Altruism, predation and the Samaritan's dilemma

Stefano Dughera, Alain Marciano

Altruism, predation and the Samaritan's dilemma

Stefano Dughera[a,b], Alain Marciano[b]

a University of Paris Nanterre, EconomiX - UMR CNRS, Avenue de la République 200 – 92001 Nanterre – France;

b University of Torino, Department of Economics and Statistics, Lungo Dora Siena 100/A – 10153 Torino;

c corresponding author, e-mail: alain.marciano@umontpellier.fr. MRE and University of Montpellier, University of Montpellier, Faculté d'économie Avenue Raymond Dugrand, CS 79606, F-34960 Montpellier cedex 2 France.

Altruism, predation and the Samaritan's dilemma

Abstract: The goal of this paper is to study the consequences of non-reciprocal or unilateral altruism, that is, of altruism between individuals who have different concern for others. By contrast to what the literature usually shows—that unilateral altruists lead egoists to cooperate, that non-reciprocal altruism destroys altruism or that it generates non-desirable exploitation—we show that unilateral altruism does *not* forcedly lead egoists to cooperate nor it destroys altruism and that, in some situations, it can even be Pareto improving. By analyzing a simple cooperation game with other-regarding preferences, we find that unilateral altruism gives birth to a Samaritan's Dilemma where egoists predate Samaritans by free-riding on their contribution. Perhaps counterintuitively, we also show that in case "exploited" Samaritans experience a higher subjective well-being than in a classical Prisoners' dilemma. Finally, we derive conditions for the evolutionary stability of both the predators' and Samaritans' behavior.

Keywords: altruism, cooperation, predation, exploitation, Samaritan's dilemma, evolutionary game theory

JEL Codes C73 H41 D64 D74

1. Introduction

Paying taxes, recycling, reducing CO2 emission by using public rather than private transports are all examples of behaviors which yield social benefits at private costs. In all these cases, individual welfare depends on public goods and thus, on the capability of communities to discourage parasitism, as the loath against free-riders may prevent cooperation to emerge and diffuse. When societies are small and group cohesion is strong, cooperation is self-sustaining, as individuals spontaneously abide with the norms of the collectivity because of their concern for others. Even in more "dysfunctional" cases when people develop antagonistic dispositions

towards their peers, the fact of interacting with acquaintances at least guarantees the fine-tuning of moral dispositions. When rules emerge through frequent and repeated interactions, in fact, they create shared values that generate social homogeneity. In other words, regardless of the possibility of ending in a cooperative or antagonistic community, interactions in small groups rule out the risk of parasitism.

Conversely, in our large and open societies, interactions tendentially occur between groups of individuals who know little or even nothing about each other. Now, to interact with a "stranger", i.e. with an individual belonging to a different group, means to interact with someone who may use different conventions or rules of conduct and therefore adopt different moral dispositions. In this case, the risk of parasitism is far more severe and room is created for a variety of situations to occur. First and more intuitive, the lack of bonds, sympathy and concern for others may simply annihilate the possibility of cooperation. In this case, the collective provision of public goods must be enforced, either privately or publicly[1]. Second, when morality is strong, cooperation may flourish despite the absence of a sense of belonging. Third and more interestingly, when people behave according to different rules, those with higher sympathy and concern for others may end up contributing to public goods despite being surrounded by egoists who free-ride on their morality. In other words, moral heterogeneity may split the collectivity into *predators* and *preys*, or, in other words, into good *Samaritans* and *parasites*.

[1] The reciprocity motive has been found key to the endogenous enforcement of cooperation (Fehr and Fischbacher, 2002a, 2002b; Fehr and Gächter,2000a, 200b). The analysis of tax evasion proposed by Antoci et al. (2014), for instance, shows that tax evasion will prevail if taxpayers are unwilling to report evaders beside honestly paying their contributions. In a companion paper, Antoci and Zarri (2015) push this intuition further and analyze a society where strong reciprocators coexist with both unconditional cooperators and unconditional defectors. In this framework, cooperators inhibit the diffusion of righteous behaviors, as they decrease the probability that defectors get punished by reciprocators. The idea, in this case, is that the existence of cooperators as prey provides benefits to defectors as predators. The key result of the model is that large-scale cooperation cannot survive unless unconditional cooperators are driven to extinction by a novel type of "very strong" reciprocators, who reprehend both defectors—as first-order free riders—and cooperators—as second-order free-riders.

This is the issue this paper analyzes. More precisely, we study the outcome of interactions among individuals who have different degrees of morality, concern for others or altruism[2]. One of the key assumptions we use in this paper is that individuals are victim of *moral illusion*. In this framework, they are unable to identify—or learn to identify—moral and unmoral individuals and thus, to adapt their behavior according to whom they interact with. As a matter of fact, recognizing *ex-ante* the moral inclination of another individual may not be easy, no more than accepting moral disillusion. Hence, an individual with certain moral dispositions may not easily refrain from abiding with the latter, even when she interacts with someone characterized by different ethical beliefs. From this perspective, when people suffer from moral illusion they continue to behave morally even when their self-interested opponents adopt egoist behaviors. In the same vein, a selfish individual may stick to her non-moral conduct despite being exposed to examples of virtuous behavior[3].

To analyze the interactions among agents with different levels of altruism, we develop a simple evolutionary game-theoretic model where individuals from morally heterogeneous groups are randomly coupled to play a one-shot Prisoners' dilemma with other regarding preferences. The Prisoners' dilemma has been long used as a framework to analyze—both theoretically and experimentally—the evolution of cooperation in public good games[4]. By

[2] We define "morality" as a form of altruism or sympathy or concern for the well-being of others without distinguishing between these neighboring concepts. We rely on what Francis Edgeworth (1881: 102) was the first—to our knowledge—to name an "effective coefficient of sympathy". In his words «between the frozen pole of egoism and the tropical expanse of utilitarianism [there is] the position of one for whom in a calm moment his neighbor's utility compared with his own neither counts for nothing, nor `counts for one', but counts for a fraction». For a detailed account of Edgeworth's treatment of altruism and its relation to more recent literature, see Collard (1975), Rotemberg (1994) and Bester and Güth (1998).

[3] Another way of legitimizing our approach is to characterize non-reciprocal altruism as an act of charity or benevolence, which has nothing to do with reciprocity and reciprocation—on this point, see Tullberg (2004). In other words, being benevolent to a stranger is equivalent to behave altruistically without expecting anything in return. This was actually the meaning of the original parable of the Good Samaritan that one finds in the New Testament (Luke, 10: 25-37): the Samaritan helps someone without any expectation to be rewarded in return.

[4] For a theoretical example, see Antoci and Zarri (2014); for experimental treatments, see Gächter and Herrmann (2011) and Carpenter et al. (2009).

focusing on "non-reciprocal" altruism, we provide with mixed results on the emergence of cooperation, as well as on a variety of other dynamical configurations.

First, we derive conditions for which either large-scale cooperation or large-scale defection may emerge as an evolutionary stable strategy. Second, we show that under alternative parametrizations, an even more interesting social situation may occur, characterized by the stable interaction between a population of altruist cooperators and one of selfish defectors. When the degree of altruism or concern for others strikingly diverge across the interacting populations, in fact, we find that the Prisoners' Dilemma is transformed into another form of social dilemma, that James Buchanan (1975) called the *Samaritan's Dilemma*—hereafter, SD. In our framework, the SD consists in a game where those with a relatively high degree of altruism cooperate, while those with a relatively low degree of altruism defect. Buchanan refers to this scenario as "exploitation", as selfish defectors enjoy higher payoffs than altruist cooperators by free-riding on their morality. Hence, a key message of our paper is that altruists interacting with egoists have no other choice than playing either a Prisoners' or a Samaritan's dilemma. In other words, they go from a social dysfunctional situation to another.

Now, to enjoy the benefits of a public good such as, say, public health. without paying the related cost can be understood as an act of *predation*, as it involves a pure redistribution of wealth from contributors to free-riders. We deem this perspective as largely complementary to Buchanan's, and further qualify the SD as a predatory situation where *altruist Samaritans become preys and get exploited by selfish predators*. Political economists have been mainly studying predation in situations of clear asymmetry of power, normally, by considering how an elite can predate a population of oppressed—for a review see Acemoglu and Robinson (2006) and Vahabi (2010, 2011, 2015). Hence, we provide a novel view on predation by showing that predatory behaviors are not always of asymmetric matter of power but also, of *asymmetric morality*.

Does it imply that the problems it raises are more delicate? We do think so. In effect, and this is the second result we reach in this paper, we demonstrate that the move from the Prisoners' to the Samaritan's dilemma always corresponds to an improvement in the welfare of *both* players, but this improvement is *subjective*. At first sight, this may seem to suggest that exploitation has not necessarily to be avoided. The justification for this counterintuitive result, as it will appear below, is that we use the players' subjective perceptions of their payoffs rather than their objective value to estimate the outcome of interactions based on asymmetric altruism. However, we also show that there exist situations where objective social welfare and subjective well-being may diverge. Which is quite important because, and this is the third point of the paper, we also find that the exploitation equilibrium can be evolutionary stable. This means that, because of the subjective perception of an improvement, the exploitation equilibrium will last.

The remainder of the paper is organized as follows. Section reviews 2 the literature on altruism and reciprocity from an interdisciplinary perspective. Section 3 presents the model's main assumptions. Section 4 analyzes the system's dynamics, while section 5 derives and comments the welfare properties of the game. Section 6 concludes.

2. Literature review

The works of Adam Smith and David Hume provides the background for our analysis, as both argue that human beings are characterized by some form of concern for others, which they call "sympathy". In addition, they insist that sympathy towards friends and acquaintances differ from the benevolence we can feel for people living at the other end of the planet—see, e. g., Khalil (2001, 2013). Hence, our idea of unilateral or non-reciprocal altruism is consistent with this intuition.

A key question concerning the interactions among individuals with different concerns for other is that of knowing if altruists will lead egoists to cooperate and behave altruistically or rather, if they will end up in a state of predation where egoists will free-ride on their morality. From this perspective, the reference point is provided by Gary Becker's seminal article, "A Theory of Social Interaction" (1974), in which Becker demonstrates the "Rotten Kid theorem" (1974), according to which selfish people tend to behave altruistically when interacting with altruists. Or, as Robert Axelrod puts it in another well-known article, cooperation can emerge in a world of egoists if players use a specific strategy, namely Tit-for-Tat (1981, 1984)—see also Stark (1989), who reaches a similar result[5]. Under certain conditions, a "surge of altruism" (Kahana, 2005) may exist even when preferences do not change, but behaviors do. In this case, altruism may spread—see the criticism of Becker's model in Bergstrom (1989); Hirshleifer, (1977) and Tullock (1977).

By considering non-assortative and assortative interactions, Sethi and Somanathan (2001: 295) conclude that even individuals with altruistic concern «may … gain pleasure from reducing the well-being of those who are perceived to be selfish or spiteful». Similarly, Kaushik Basu (2010: 20) analyzes interactions in the context of a heterogenous population and notes that «the injection of one habitual noncooperator results in a total breakdown of cooperation. In other words, the addition of a new person who is innately non-cooperative, vitiates the atmosphere for those individuals in society who were close to the borderline» and that «[o]ne persons change of preference can cause a change in the behavior of all other persons in society, despite their preferences remaining unaltered». Thus, both Sethi and Somanathan and Basu tend to conclude that non-reciprocal altruism tends to destroy altruism.

[5] Beyond Axelrod, we must cite the literature—particularly large —on private orderings and the emergence of norms among individuals through repeated interactions. Among the important references, let us mention: Benson (1989, 1990, 1991); Bernstein (1992); Greif (1989, 1993); Greif, Milgron and Weigast (1994); Milgrom, Roberts and Weingast (1994); Stringham (2005); Stringham and Bowel (2009).

To a certain extent, this result is similar to the one put forward about martyrs and martyrdom. A large set of works were produced in the 1960s and 1970s—and developed rather independently from the previous ones—by game theorists who studied the prisoner's dilemma showing that some individuals tend to systematically cooperate even in interactions with people who systematically defect. Martyrs were not always successful in changing the behaviors of their opponents: "the real subjects tend to split into two approximately equal populations, those who exploit the martyr and those who cooperate" (Rapoport, 1975: 663; see also Rapoport, 1962). But, in the long run, the strategy of martyrdom might pay: "martyrs, while they may be unsuccessful against present oppressors, do indeed demonstrate their benevolence to observers of the martyrdom, resulting in less exploitation from these later interactants" (Braver and Rohrer, 1975: 653). Martyrs are very similar to our Samaritans.

Finally, there exist a wide literature at the crossroad of biology, economics and political science which distinguishes between "genuine" and "reciprocal" altruism[6]. Robert Trivers was one of the firsts to contribute to this research—see also Hamilton (1964), Sober and Wilson (1998) and Maynard Smith (1998). In a seminal paper (Trivers, 1971), he focused on *reciprocal altruism* and *conditional cooperation*, that is, on actions that confer a benefit to others at a cost to oneself under the expectation of a subsequent reciprocal benefit sufficient to offset the cost. He contrasted reciprocal altruism and conditional cooperation, to *indirect reciprocity* and *indirect cooperation*, which entail conferring benefits on those who have benefitted from others and receiving benefits in return. By contrast to these "intelligent ways of being selfish", strong reciprocity motives guided by social preferences may induce behaviors that are altruistic in the biologists' sense, conferring benefits to others in one's group at a cost to oneself. Hence altruism differs from reciprocity, as it is not conditioned on the type or actions

[6] For a history of how economists studied altruism and interacted with other disciplines, see Fontaine (2007a, 2007b). On the interrelation between altruism, economics and sociobiology see also Becker (1976).

of the others[7]. In this paper, we deliberately ignore the reciprocity motive to focus on the outcome of non-reciprocal altruism. Despite our results could be weakened by considering the interrelation between altruism and reciprocity—as in Bowles and Hwang (2012)—the quality of our message would remain unvaried.

We contribute to such neighboring streams of literature in several ways. First, we derive conditions for which large-scale cooperation may emerge even in the absence of strong reciprocators, despite we do not rule out the possibility that mass defection may prevail as an evolutionary stable strategy. Hence, we differ both from the optimistic literature close to Becker's theorem and from the pessimistic approach *à la* Basu or *à la* Sethi and Somanathan. Second and most important, we analyze situations where a population of altruist cooperators stability interact with one of selfish defectors. We show that non-reciprocal altruism may lead to another form of social dilemma, namely, a Samaritan's dilemma[8]. In our framework, the SD is a game where a population of egoists enjoy the benefits of a public good without contributing to the latter. Our contention is that this can be understood as an act of *predation*, as selfish defectors exploit altruist Samaritans by free-riding on their contribution. As a byproduct, we thus show that predatory behaviors may extend beyond the relationship between citizens and rulers. Indeed, predation does not need asymmetric power to occur, as it may also emerge in all situations characterized by asymmetric morality.

[7] The commonly observed rejection of positive offers in the Ultimatum Games is an example of the reciprocity motive. Experiments conducted in in the United States, Slovakia, Japan, Israel, Slovenia, Germany, Russia, Indonesia, and many other countries (Fehr and Gächter 2000b) support the importance of the reciprocity motive on the part of the responder, who accept to bear a cost and forego a positive payoff to punish the proposer for making an unfair offer.

[8] The Samaritan's Dilemma has been studied in private settings like families (Futagami, Kamada and Sato, 2004) but also in public situations like redistribution and poor relief (Wagner, 2005), medical care and national health insurance, international aid in case of natural disaster (Gibson et al., 2005; Williamson, 2010; Boone, 1996; Stone, 2008; among others) and social programs, where it has been presented as an argument against Basic Income Programs (Boettke and Martin, 2010). In all these circumstances, Samaritanism gives birth to problems such as, to name but a few, short-term benefits but long-term harm, loss of self-reliance, increase in the number of beneficiaries and decrease in the wealth of the beneficiaries, etc.

Another original aspect of our paper is that we generalize the possibility of the SD, showing that there is no need to make any specific assumptions to reach this result. For instance, we do not need to assume that the population is structured in a particular way, as in Basu (2001), or that «reciprocation can bring benefits through reputation-building», as in Roberts (1998: 428-429). In our framework *exploitation occurs in very simple and straight-forward situations, namely each and every time two individuals with different degrees of altruism interact with each other*.

Finally, we also depart from the existing literature on the SD who usually shows that "exploitation" is problematic, both *per se*—exploitation it has never been reputed for being desirable—and for its consequences—the equilibrium with exploitation is inefficient. In our paper, we show the existence of a set of parameter's value where exploitation is, perhaps counterintuitively, Pareto-improving. The intuition is straightforward: individuals with "excessive" concern for others indirectly benefit from their opponent's unilateral defection, as this provide the latter with higher well-being than bilateral cooperation. However, we wish not to emphasize this as a positive result. On the contrary, by insisting on the difference between objective social welfare and the subjective perceptions of the latter, we highlight the existence of a "dysfunctional" level of altruism whereby cooperators have higher payoffs when interacting with a defector than with another cooperator. In a game with "more rationality" where trajectories are both Nash equilibria and Pareto-optimal, this could lead to situations where the predatory equilibrium is selected by the interacting agents. Hence, we see this result as corroborating rather than weakening the idea that altruism may generate parasitism, as individuals who actually "enjoy" being exploited may create a vicious circle where virtuosity on the one side creates opportunism on the other—see the end of section 5 for further discussion.

3. The model

3.1 Assumptions

Consider a model economy populated by two groups of individuals, indexed by 1 and 2 respectively. At each moment in continuous time, there are many random encounters between member of the two populations. In each encounter, agents choose whether to cooperate (strategy C) or defect (strategy D). As it is common in the literature on cooperation, we think of this as representing a situation where a collectivity enjoys the benefits a public good and each individual must therefore choose whether to bear a cost and contribute to the latter or "play smart" and free-ride on the efforts of his peers—see e.g., Antoci and Zarri (2014). As anticipated, we rationalize the decision to free ride as an act of *predation*, as it involves the pure redistribution of a part of the wealth created by the productive to the unproductive, who, as we shall see in a moment, decide to stay so because of their level of concern for others.

In choosing their strategy, we assume that agents weight the material returns to their action for their degree of altruism. Hence, the players' moral dispositions *distort* their perceptions of the objective implications of playing a given strategy at a given state. To allow for conformism, we assume that groups are group homogenous with respect of their moral inclinations, though we allow the latter to differ across populations. In other words, altruism is group-specific and thus qualifies the individuals of each population.

In formal terms, we model altruism through utility-interdependence and specify the agents' returns as a function of their payoffs – weighted for their selfishness – and the payoffs of their opponents – weighted for their altruism. Hence, the utility of the individuals from group 1 is given by $U_1 = \alpha\pi_1 + (1-\alpha)\pi_2$, where $0 \leq \alpha \leq 1$ is the degree of altruism established in population 1, while π_1 and π_2 are the player's objective payoffs computed at each of the four states belonging to the strategy set $\{CC, CD, DC, DD\}$. Similarly, we write the utility of the

individuals from group 2 as $U_2 = \beta\pi_2 + (1-\beta)$,where $0 \leq \beta \leq 1$ is the degree of altruism established in population 2, while π_1 and π_2 have the same interpretation as above. If $\alpha > 1/2$ (resp., $\beta > 1/2$), we say that individuals from group 1 (resp., group 2) are *intrinsically* altruistic, as they deem their counterparts' well-being more important than theirs. Conversely, if $\alpha < 1/2$ (resp., $\beta < 1/2$), we say that individuals from group 1 (resp., 2) are *intrinsically* selfish, as they deem their own well-being more important than their counterparts'. In addition, if $\alpha > \beta$, we say that players from group 1 are more altruistic than their counterparts from group 2, and vice-versa, if $\alpha < \beta$.

The players'μ payoff matrix is reported in (1), Where agents from group 1 are row players and agents from group 2 are columns players.

	C	D
C	R, R	$(1-\alpha)E + \alpha T; (1-\beta)T + \beta E$
D	$(1-\alpha)T + \alpha E; (1-\beta)E + \beta T$	P, P

(1)

Following previous contributions in the evolutionary literature on cooperation and public goods provision—see, e.g., Antoci and Zarri (2014)—we assume that the objective game has the structure of a Prisoners' dilemma. Observe, however, that the introduction of other-regarding preferences allows for a variety of equilibrium configurations that are not reachable in a standard cooperation game. We shall return on this later. We refer to T as measuring the objective temptation from unilateral defection; to P as measuring the objective punishment to bilateral defection; to R as measuring the objective reward from bilateral cooperation; and to E as measuring the objective *exploitation* from unilateral cooperation. To model the idea that cooperation is socially efficient (in the sense of Pareto), we further assume that there are gains to cooperation, that is, that $2R > T + E$. Hence, objective social welfare is greater under bilateral than under unilateral cooperation.

3.2. Payoffs

We analyze the game under two alternative parametrizations. In the first, we have that $T - P > R - E$. In this case, we say that the *temptation to defect is strong*, as the difference between the objective gains from unilateral defection and the objective punishment from bilateral defection is greater than the difference between the objective reward from bilateral cooperation and the objective exploitation from unilateral cooperation. Conversely, under the assumption $T - P < R - E$, the *temptation to defect can be said to be weak*, as the difference between the objective gains from unilateral defection and the objective punishment from bilateral defection is smaller than the difference between the objective reward from bilateral cooperation and the objective exploitation from unilateral cooperation. As we shall see, the system displays different dynamic behavior depending on these assumptions.

At each moment in continuous time, we denote by $0 \leq x \leq 1$ the share of compliers in group 1 and by $0 \leq y \leq 1$ the share of compliers in group 2. Hence, $(1 - x)$ and $(1 - y)$ measures the share of defectors in groups 1 and 2 respectively. From matrix (1), we compute the expected utilities from complying and defecting for individuals of group 1, which are given, respectively, by: $U_1^C = Ry + [(1 - \alpha)E + \alpha T](1 - y)$, and $U_1^D = [(1 - \alpha)T + \alpha E]y + P(1 - y)$, so that the payoff-difference between the two strategies writes:

$$U_1^C - U_1^D = [R - (1 - \alpha)T - \alpha E]y + [(1 - \alpha)E + \alpha T - P](1 - y)$$

From which it is easy to derive the nullcline along which $U_1^C - U_1^D = 0$, whose equation writes:

$$y = \frac{P - (1 - \alpha)E - \alpha T}{P - (1 - \alpha)E - \alpha T + R - (1 - \alpha)T - \alpha E} \tag{2}$$

Similarly, we compute the expected utilities from complying and defecting for group 2, which are given, respectively, by: $U_2^C = Rx + [(1-\beta)E + \beta T](1-x)$ and $U_2^D = [(1-\beta)T + \beta E]x + P(1-x)$, so that the payoff-difference between the two strategies writes:

$$U_2^C - U_2^D = [R - (1-\beta)T - \beta E]x + [(1-\beta)E + \beta T - P](1-x)$$

From which it is easy to derive the nullcline along which $U_2^C - U_2^D = 0$, whose equation writes:

$$x = \frac{P - (1-\beta)E - \beta T}{P - (1-\beta)E - \beta T + R - (1-\beta)T - \beta E} \quad (3)$$

The next step is to describe how the system may evolve under alternative sets of parameters' values.

4. Dynamics

4.1. Equilibria and stability

We model the diffusion of cooperation in both populations via the standard replicator dynamics derived by Taylor and Jonker (1978). The replicator dynamics is a learning-by-imitation model which postulates that players are boundedly rational, they learn from each other, and they tend to adopt the strategy that performs better than the other. In this framework, relatively successful behaviors are replicated, while unsuccessful behaviors are abandoned. The idea, in our case, is that the players' moral dispositions are private information and cannot be signaled nor inferred when entering the Prisoners' dilemma. Hence, each individual initially behaves according to her personal inclination and subsequently review her strategy by best-responding to payoff difference in the past. This dynamics allows alternative codes of behavior to emerge at the group level, leading the system towards configurations where all individuals abide with the prevailing *social norm* established in their population. As we shall see, such norms are dichotomic and may either take the form "always defect when interacting with a stranger" or

"always cooperate when interacting with a stranger", depending on the players' moral dispositions and on the initial compositions of the two populations. The system's dynamics are given by:

$$\begin{cases} \dot{x} = x(1-x)(U_1^C - U_1^D) \\ \dot{y} = y(1-y)(U_2^C - U_2^D) \end{cases} \tag{4}$$

where $\dot{x}$ and $\dot{y}$ are the time derivatives of x and y respectively. Dynamics (4) is defined in the unit square $Q = [0,1]^2$. As usual with replicator dynamics, all edges of the square are invariant[9] and the four vertices (0,0), (0,1), (1,0) and (1,1) where both populations are homogenous—they are both composed of one type only—are always stationary states.

In addition, dynamics (4) may admit another stationary state—indicated as $(x^*, y^*)$with $0 < x^* < 1$ and $0 < y^* < 1$—corresponding to the intersection, when existing, of the nullclines defined by (2) and (3). In such state, all four types of players coexist. Observe that $\dot{x} = 0$ holds along the curve defined by (2) and along the edges of Q where $x = 0$ and $x = 1$, while $\dot{y}$ holds along the curve defined by (3) and along the edges where $y = 0$ and $y = 1$. Evaluating the Jacobian matrix of system (4) at each stationary point, we derive the dynamic's topological properties, which are summarized in the following Proposition[10]:

Proposition 1 – *The internal* equilibrium (x^*, y^*)*, when existing, is always a saddle. In addition:*

(i) *The stationary state where all players defect—corresponding to the corner* (0,0) *of* Q*—is attractive if* $\alpha < \frac{P-E}{T-E}$ *and* $\beta < \frac{P-E}{T-E}$*, that is, when both groups have weak other-regarding preferences – i.e., they are composed of "rather" selfish individuals.*

[9] Meaning that all trajectories starting from an initial pair $(x_0, y_0) = (1, \hat{y})$, $(x_0, y_0) = (0, \hat{y})$, $(x_0, y_0) = (\hat{x}, 0)$ and $(x_0, y_0) = (\hat{x}, 1)$ will lie on the side with $x = 1$, $x = 0$, $y = 0$ and $y = 1$ respectively, where $0 \leq \hat{x} \leq 1$ and $0 \leq \hat{y} \leq 1$.

[10] The proof is routine, so it is omitted, but it is available from the author upon request.

(ii) *The stationary state where all players cooperate—corresponding to the corner* $(1,1)$ *of Q—is attractive if* $\alpha > \frac{T-R}{T-E}$ *and* $\beta > \frac{T-R}{T-E}$*, that is, when both groups have strong other-regarding preferences – i.e., they are composed of "rather" altruistic individuals.*

(iii) *The stationary state where players from group 1 cooperate while players from group 2 defect—corresponding to the corner* $(1,0)$ *of Q—is attractive if* $\alpha > \frac{P-E}{T-E}$ *and* $\beta < \frac{T-R}{T-E}$*, that is, when other-regarding preferences are strong in group 1 but weak in group 2.*

(iv) *The stationary state where players from group 1 defect while players from group 2 cooperate—corresponding to the corner* $(0,1)$ *of Q—is attractive if* $\alpha < \frac{T-R}{T-E}$ *and* $\beta > \frac{P-E}{T-E}$*, that is, when other-regarding preferences are strong in group 2 but weak in group 1.*

The results from Proposition 1 are intuitive and predict that different codes of behavior will emerge across the interacting populations depending on the moral disposition of the two groups. In simple words, when the level of altruism in a given population is relatively high, the social norm "always cooperate when interacting with a stranger" will prevail; conversely, when the level of altruism in a given population is relatively low, the social norm "always defect when interacting with a stranger" will prevail. This generates two sets of intuitive results. When other-regarding preferences are homogenous across populations, a state of generalized cooperation—see point (i) of Proposition 1—or generalized defection—see point (ii) of Proposition 1—will eventually result. This means that a little bit of altruism is not sufficient to induce individuals to cooperate, and, complementarily, that altruism has not to be "pure" to lead towards a state of generalized cooperation. This confirms a result put forward by Stark (1989).

More original is the second set of results. Indeed, when other-regarding preferences are heterogenous across populations, the social norm "always cooperate when interacting with a stranger" establishes in a population while the social norm "always defect when interacting

with a stranger" establishes in the other. This is precisely the situation that Buchanan (1975) called a *Samaritan's dilemma*, which, as anticipated, corresponds to a scenario where benevolence backfires on moral individuals and encourages parasitism and predation from less moral ones. In other words, and this is one of the key originalities of the paper, we endogenize the origins of the Samaritan's dilemma by showing that Samaritans can do nothing to impede predators to free-ride on their contributions because of their excessive concerns for others. However, to fully comment on such result, we must first analyze the different dynamic regimes allowed for by Proposition 1. This will be done in the following section.

4.2. Dynamic regimes

From Proposition 1, it is easy to check that at most two equilibria may simultaneously attract. When a single stationary point is globally attractive, we say that the corresponding dynamic regime is monostable—see fig. 1. Conversely, when two stationary points are locally attractive, we say that the corresponding dynamic regime is bistable—see fig. 2 and 3. Hence, we have four possible states. To refer to the latter in an intuitive way, we call the situation where both players cooperate "Cooperation"—see point (i) of Proposition 1—while we call the situation where both players defect "Defection"—see point (ii) of Proposition 1. In addition, we take population 1 as our "focal group" and call the situation where individuals from group 1 cooperate but individuals from group 2 defect, "Exploitation"—see point (iii) of Proposition 1—while we call the situation where individuals from group 2 cooperate but individuals from group 1 defect "Predation"—see point (i) of Proposition 1. Needless to say, such labels should be reverted when focalizing on group 2. In what follows, we analyze further the most interesting of the dynamic regimes allowed for by Proposition 1, i.e. when the system exhibits bistable behaviors. The key implication of multiple equilibria is that "history matters", so that the system's eventual configuration does not solely depends on parameters, but also on the populations' initial composition.

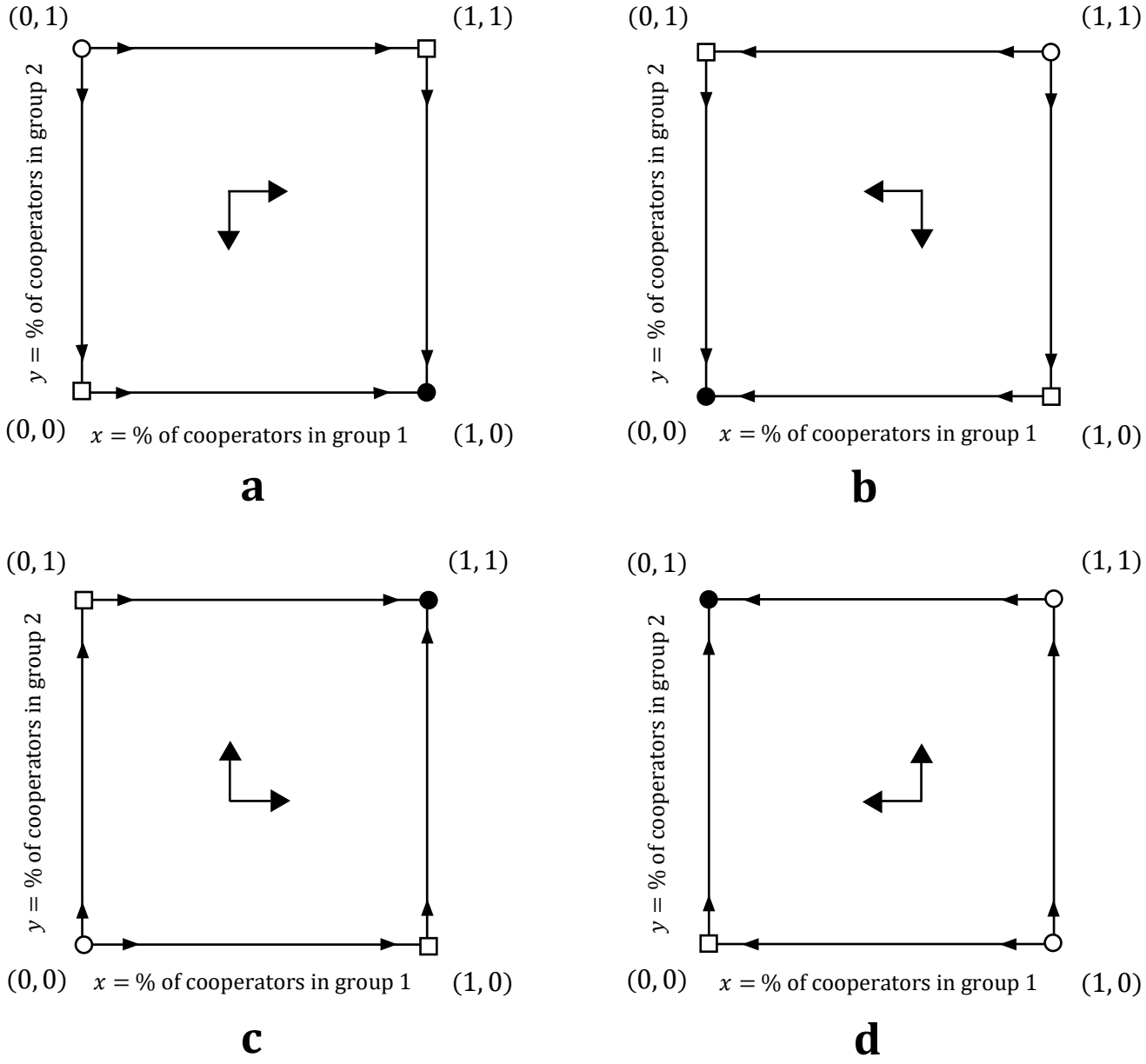

Fig. 1: Phase portrait of the four monostable regimes admitted for by dynamics (4). Filled dots represent attractors; empty dots represent repellors; empty squares represent saddle points

The learning-by-imitation mechanism of the replicator model, in this case, may lead towards socially inefficient situations, as individuals may mimic Pareto-inferior strategies for the fact of being exposed to a malfunctioning social environment. Hence, this path-dependent property may allow the emergence of dysfunctional situations characterized by both asymmetric and/or inefficient social norms.

4.2.1. The "cooperate or defect" regime

In the first bistable regime, the Cooperation and Defection equilibria simultaneously attract—see figure 2. The associated parametrization imposes two limitations on the set of parameter's values. The former concerns the level of other regarding-preferences, which must be neither too high nor too low. In formal terms, $(T-R)/(T-E) < i < (P-E)/(T-E)\,, i = \alpha, \beta$. The latter is derived as a necessary condition for intermediate regarding-preferences to exist and requires that temptation is weak. In formal terms, $T - P < R - E$. The first inequality states that cooperators have no incentive to defect when matched with another cooperator, since other-regarding preferences are relatively strong. This guarantees that in a state of generalized cooperation, no individual has an incentive do deviate and engage in predatory behaviors. The second inequality, on the other hand, states that agents will impede predation to occur, as bilateral defection provides higher payoffs than unilateral cooperation—other-regarding preferences are relatively weak. As we shall see in a moment, there exits parameterizations where an "excess" of altruism leads the way to predation, as the individuals from the group with strong concern for others cooperate despite the fact of interacting with a group of free-riders. We shall return on this later.

In the literature on cooperation, the simple choice to defect in a Prisoners' dilemma is often referred to as an implicit means of (costly) punishing defectors—see for instance, Antoci and Zarri (2014). In this regime, individuals wholesomely welcome bilateral cooperation but are unwilling to allow defectors to "get away" with their misbehavior. This can be further appreciated from the fact that the growth rate of cooperators are negatively correlated across groups, since $\partial \dot{x}/\partial y > 0$ and $\partial \dot{y}/\partial x > 0$ under the assumption $T - P < R - E$. In this framework, individuals who may be willing to cooperate with cooperating strangers may be discouraged to do so by a sufficient presence of defectors the other population. To some extent, this social mechanism resembles a sort of "tit for tat".

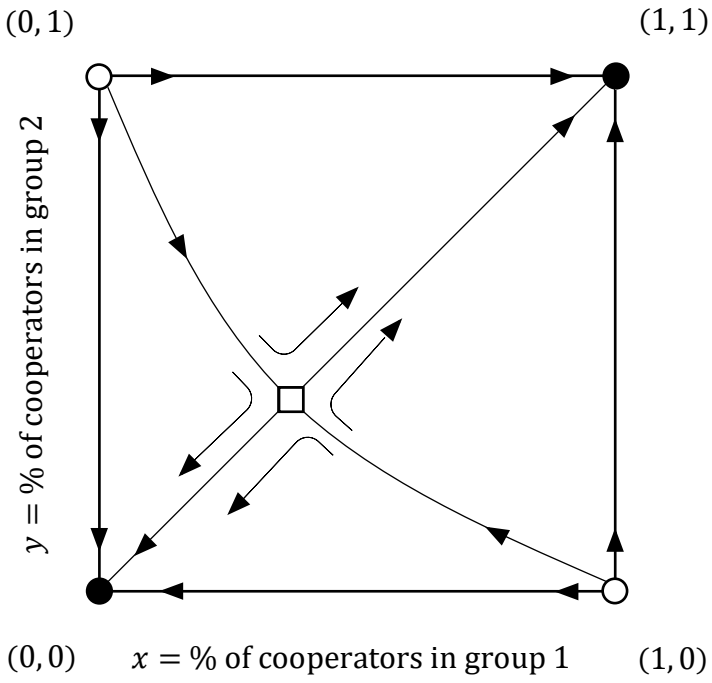

Fig. 2: Phase portraits of replicator dynamics (4) in the "Cooperate of defect" regime. Filled dots represent attractors; empty dots represent repellors; empty squares represent saddle points and the two intersecting lines are the trajectories belonging to the stable and unstable branch of the internal saddle.

The key implication of this social mechanism is that both equilibria are characterized by symmetric social norms: if the share of defectors in the economy as a whole is initially large, "always defect when interacting with a stranger" will prevail as the dominant institution: conversely, if the share of defectors in the economy as a whole is initially small, "always cooperate when interacting with a stranger" will prevail as the dominant institution.

4.2.2. The "predate or get exploited" regime

In the second bistable regime, the Exploitation and Predation equilibria simultaneously attract—see figure 3—and asymmetric social norms emerge across populations. As for the "Cooperate or defect" regime, the level of altruism in both population is intermediate, though temptation is strong in this scenario. In formal terms, In formal terms, $(P-E)/(T-E) < i < (T-R)/(T-E)$, $i = \alpha, \beta$ and $T - P > R - E$.

The key distinction between this and the previous regime is that individuals now have an incentive to predate, since since $R < (1-a)T + aE$ – other-regarding preferences are relatively weak. However, the incentives to "punish" unilateral defections by not cooperating are absent in this regime, as unilateral cooperation provides higher payoffs than bilateral cooperation since $(1-a)E + aT > P$ – other regarding preferences are relatively strong. This can be further appreciated from the fact that the growth rates of cooperators are negatively correlated across groups, since $\partial\dot{x}/\partial y < 0$ and $\partial\dot{y}/\partial x < 0$ under the assumption $T - P > R - E$. In this framework, other-regarding preferences generate, so to say, "unwholesome" situations, where cooperators end up being exploited when the rate of defectors in the other population is sufficiently high and defectors end up predating when the rate of defectors in the other populations is sufficiently low.

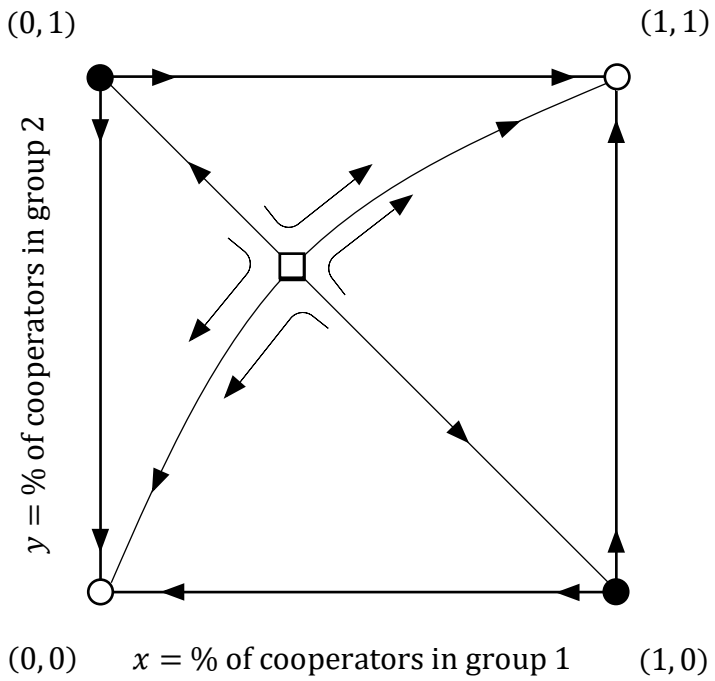

Fig. 3: Phase portraits of replicator dynamics (4) in the "Exploit or get exploited" regime. Filled dots represent attractors; empty dots represent repellors; empty squares represent saddle points and the two intersecting lines are the trajectories belonging to the stable and unstable branch of the internal saddle.

The socially dysfunctional character of this regime becomes clearer when we look at the dynamic behavior of the system in the neighborhood of its unstable states—see fig. 3. Imagine a situation where the economy is temporarily settled at the Defection equilibrium. Starting from this state, even the tiniest mutation in the population's composition is sufficient to transport the system towards either of the two "dysfunctional" attractors. More precisely, if the rate of cooperators in group 2 (resp., group 1) exogenously increases, all defectors in group 1 (resp., group 2) will have an incentive to stick to their strategy and predate their counterparts in group 2 (resp., group 1), so that the system will eventually snowball towards the Predation (resp., Exploitation) equilibrium.

Complementary remarks can also be drawn from the opposite situation. Consider a scenario where all players across populations initially cooperate. As before, even the tiniest mutation in the population's composition is sufficient to transport the system towards either of the two "dysfunctional" attractors. The learning-by-imitation mechanism behind the replicator dynamics works differently in this case, as the relatively low level of altruism in both groups pushes individuals to predate their opponents. Unsurprisingly, temptation is strong in this regime. The fact that the agents in the other group do not retaliate against such misconducts is due the lack of incentives. Hence, when exogenous variations in either of the two populations are not coupled by counterbalancing mutations in the other, a single negative shock may suffice to transport the system from an initial state of unstable cooperation to a final state of stable exploitation. To qualify further the quality of these regimes we need to inquire into the welfare properties of the system.

5. Welfare

In this section, we analyze the welfare properties of the game from the viewpoint of the individuals in group 1—all results can be extended to population 2 by substituting α with β in

the following discussion. To measure welfare, we compute the agents' average payoff in the four attractive states admitted for by Proposition 1, which are given, respectively, by: $U_{(0,0)} = P$; $U_{(0,1)} = (1-\alpha)T + \alpha E$; $U_{(1,1)} = R$ and $U_{(1,0)} = (1-\alpha)E + \alpha T$. According to Pareto's classic definition, a given state is comparatively less efficient than another when moving from the former to the latter the utility of at least one individual increases. Accordingly, a given equilibrium is Pareto-optimal when moving from the former to any other state the utility of all individuals decreases. In what follows, we shall refer to a stable equilibrium which is Pareto-dominated by at least another state of the system as a *poverty trap* (Carrera, 2018). Observe that an equilibrium need not be attractive to be more efficient than another. Hence, we formulate the following Proposition[11]:

Proposition 2 – *When the Defection equilibrium is attractive, it is always a poverty trap. In addition:*

(i) *When the Cooperation equilibrium is attractive, it is always more efficient than the Temptation equilibrium. However, it is less efficient than the Exploitation equilibrium if* $\alpha > \frac{R-E}{T-E}$.

(ii) *When the Exploitation equilibrium is attractive, it is less efficient than the Temptation equilibrium if* $\alpha < \frac{1}{2}$ *and it is less efficient than the Cooperation equilibrium if* $\alpha < \frac{R-E}{T-E}$.

(iii) *When the Temptation equilibrium is attractive, it is always more efficient than both the Cooperation and the Exploitation equilibria.*

A first remark from Proposition 2 is that, when the Exploitation equilibrium is attractive, it is also more efficient for the agent who gets exploited—that is, for the Samaritan—than the Punishment equilibrium. In this case, sufficiently altruistic individuals prefer "being exploited" in a Samaritan's dilemma than remaining trapped in a state of mutual defection. The reason is that the objective utility loss they experience when moving from a Prisoners' to a Samaritan's

[11] The proof is trivial so it is omitted, but it is available from the authors upon request.

dilemma is more than compensated by the utility gain obtained by the predator they interact with. However, recalling that the players' subjective perceptions and objective social welfare may diverge, this may allow for situations where the individuals' subjective perception of the Exploitation state is inconsistent with its actual social efficiency. To prove this point, recall that the Punishment equilibrium is *objectively* more efficient than the Exploitation equilibrium if if $2P > R + E$. From a subjective viewpoint, however, we know that the Exploitation equilibrium is attractive if $(1 - \alpha)E + \alpha T > P$. Putting together these two conditions, we see that objective social welfare and the Samaritan's perception of the latter are misaligned if $(T + E) \,/\, 2 < P < (1 - \alpha)E + \alpha T$. Solving this expression for α, we see that a necessary condition for this to hold is that individuals are intrinsically altruists, which requires that $\alpha > 1/2$.

The second insightful remark that can be drawn from Proposition 2 is that, in the "predate or get exploited" regime, both equilibria are always efficient for the population who exploits and always inefficient for the population who gets exploited. Under the bistability requirement $(P - E)/(T - E) < \alpha < (T - R)/(T - E)$, in fact, individuals are always intrinsically selfish, since $(T - R)/(T - E) < 1/2$ is always satisfied under the assumption $2R > T + E$. Combining this with the Pareto-ranking of the Exploitation and Punishment equilibria derived in the above, we see that the players' altruism creates a paradoxical situation in this regime. Indeed, despite individuals would prefer predating instead of behaving as lone cooperators, their other-regarding preferences are strong enough to prevent them from "punishing" defectors by start defecting themselves. Hence, once they are trapped in a Samaritan's dilemma, they cannot escape the latter, despite its Pareto-inefficiency.

Conversely, the welfare properties of the "cooperate or defect" regime are, so to say, more "coordinated", as the Defection equilibrium is always a poverty trap for both populations. Observe, however, that even in this scenario, the Cooperation equilibrium may be Pareto-dominated by the Exploitation equilibrium. In particular, this occurs when the players' other-

regarding preferences is above a critical threshold that separates what we call "functional" from "dysfunctional" altruism. Formally, this cutoff is given by $(R - E)/(T - E)$. Similarly to the situation described in the above concerning the move from a Prisoners' to a Samaritan's dilemma, such threshold always creates a situation of misalignment between objective social welfare and individual well-being. In this scenario, "dysfunctional" Samaritans prefer to feel as "lone" cooperators rather than to live in a state of generalized cooperation, as they are better off in the Exploitation than in the Cooperation equilibrium. From an objective perspective, however, the latter is always more efficient than any other state of the system. In this case, although the players are still facing a Samaritans' dilemma, the situation may seem less problematic than the literature usually argues, as exploitation is not utility-depressing, but rather, utility-enhancing.

The possible situations where exploitation is utility-enhancing from the viewpoint of she who gets "exploited" may not be as unrealistic as it looks at first sight. Indeed, it may be used to comment on complementary behavioral motives that, despite being left outside the model, may have a role in the emergence of such "dysfunctional" altruism. Without overstressing such line of reasoning—which may be one in a series of possible others—excessive concerns for others may have negative implications for psychological well-being. Thus, as claimed by Marciano (2020), the situation that corresponds to a Samaritan's dilemma is typically a form of masochism. Indeed, the Samaritan cares sufficiently for the recipient to sacrifice her utility. The utility she gets from sacrificing her pleasure is compensated by the pleasure obtained by the "parasite"—see Khalil (2001, 2004) and Nida-Rümelin (1991). Thus, in contrast to what Buchanan and the literature after him argued, asymmetrical altruism may yield subjective utility gains to both parties, even when it leads to exploitation—see also Singh (1995). As anticipated, however, this may create scenarios where the Samaritan's perception of the

objective situation is distorted by her altruism, perceiving a given state as more efficient than it actually is.

In the same vein—since the logic behind the behaviors the same—when cooperators are so unselfish that they actually enjoy "being exploited" by defecting strangers, more or less conscious feelings of "moral elitism" may complement their concerns for others, as they may enjoy idealizing themselves as virtuous individuals capable of moral conducts in an otherwise unmoral society. From these two perspectives, it is clear that Samaritans need parasites. Masochism or elitism leads to a form of sadism, in the sense that Samaritans may «welcome tragedies» (Khalil, 2004: 102). Indeed, as Khalil (ibid.) puts it, «[t]he altruist qua masochist may not abhor natural disasters befalling others. While such an altruist may refrain from expediting disasters, he would celebrate the opportunities such disasters afford him». Or, in other words, «Becker's model entails that altruists should feel joyful over the prospect of the miseries of others because such miseries occasion for them the opportunity to be aroused» (ivi: 431). This kind of altruism might not be encouraged. Hence, the claim put forward by Antoci and Zarri (2014) that punishing unconditional cooperators who act as second-order free riders is vital for the promotion of large-scale cooperation has, so to speak, a "nudging" implication, as it may discourage the formation of what we call "dysfunctional" altruism.

A further implication of allowing for misalignments between objective and subjective returns is that individuals may avoid engaging in socially desirable actions because of their "excessive" altruism. As recalled by Antoci and Zarri (2014), in fact, the simple choice to defect in a Prisoners' dilemma can be seen as an implicit means of (costly) punishing defectors and, more generally, of discouraging parasitism. Imagine a father who is uncapable of reprehending his offspring because of his excessive concern for the well-being of the latter. In this case, our model predicts that the genitor's excess of "altruism" will end up encouraging his offspring's misbehavior. What is left outside the model is that the long-run implications of encouraging

misconducts of this sort may have negative repercussion not only on the community but on defectors themselves. Hence, our model supports previous findings on the "dark side" of altruism which postulate that excessive concerns for others may, perhaps counterintuitively, undermine rather than promote large-scale cooperation—see Bowles and Hwang (2012) and Antoci and Zarri (2014).

6. Conclusion

How do we interact with strangers? Or, in other words, how do we interact with people belonging to a different group than ours? Do we cooperate with them as much as we do with the people from our group? In this paper, we try to answer this set of questions, which are all the more important in a world which is increasingly fragmented into subgroups. To do so, we develop a simple evolutionary game theoretic model where group-membership is characterized by an idiosyncratic degree of altruism or concern for the welfare of others. In this framework, the outcome of interactions between individuals who—in all likelihood—have different moral inclinations allows for a variety of dynamic configurations. Among the latter, we dedicate particular attention to the interactions occurring between an altruist and an egoist, that is, between a Samaritan and a non-Samaritan. With this respect, we endogenize the origin of the Samaritan's dilemma by showing that this occurs in simple situations where the strangers degrees of altruism are sufficiently different from one another. In this case, she who has the highest degree of concern for others is exploited, that is, she cooperates while her opponent other defects.

More precisely, we show that generalized defection emerges as an evolutionary stable strategy when the degree of altruism of both populations is below a critical threshold, suggesting that altruism does not always lead to cooperation. Conversely, when the degree of altruism of both populations is above that critical threshold, a state of generalized cooperation

occurs even in the absence of strong reciprocators. We thus show that the prediction whereby cooperation cannot flourish in the absence of reciprocity—see e.g. Antoci et al. (2014); Antoci and Zarri (2105)—may be dropped if individuals develop "coordinated" levels of concern of others, or, less strongly, that bilateral altruism could mitigate the need of reciprocity as an endogenous enforcement mechanism. On the other hand, as recalled by Bowles and Hwang (2012), altruism could also weaken the reciprocity motives, hence impeding rather than promoting cooperation. Whether the former or the latter of these mechanism will prevail, is largely an empirical question.

Finally and more importantly, when populations develop diverging levels of other-regarding preferences, our model predicts that the altruistic individuals will stably cooperate while egoists defect. Hence, we show that the Samaritan's dilemma can be evolutionary stable. Indeed, "there is no obvious escape from this dilemma" (Lee, 1987: 162) because "[a]ctions that are motivated by feelings of compassion are difficult to resist even if the long-run effects are known to be detrimental to those who are the object of our compassion". This is the first result of this paper. Observe that we reach the latter without making any assumption about the structure of the population, or the probability to meet an egoist, or the income level of the people involved in the interaction. In addition, and this is the second result of the paper, we show the existence of a set of parameters' value for which exploitation is utility-enhancing from the subjective viewpoint of the Samaritans, despite this may be at odds with objective social welfare.

The key limitation of our model is that we do not consider how the players' payoffs may evolve in response to the respective behaviors of the players involved. In other words, payoffs in our model remain exogenous. Even under this restrictive assumption, we believe that these results are of interest for an analysis of private orderings and spontaneous orders. If our results are valid, morality if asymmetric and non-reciprocal can be an obstacle for social cooperation.

Morality does not remove social dilemmas but simply change their nature. This therefore may create problems for interactions in heterogenous populations that become risky and costly. Codes and constitutions, "rules of the game" are needed—see also Leeson and Skarbek 2010; Skarbek, 2016. Here, we add the claim that these rules are aimed at preventing predation, even more so when this is paradoxically welcomed by the victims of predation, who thus do not engage in actions which may be functional for the society as a whole, like, for instance, that of punishing free-riders. This was also the claim made by Buchanan in *The Limits of Liberty* (1975b).

References

Acemoglu, Daron and James A. Robinson. (2006). *The Economic Origins of Dictatorship and Democracy*. Cambridge University Press, Cambridge.

Antoci, Angelo and Lucca Zarri. (2015). Punish and perish? *Rationality and Society*, 27(2): 195-223.

Antoci, Angelo, Paolo Russu and Lucca Zarri. (2014). Tax evasion in a behaviorally heterogeneous society: An evolutionary analysis. *Economic Modelling*, 42: 106–115.

Axelrod, Robert. (1981). The emergence of cooperation among Egoists. *American Political Science Review*, 75 (2): 306-318.

Axelrod, Robert. (1984). *The evolution of cooperation*, New York, Basic Books.

Basu, Kaushik. (2010). The moral basis of prosperity and oppression: altruism, other-regarding behaviour and identity. *Economics and Philosophy*, 26 (2): 189-216.

Becker, Gary S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6): 1063-1093.

Becker, Gary S. (1976). Altruism, egoism, and genetic fitness: economics and sociobiology. *Journal of Economic Literature*, 14: 817-826.

Benson, Bruce L. (1989). Enforcement of Private Property Rights in Primitive Societies: Law Without Government. *Journal of Libertarian Studies*, 9(1): 1-26.

Benson, Bruce L. (1990). Customary Law with Private Means of Resolving Disputes and Dispensing Justice: A Description of a Modern System of Law and Order Without State Coercion. *Journal of Libertarian Studies*, 9(2): 25-42.

Benson, Bruce L. (1991). An Evolutionary Contractarian View of Primitive Law: The Institutions and Incentives Arising Under Customary American Indian Law. *Review of Austrian Economics*, 5(1): 41-65.

Bergstrom, Theodore C. (1989). A Fresh Look at the Rotten Kid Theorem – and Other Household Mysteries. *Journal of Political Economy*, 97 (5). 1138-1159.

Bernstein, Lisa. (1992). Opting out of the legal system: Extralegal contractual relations in the diamond industry. *Journal of Legal Studies*, 21(1). 115-157.

Bester, Helmut and Werner Güth. (1998). Is Altruism evolutionary stable? *Journal of Economic Behavior and Organization*, 34 (2): 193-209.

Boettke, Peter and Adam Martin. (2010). Exchange, production, and Samaritan dilemmas, mimeo, http://mpra.ub.uni-muenchen.de/33199/

Boone, P. (1996). Politics and the effectiveness of foreign aid. *European Economic Review*, 40: 289–329.

Braver, Sanford L. and Van Rohrer. (1975). When Martyrdom Pays: The Effects of Information concerning the Opponents' past Game Behavior. *The Journal of Conflict Resolution*, 19(4): 652-662.

Braver, Sanford L. and Van Rohrer. (1975). When Martyrdom Pays: The Effects of Information concerning the Opponents' past Game Behavior. *Journal of Conflict Resolution*, 19 (4): 652-662.

Buchanan, James M. (1975a). The Samaritan's dilemma, in E.S. Phelps (ed.), *Altruism, morality and economic theory*, New-York, Sage Foundation: 71-85.

Buchanan, James M. (1975b). *The Limits of Liberty. Between Anarchy and Leviathan*, University of Chicago Press.

Carpenter, Jeffrey, Samuel Bowles, Herbert Gintis, and Sung-Ha Wand. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior and Organization*, 71(2): 221–232.

Carrera, Edgar J. Sanchez. (2019). Evolutionary dynamics of poverty traps. *Journal of Evolutionary Economics*, 29(9): 1-20.

Collard, David. (1975). Edgeworth's Propositions on Altruism, *Economic Journal*, 85 (338): 355-360.

Edgeworth, Francis Y. (1881). *Mathematical Psychics. An Essay on the Application of Mathematics to the Moral Science*, London, Kegan Paul.

Fehr, Ernst and Simon Gächter. (2000a). Cooperation and punishment. *American Economic Review*, 90 (4): 980–994.

Fehr, Ernst and Simon Gächter. (2000b). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, 14 (3): 159-181.

Fehr, Ernst and Urs Fischbacher. (2002a). Why Social Preferences Matter. The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives. *Economic Journal*, 112 (478): C1-C33.

Fehr, Ernst and Urs Fischbacher. (2002b). Altruistic Punishment in Humans. *Nature*, 415: 137–40.

Fontaine, Philippe. (2007). From Philanthropy to Altruism: Incorporating Unselfish Behavior into Economics, 1961-1975. *History of Political Economy*, 39 (1): 1-46.

Futagami, Ritsuko, Kimiyoshi Kamada and Takashi Sato. (2004). Government Transfers and the Samaritan's Dilemma in the Family. *Public Choice*, 118: 77-86.

Gächter, Simon and Benedikt Herrmann. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review,* 55(2): 193–210.

Gibson, Clark C., Krister Andersson, Elinor Ostrom, and Sujai Shivakumar. (2005). *The Samaritan's Dilemma. The Political Economy of Development Aid*, Oxford University Press.

Greif, Avner, Paul Milgrom, and Barry Weingast. (1994). Coordination, Communication, and Enforcement: The Case of the Merchant Guild. *Journal of Political Economy* 102 (4): 745-776.

Greif, Avner. (1989). Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History*, XLIX: 857–882.

Hamilton, W. D. (1964) The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, 7, 1–32.

Hirshleifer, Jack. (1977). Shakespeare Versus Becker on Altruism: The Importance of Having the Last Word. *Journal of Economic Literature*, 15 (2): 500-502.

Hwang, Sung-Ha and Samuel Bowles. (2012) Is altruism bad for cooperation? *Journal of Economic Behavior and Organization*, 83(3): 330–341.

Jonker, Leo B. and Peter D. Taylor. (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Bioscience*, 40: 145-156.

Kahana, Nava. (2005) On the Surge of Altruism, *Journal of Population Economics* 18 (2): 261-266.

Khalil, Elias. (2001). Adam Smith and Three Theories of Altruism, *Recherches Economiques de Louvain*, 67 (4): 421-435.

Khalil, Elias. (2004). What is altruism? *Journal of Economic Psychology*, 25: 97-123.

Khalil, Elias. (2013). What determines the boundary of civil society? Hume, Smith and the justification of European exploitation of non-Europeans. *Theoria*, 60(134): 26-49.

Lee, Dwight R. (1987). The Tradeoff between Equality and Efficiency: Short-Run Politics and Long-Run Realities, *Public Choice*, 53 (2): 149-165.

Leeson, Peter and David Skarbek. (2010). Criminal constitutions. *Global Crime*, 11(3): 279–298.

Lindbeck, Asar and Jörgen Weibull. (1988). Altruism and efficiency: the economics of fait accompli. *Journal of Political Economy* 96: 1165–82

Marciano, Alain. (2020). Sado-masochistic altruism in a Samaritan's dilemma, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2543073

Maynard Smith, Jhon. (1998). The origin of altruism. *Nature*, 393: 639–640.

Nida-Rümelin, Julian. (1991). Practical reason or metapreferences? An undogmatic defense of Kantian morality. *Theory and Decision*, 30: 133-162.

Rapoport, Anatol. (1962). Formal Games as Probing Tools for Investigating Behavior Motivated by Trust and Suspicion. *Journal of Conflict Resolution*, 7 (3): 570-579.

Rapoport, Anatol. (1975). Comments on "When Martyrdom Pays". *Journal of Conflict Resolution*, 19 (4): 663-664.

Roberts, Gilbert. (1998). Competitive Altruism: From Reciprocity to the Handicap Principle. *Biological Sciences*, 265 (1394): 427-431.

Rotemberg Julio J. (1994). Human Relations in the Workplace. *Journal of Political Economy*, 102 (4): 684-717.

Sethi, Rajiv and Eswaran Somanathan. (2001). Preference Evolution and Reciprocity. *Journal of Economic Theory*, 97 (2): 273-297.

Singh, Nirvikar. (1995). Unilateral altruism may be beneficial: A game-theoretic illustration, *Economics Letters*, 47 (3-4), 275-281.

Skarbek, Emily. (2016). Aid, ethics, and the Samaritan's dilemma: strategic courage in constitutional entrepreneurship. *Journal of Institutional Economics*, 12 (2): 371-393.

Sober, Elliot and David Sloan Wilson. (1998). *Unto others. The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.

Stark, Oded. (1989). Altruism and the quality of life. *American Economic Review*, 79: 86-90.

Stone, Deborah. (2008). *The Samaritan's Dilemma: Should Government Help Your Neighbor?* Nation Books.

Stringham, Edward P. (ed.). (2005). *Anarchy, state, and public choice*. Cheltenham: Edward Elgar.

Stringham, Edward P. and B. Powel. (2009). Public choice and the economic analysis of anarchy: A survey. *Public Choice*, 140: 503-538.

Trivers, Robert L. (1971) The Evolution of Reciprocal Altruism. *Quarterly Review of Biology* 46: 35–57.

Tullberg, Jan. (2004). On Indirect Reciprocity the Distinction Between Reciprocity and Altruism, and a Comment on Suicide Terrorism. *American Journal of Economics and Sociology*, 63 (5): 1193-1212.

Tullock, Gordon. (1977). Economics and Sociobiology: A Comment. *Journal of Economic Literature*, 15 (2): 502-506.

Vahabi, Mehrdad. (2010). Integrating social conflict into economic theory, *Cambridge Journal of Economics*, 34 (4): 687-708.

Vahabi, Mehrdad. (2011). Appropriation, violent enforcement, and transaction costs: a critical survey, *Public Choice*, 147 (1/2): 227-253.

Vahabi, Mehrdad. (2015). *The Political Economy of Predation: Manhunting and the Economics of Escape*. Cambridge: Cambridge University Press.

Wagner Richard E. (2005) Redistribution, Poor Relief, and the Welfare State. In: Backhaus J.G., Wagner R.E. (eds) *Handbook of Public Finance*. Springer, Boston, MA.

Williamson, Claudia. (2010). Exploring the failure of foreign aid: The role of incentives and information. *Review of Austrian Economics*, 23: 17-33.