Happiness, Morality, and Game Theory

Luca Zarri[1]

Department of Economics
University of Verona

December 2006

Keywords: Non-cooperative Games; Happiness; Morality.

JEL Classification: B41; C72; C91; D64.

* I would like to thank Luigino Bruni, Pier Luigi Porta and participants in the International Conference on "The Paradoxes of Happiness in Economics", University of Milano-Bicocca, where an earlier version of this article was presented. I am also indebted to Robert Sugden for his invaluable and constant suggestions throughout this work as well as to Shaun Hargreaves Heap, Piergiovanna Natale, Paolo Vanin, Stefano Zamagni, seminar audience at UEA and participants in the Workshop on "Social Preferences and Happiness", University of Verona, for helpful comments and insights. The usual disclaimers apply.
[1] Email: luca.zarri@univr.it

1. Introduction

As far as contemporary economically advanced societies are concerned, it would be hardly deniable that people's search for happiness is significantly affected not only by the satisfaction of material needs, but also by several non-material sources such as psychological and social factors, as well as by the pursuit of complex, morally-charged goals, as a growing body of experimental and empirical contributions tends to confirm (see e.g. Easterlin, 2001; Fehr and Gächter, 2002; Rabin, 2002). Recent evidence suggests that money is less and less able to buy happiness and, in this light, Rabin (1993) correctly remarks that "Welfare economics should be concerned not only with the efficient allocation of material goods, but also with designing institutions such that people are happy about the way they interact with others". These two types of objectives (i.e. material and non-material ones) seem to interplay in complex ways, as, for instance, it is often the case that the pursuit of non-material ends such as the search for social prestige or freedom of choice crucially passes through the attainment of monetary gains. As an example of this, we may think of a status-seeking agents deciding to buy a luxury car or an expensive yacht in order to more effectively signal a given status level (regardless of its reflecting his actual social position or not): in his view, status acts as a source of (positional) utility directly provided by the (instrumental) relationship established with other subjects belonging to his 'reference group'. It seems clear, then, that, insofar as we aim at getting significant insights over the often paradoxical meanings of a multifaceted notion such as 'happiness' within advanced economic systems, the above recalled interplays need to be seriously taken into account.

Within such a complex framework, the specific aim of this essay is to provide a contribution to the understanding of the aforementioned relationship (between material and non-material determinants of individual happiness) in the context of non-cooperative game theory. The point is that while both theoretical and empirical studies on happiness have been rapidly growing in the last years, we still lack researches focusing on the attempt to reconcile happiness with game theory, i.e. to analyze happiness within strategic interaction scenarios. As Section 2 will show, exploring the connection between happiness and non-cooperative game theory is crucial as, in strategic interaction situations, players' subjective conception of happiness alters their evaluation of each possible outcome. One of the major purposes of this methodological work is to shed light on the primitive concepts constituting

two-player, simultaneous-move non-cooperative games in order to properly account for the crucial interplays taking place between (suitably defined) 'preferences' and *moral principles*. In order to do so, the following issues will be specifically addressed: what kind of difficulties arise when a relevant, non-material source of individual happiness such as morality is included into the *pay-offs* of the game? Can we incorporate *any* moral principles within individual pay-offs?

The structure of the remainder of the essay is the following. In Section 2 we briefly recall some of the main stages in the history of utility and happiness in economics, with a special focus on the relevance of the so called 'Pareto turn' and on the most significant features of the 'revealed preference' approach. Section 3 shows that non-cooperative game theory can deal with *some* moral principles (the ones we label as 'preferential' moral principles) through a proper respecification of individual pay-offs (shifting from standard to 'extended' pay-offs). In Section 4, however, we make clear that 'non-preferential' moral principles, such as Kantian principle of universalizability, cannot be satisfactorily modeled by simply respecifying players' pay-offs: with regard to this set of moral principles, we suggest to take a step forward, as the very nature of principles of morality other than utility maximization calls for non-utilitarian solution concepts. In Section 5, we shed light on what we term the 'as if paradoxes of happiness in economics', i.e. on the paradoxical implication according to which, in some social settings, even utility-maximizers get better results by acting 'as if' they were driven by a non-utilitarian moral principle. Section 6 concludes.

2. Rational Choices and Revealed Preferences in Non-cooperative Games

Before directly entering into the major theoretical issues addressed in this work, a preliminary methodological clarification is in order. In dealing with the themes specified above, we will often have recourse to what is probably the most well-known among two-person, non-cooperative games: the Prisoner's Dilemma (PD)[2]. However, we maintain that most of the considerations developed in Sections 2, 3, 4 and 5 hold also insofar as we refer to other relevant strategic interaction scenarios

[2] Strictly speaking, several real-life situations where collective action problems are likely to arise would call for the generalized version of the PD, i.e. the *N-person Prisoner's Dilemma* (or *Social Dilemma*). However, for simplicity, we will keep on exclusively referring to the classic one-shot PD setting, though we believe that most of the qualitative considerations developed here extend naturally to larger and more complex social environments.

that lend themselves to be modeled in game-theoretic terms. The main reason for choosing the PD as the reference social structure is twofold. First, many of the game theorists, economists and philosophers quoted and critically commented here with regard to the methodological issues at stake (such as Ken Binmore, Amartya Sen, Elizabeth Anderson and Robert Sugden) have often clarified their positions by directly referring to this game. In this light, shifting from one author to another by always considering the same game structure should help in expositional terms. Second, the PD can be seen as the metaphor *par excellence* of social situations where (a) individualistic rationality fails and (b) even utility-maximizing agents would reach better results by acting on principles of action other than Nash behavior (or, at least, by acting *as if* they were driven by 'proper' non-utilitarian principles; see on this Section 5). Therefore, such a game structure will play a critical role in making clear the main implications drawn in the present contribution with regard to a characterization of happiness within strategic interaction scenarios[3].

According to the 'revealed preference' approach, individual preferences are to be interpreted in terms of choice, as choices 'reflect' preferences. As Sen (1982) observes: "Preference here is simply *defined* as the binary relation underlying consistent choice. In this case 'counter-preferential' choice is not empirically different, but simply impossible. *Non*-preferential choice is, of course, possible, since the choices may lack the consistency needed for identifying a binary relation of preference, but obviously it cannot be the case that such an identified preference relation exists and the choices are 'counter' to it". In this light, we are already able to draw a simple, direct implication from this characterization of preferences: choice, not preference, ends up being the basic, 'salient' concept of any theory grounded on such definition of preferences. Savage's formal theory seems to prove this as "although Savage's axioms are formulated in terms of the concept of preference, it seems that he

[3] At this stage, though this will appear to many as an obvious remark, it is worth pointing out that, while hereafter and throughout the paper we will take for granted that in a PD the Pareto-efficient outcome of mutual cooperation is individually and socially desirable, we are not claiming that, with regard to real-life situations having the PD structure, the same proposition holds from the point of view of society as a whole. Such a broader conclusion would hold only if (i) either the set of players coincided with the set of all citizens or (ii) interaction within this setting brought about beneficial effects for society as a whole (as it is the case when, say, the voluntary provision of a public good such as education, environment preservation or health is concerned). See on this Sacco and Zarri (2002).

regards *choice* as *the more fundamental concept*: the idea is to construct a theory of rational choice, not of rational preferences" (Sugden, 1991; italics added).

However, it is important to point out that the reasons behind the greater salience of 'choice' with respect to 'preference' are not purely theoretical, but also historical. In this regard, the decisive moment in the history of economic thought coincides with the so called 'Pareto turn' (or 'indifference-curve' or 'ordinalism revolution'), the event that starts driving a wedge between neoclassical thinkers and contemporary scholars as to the interpretation of 'utility'. The early neoclassical economists tended to explain preferences by postulating a *one-dimensional* scale of inner psychological experience (referred to in terms of 'happiness' or 'pleasure'), but the difficulties arising from the search for such measure pushed economists to gradually adopt a revealed preference approach to utility, where "preferences are whatever dispositions lie behind observed choices, and the formal properties of preference which guarantee an ordering – completeness, reflexivity and transitivity – are postulated *as properties of rationality*" (Sugden, 2001). Sugden (2002) maintains that, after the Paretian turn, the concept of the utility function has been retained, but has been re-interpreted as a representation of individual preferences, which in turn are seen as whatever the agent takes to be choice-relevant reasons. Ng (1997) correctly observes that the main motives justifying such methodological revolution – related to the attempt to make economic analysis based on more objective grounds – are quite clear and sound, but he also adds that all this has been carried to an excess, preventing economics from successfully dealing with several important issues. He further interestingly asserts that psychology went through a similar process, due to the Watson-Skinner behaviorist revolution, recently resulting in the well-known cognitive turn.

The above argument clearly indicates that the historical process briefly described, while making choice the central concept of economic theory, has started assigning a less and less relevant role to the classical notion of 'utility'[4], certainly the conceptual category which turned out to be more closely related to the idea of

[4] In his pioneering contribution to the theory of 'revealed preferences', Samuelson (1938) argued that his aim was "to develop the theory of consumer's behaviour freed from any vestigial traces of the utility concept" (quoted in Sen, 1973). With reference to this approach, Little (1949) asserted that "the new formulation is scientifically more respectable [since] if an individual's behaviour is consistent, then it must be possible to explain that behaviour without reference to anything other than behaviour" (quoted in Sen, 1973).

'happiness'. As a consequence, happiness itself has never played thereafter a significant role in economics: therefore, if today we wish to re-descover the importance of such a concept and to incorporate it into the formal structure of economic theory – in the light of a rather stimulating wave of empirical and experimental studies centered either directly on happiness or on specific components of it – we need to understand whether (re)introducing happiness in our theory is compatible with the maintenance of a 'revealed preference' approach. In this regard, from the point of view of the historical evolution of ideas on the theme, it seems to be the case that bringing happiness back into economics will take the form of a sort of theoretical counter-revolution, with respect to the Pareto turn recalled above.

According to Sen and Williams (1982), however, the whole picture is even more complex and blurred, as the two different interpretations of preference (i.e. the pre-Paretian, Benthamite perspective and the post-Paretian, choice-centered view previously mentioned) still ambiguously coexist within contemporary economic theory. Further, they observe that "The ambiguity of the term 'preference' facilitates this dual picture of utility, since linguistic convention seems to permit the treatment of 'preferring' as choosing as well as taking what a person (really) 'prefers' as what would make him better off". Sen (1994) expresses a similar view by claiming that both a *choice-salience* interpretation and a *well-being* interpretation of preference are present in contemporary economics. In this regard, we may add that Kahneman et al.'s (1997) well-known distinction between *decision utility* and *experienced utility* seems to conceptually parallel the same dichotomy. On the same vein, Rabin (1997) asserts that "For positive analysis, the usefulness of the utility-maximization framework depends on whether choice data can be usefully organized by positing that people maximize stable utility functions. A more controversial and rarer question about this framework is whether the preferences which people seem to maximize correspond to the well-being they actually experience. Many economists consider such a question off limits, feeling that 'by definition' the actions of informed people reflect what makes them happy. But there is a coherent sense in which even outcomes intentionally chosen may not maximize a decision maker's experienced well-being".

As far as game theory is specifically concerned, it is important to notice that this discipline was born *after* the crucial Pareto turn occurred in economics: as a consequence, in its intense development during the last decades, the process of progressive construction of its formal structure has been deeply rooted within a

behavioristic, revealed preference approach. As Sugden (1991) points out: "When economists and game theorists feel obliged to justify their use of these concepts, they still turn to Savage". However, while recognizing such a close connection between behaviorism and game theory is not a difficult task, justifications of it on methodological grounds are not so easy to find. A rather relevant (and probably the most significant) exception is Binmore's massive work (1994; 1998), which contains inter alia a systematic attempt to explain in which sense the methodological foundations of non-cooperative game theory necessarily lie in a strong version of the revealed preferences perspective. More specifically, it is the case that, insofar as one mechanically transfers this approach to game theory, then Binmore's position automatically follows: formally, there are no reasons why we should not incorporate 'whatever affects choice' into players' preferences, i.e. into game pay-offs.

3. 'Extended Pay-offs' and 'Preferential' Moral Principles

Binmore (1994) affirms that "It is a common source of misunderstanding for it to be thought that game theorists intend a payoff to be some naïve measure of a player's individual welfare, like a sum of money. However, game theory is based on the principle that the players act as though seeking to maximize the payoff they receive at the end of the game. A naïve view of the nature of a payoff will therefore sometimes not suffice. For example, it is easy to quote situations, especially in a moral context, where almost nobody would regard the amount of money that he gets as being the major determinant in deciding what to do. Game theorists therefore understand the notion of a payoff in a sophisticated way that makes it tautologous that players act as though maximizing their payoffs. Such a sophisticated view makes it hard to measure payoffs in real-life games, but its advantage in keeping the logic straight is overwhelming".

Our judgments are often biased by the belief that well-known categories such as, say, 'goods' or 'externalities' represent objective features of economic interactions. By claiming this, we tend to forget that, by contrast, characterizing a certain 'object' as a good or a bad, as a private or public good, as a relational or positional good or as a positive or negative externality always critically depends on the subjective *value-system* of the agents we are referring to in our analysis. The point is that while in a homogeneous and relatively simple society it will be quite easy for

its members to agree one with the other at this level, in a heterogeneous and relatively complex one differences in this regard are likely to be rather important and non-negligible. Such a subject-dependence clearly holds also as far as the concept of 'game' is concerned: the ranking of all possible outcomes of a non-cooperative game crucially depends on each agent's value-system. For example, the same 'material pay-off' can yield completely different consequences in terms of overall 'subjective pay-offs' depending on players' 'motivations', i.e. attitudes towards others. Properly speaking, each cell of a normal form game contains 'utilities', which can or cannot be a (more or less complex) function of some material pay-offs the game theorist considers significant for a proper description of the game. In this light, as far as individual goals are concerned, utilities can incorporate both an *objective* dimension (i.e. material rewards such as a sum of money) and a *subjective* one (critically dependent on each player's values): therefore, via utilities, both dimensions can be simultaneously accounted for by the game theorist.

In other words, it is the case that the same monetary pay-offs may translate into individuals''utilities' in different ways, according to each agent's 'happiness technology' (for this expression, see Menicucci and Sacco, 1996), in a world of motivationally heterogeneous players. We may add that, the more a society is motivationally heterogeneous, the greater the difference between monetary pay-offs and utilities and the more the description of the game in terms of 'objective' pay-offs is lacking and unsatisfactory. So far, our reasoning is perfectly compatible with Binmore's revealed preference view: *game pay-offs need not coincide with some naïve measure of individual welfare like a sum of money*. If this has often been the case, the reason has basically to do with the well-known, strong historical salience of the self-interest assumption in both economics and game theory. It is only when material net benefits perfectly coincide with the underlying 'utilities' that it is totally legitimate to describe a game by identifying individual pay-offs with the physical consequences faced by each player in correspondence with each possible outcome[5].

In recent years, with the development of behavioral and experimental economics, more and more studies have started focusing on 'extended pay-offs' incorporating not only material benefits but also psychological factors, social rewards

[5] Even in this respect, the PD provides a clear example of such a coincidence: here, insofar as players are assumed to exclusively care about their material pay-offs, it makes sense to specify the game by inserting in each cell the 'years of prison' each of them would get as a consequence of the combination of his and his opponent's strategic choices.

and moral principles: let us think of the literature on psychological games (Geanakoplos et al., 1989; Rabin, 1993) as well as of the related experimental works on positive and negative reciprocity (see, e.g., Fehr and Gächter, 1999; 2002; Fehr and Fischbacher, 2002). Inserting such non-material benefits into the pay-offs of the game is a fully legitimate operation right because of the above reasoning: insofar as these non-monetary goals are assumed to be a component of players' 'objective function', there are *no formal reasons* preventing the model builder from taking them into account. In other words, we may assert that the relevance of the self-interest assumption in the history of economic thought has tended to make us forget that, on purely formal grounds, game theory is well equipped to deal with other, more complex and sophisticated objective functions as well. As Hollis (1994) observes, "strictly, the standard first principle assumes only that agents are guided by their own preferences. In this sense are as 'self-interested' as sinners and the theory of rational choice is not committed to any view about how saintly or sinful we are"[6].

We ought to proceed even farther along this path, in order to lay stress on the following point: not only the model builder has access to formal tools which, in principle, are capable of properly taking such non-material rewards into account; insofar as such factors are considered important by the players themselves, he *has to* incorporate them into their 'extended pay-offs' (or 'utilities'). When they are deemed important by the players, but, for some reasons, the game theorist expresses the pay-offs in purely material terms, then *the resulting game will necessarily turn out to be improperly specified*, as what matters in a game are *players*' preferences, not modeler's ones. This observation seems to be in line with Ng's (1997) view, according to which happiness is far more important than more objective concepts such as choice, preference and income as "happiness is the ultimate objective of most, if not all people (…) We want money (or anything else) only as a means to increase our happiness. If having more money does not substantially increase our happiness, then money is not very important, but happiness is".

With reference to Binmore's objections to Sen's (and also, implicitly, to Sugden's) position, the major thesis defended here can be summarized as follows: on

[6] On the same vein, Camerer (2003) claims that "The payoffs are *utilities* for consequences. That is, in the original game the consequences may be money, pride, reproduction by genes, territory in wars, company profits, pleasure, or pain. A key assumption is that players can express their satisfaction with these outcomes on a numerical utility scale. The scale must at least be ordinal – i.e. they would rather have an outcome with utility 2 than with utility 1 – and when expected utility calculations are made the scale must be cardinal (i.e., getting 2 is as good as a coin flip between 3 and 1)".

the one hand, *insofar as we adopt a purely utility-maximizing, instrumental view of rationality* (where preferences – and, consequently, pay-offs – are broadly defined as 'whatever agents maximize' when strategically interacting with each other), we may easily agree with Binmore that *any* preferences (including 'moral preferences') ought to be already embedded in the extended pay-offs of the game *before* the game starts. Hausman and McPherson (1994) seem to adopt a similar (though more prudent) view as they argue that "The standard theory of rationality says that A's choices are determined by A's preferences. This sounds a bit like the claim that A is self-interested, but the impression is misleading. To say that A is self-interested is to make a claim about *what* A prefers. Utility theory does not rule out preferences for acting on moral principles or preferences for serving the interests of others. We are not claiming here that all moral theories are compatible with the standard theory of rationality. Our point is only that utility theory does *not* imply self-interest". We perfectly agree with Binmore as well as with Hausman and McPherson that *some* moral principles are compatible with a maximizing framework, in both strategic and non-strategic interaction scenarios.

In this light, if, say, a pre-play, material game takes the form of a PD but players are *altruists*, then, depending on their degree of altruism, the 'proper' or 'right' game (i.e. the one that takes account of *all* the relevant dimensions and not only of the material one) may well take the form of either an Assurance Game (AG), where, both (cooperate, cooperate) and (defect, defect) are Nash Equilibria in pure strategies, or an Other-Regarding game (OR), where cooperation is the dominant strategy for both agents. Sen (1973) himself agrees on this point, as he has no difficulties in admitting that "the entire problem under discussion can be easily translated into the case in which each person does worry about the other's welfare as well and is not concerned only with his own welfare. The numbers in the pay-off matrix can be interpreted simply as welfare indices of the two persons and each person's welfare index can incorporate concern for the other". What Sen is clarifying here is that insofar as we deal with a moral motivation such as altruism, defined as 'concern about the other's welfare' (elsewhere, he qualifies this motivation as 'sympathy'; see e.g. Sen, 1974), we can easily account for such moral category through a simple respecification of individual pay-offs, without altering any other element of the game as a whole. Selfish, altruistic or other types of preferences can be easily accommodated in the formal framework of non-cooperative game theory, as

game pay-offs are to be interpreted as 'welfare indices', reflecting players' (possibly heterogeneous) motivational systems.

4. 'Non-preferential' Moral Principles, Solution Concepts and Happiness

The considerations developed in Section 3 show that not only selfish preferences but also pro-social (like altruism) or anti-social ones (like envy) can be easily modeled in game-theoretic terms, through a proper specification of individual pay-offs. However, it is worth asking the following question: do people systematically *act* on the basis of their *preferences*? Are all possible principles of action based on the attempt to maximally satisfy one's preferences? Our point is that, insofar as we properly define individual preferences, we need to provide a negative answer to such questions. In particular, we claim that, for agent A, acting on her *preferences* may entail deciding *not* to act *morally* (and viceversa): in other words, it is intuitive to consider 'preferential behavior' as a form of action that may not be in line with a person's moral system, but that, by contrast, is directly linked to her non-rational impulses and inclinations, so that, as we all know, serious inner conflicts may arise between personal preferences and moral prescriptions. In this light, the qualitative difference between the two types of action is well captured by the well-known Kantian distinction between *autonomous* and *heteronomous* behavior: as Van Hees (2003) observes, in such perspective "an individual can either act morally, or be under the sway of her inclinations, desires etc. If she acts morally, i.e. if she acts 'from the moral law', she is said to act autonomously. On the other hand, if she acts on the basis of her impulses, inclinations, etc., she is acting heteronomously".

In such interpretation, a person's autonomy is closely related to the frequency of her morally justified actions, whereas forces such as desires and inclinations (that is, in our interpretation, 'preferences') would tend to make her behavior heteronomous. For example, with regard to Sen's (1977) distinction between *sympathy* and *commitment*, we may argue that while sympathy can be considered as a 'preferential' moral principle, commitment appears as a non-preferential one, as committing to a given behavior implies, by definition, acting against one's (properly defined) preferences. At this stage, we maintain then that a meta-principle through which both forms of 'preferential' and 'non-preferential' behavior can be incorporated within a unifying framework is provided by Anderson's (2001) priority claim:

The Priority of Identity to Rational Principle: what principle of choice it is rational to act on depends on a prior determination of personal identity, of who one is[7].

Anderson's claim highlights that there is something prior to behavioral choice*: the choice of which principle of choice to act on*. It is at this meta-choice level that agents have to decide whether to be, say, Kantian players, team thinkers (see Sugden, 2000) or utility maximizers. She is not arguing that only one principle of choice is rational, but simply asserting that the outcome of such a meta-choice crucially depends on the agent's self-conception.

Can we account for Anderson's priority claim in game-theoretic terms? Let us assume that, with respect to a given strategic interaction scenario, some players' identity is affected by preferential moral principles only but also that, say, other players' identity depends on acting on non-preferential moral principles: are we allowed to model such situation in game-theoretic terms? Sugden (2001) criticizes Binmore's (1994) interpretation of utility indices in games on the grounds that he has recourse to "a particularly strong form of revealed preference theory, in which it is a matter of definition that an individual's choices *always* reveal her preferences. Thus, once a game has been specified, with utility indices for the various possible outcomes, certain propositions about what a player will do (for example, that she will not choose a dominated strategy) are *necessary truths*, and not merely the implications of particular solution concepts which game theorists are free to dispute (I: 104-110). I am not convinced that this is the most useful – or indeed the conventional – way of interpreting utility in games". In the light of the above reasoning on Anderson's priority claim and of Sen's and other scholars' defence of non-utilitarian moral concepts such as commitment or duty, Sugden's critique to (strong forms of) the revealed preference approach sounds rather plausible: is it necessarily part of the definition of 'pay-offs' that players choose the strategy yielding the highest pay-off value, given the opponent's strategy? Is such a definition implied by the formal structure of the game? In our view, we are allowed to provide a negative answer to

[7] On the same vein, Zamagni (2003) asks the following question: "How can the idea of an agent who chooses autonomously and rationally be reconciled with the idea that happiness has to do not only with the satisfaction of preferences (utility) and thus of interests, but also with affections, emotions, moral dispositions – in a word, with personal identity?".

this question, as, in principle, we may have recourse to *different* solution concepts with regard to the *same* game structure, i.e. we can assume that players choose by relying on principles of action other than a criterion based on a purely instrumental account of rationality such as utility maximization.

Sen (1994) is very clear in stating that (a) there are deep reasons inducing us not to systematically respecify the pay-offs insofar as we want to incorporate in our formal structure concepts such as commitment and also that (b) these reasons, far from being exclusively formal, have a mainly substantive nature. In other words, matters like 'what is the game' correctly describing a given situation and 'how should agents play it' ought to be treated *separately*, as, with respect to a given game, different moral principles may be captured by distinct solution concepts. While knowing players' (extended) pay-offs is crucial in order to correctly specify what game is going to be played, such information in itself is not sufficient to tell us how rational agents will play that game: insofar as we interpret extended pay-offs as reflecting agents' preferences, their choices *need not be mechanically driven by (extended) pay-offs only*[8]. We believe that Anderson's priority claim allows us to draw such important implications for game theory: insofar as *identity* – and not preferences – is seen as the 'primum movens' of individuals' choice process, we cannot rule out, ex ante, that a certain game's pay-offs are common knowledge among the players but that such players, without behaving inconsistently, decide to act on a principle of action other than utility maximization. This is equivalent to assert that agents' principle of action is not part of the definition of game pay-offs and that, therefore, the maximizing format need not universally apply.

In particular, it seems to us to be misleading to have recourse to such format when players' decision process is significantly affected by non-preferential moral criteria. Further, in our view the above reasoning also entails that, other things being equal, a given game structure (say, a PD) keeps on being the same even if players'

[8] Camerer (2003) defines a game as consisting "of the 'strategies' each of several 'players' have, with precise rules for the order in which players choose strategies, the information they have when they choose, and how they rate the desirability (or 'utility') of resulting outcomes". As we can see, how to play is *not* part of the definition of the game. He further adds: "It is important to distinguish *games* from game *theory*. Games are a taxonomy of strategic situations, a rough equivalent for social science of the periodic table of elements in chemistry. Analytical game *theory* is a mathematical derivation of what players with different cognitive capabilities are likely to do in games". That he believes, as we do, that action may derive from decision criteria other than utility-maximization, can be inferred from the following statement: "Dominance is important because, if utility payoffs are correctly specified (…) and *players care only about their own utility*, there is no good reason to violate strict dominance" (italics added).

identity induces them to adopt a non-utilitarian mode of rationality: the point is that the departure from a logic of play such as utility maximization occurs at a level which is different from the (pre-play) level of (suitably defined) individual preferences (captured by properly specified game pay-offs)[9]. According to the interpretation suggested in this essay, it would be unsatisfactory to avoid such departure in terms of logic of play by accomodating the moral principle under study through a simple respecification of the game's pay-offs: the problem is that – unlike situations where morally-charged preferences such as altruism (or sympathy) are involved – by so doing *we would simultaneously alter the very nature of such non-preferential moral criteria* and, therefore, we would not do justice to them.

By (a) introducing a rationality concept other than utility maximization and (b) preserving the original game structure, we are allowed to make such a non-utilitarian principle of action choice-relevant while at the same time making clear the distinction between preferential and non-preferential factors. Further, this approach is useful in order to comparatively analyze what actually happens (when a non-preferential solution concept is involved) and what would happen if the players were driven by their preferences only (i.e. if they decided by simply choosing the strategy yielding the highest pay-off): this clearly entails that *counterpreferential choices* may well occur within this scenario. As anticipated above, a further advantage of interpreting (a) pay-offs as reflecting individual preferences and (b) solution concepts in the light of the principles of action adopted by the players, without establishing any mechanical and necessary connection between (a) and (b), is the following: such a framework allows us to incorporate potential inner conflicts between preferences and non-preferential moral principles in the formal structure of a non-cooperative game, i.e. to explicitly consider the complex interplays taking place between different versions of rationality. This implies that, with reference to a given game structure, fruitful links between solution concepts and moral principles might be established,

[9] In the light of these considerations, we do not agree with Binmore's (1994) critique of Sen: according to Binmore, Sen is confusing "what has to be analyzed with how the analysis is conducted. When a game comes to be analyzed, intelligibility demands that matters like those raised by Sen should *already have been incorporated* into the structure of the game. If the players have the power to *alter their preferences* to commit themselves to behaving in certain ways *before* the play of the Prisoners' Dilemma, then it is *not the Prisoners' Dilemma* that they are playing, but some more complicated game. (…) Players cannot alter the game they are playing. If it seems like they can, it is because the game has been improperly specified".

while at the same time making clear that *both* preferences and non-preferential moral criteria affect, to some degree, players' choices.

5. The paradoxes of happiness in strategic interaction scenarios

Amartya Sen, in one of his classic contributions (Sen, 1985), focuses on three notions of 'privateness': (i) *self-centered welfare*, (ii) *self-welfare goal*, (iii) *self-goal choice*, claiming that such concepts are quite independent of each other: (i) a choice is not necessarily driven by the pursuit of a given goal; (ii) a goal is not necessarily aimed at increasing the person's own welfare and, further, (iii) aiming at increasing one's welfare does not always entail increasing one's consumption levels. Kahneman et al.'s (1997) more recent and well-known distinction between decision utility (basically the choice-based characterization of utility Sen refers to) and experienced utility (conveying a Benthamite, content-based characterization of utility) seems to be partly related to Sen's considerations: as far as individual 'welfare' or 'happiness' is concerned, choice may well happen to be 'externally inconsistent' (Hsee, 2003), i.e. it may reveal itself unable to increase the chooser's happiness level[10]. As Rabin (1997) observes, "Not knowing your own 'experienced utility function' is obviously important for the welfare implications of choice, and the main lesson of this material is that economists ought recognize that people may not correctly predict what makes them happy".

The point we would like to make here is that the above recalled wedge between choice and happiness is quite a general phenomenon, observable in both strategic and non-strategic interaction scenarios. As far as social environments where choice occurs under parametric conditions, several explanations can be simultaneously considered, like evolving aspirations (see Rabin, 1997; Easterlin, 2001) and hedonic adaptation (see Frederic and Loewenstein, 1999). A further and not necessarily alternative explanation of the choice-happiness wedge is the one suggested by Hsee (2003), referring to this possibility as to a form of 'choice-consumption inconsistency': "In situations where people choose between hedonic consumption options, the external consistency question becomes whether the option people choose delivers the best consumption experience. (…) If people choose an

[10] It is interesting to remark that such risk of external inconsistency of choice, with respect to a purpose such as happiness, had been anticipated by Kant.

option that delivers worse consumption experience over one that delivers better consumption experience, holding costs constant, we say that they exhibit a choice-consumption inconsistency".

Hsee's (2003) explanation lies in the so called JE-SE mode distinction. He maintains that an important, though largely neglected, contributor to choice-consumption inconsistency is *evaluation mode*, in the sense that choice is usually made in the Joint Evaluation (JE) mode, with several goods to be compared, whereas consumption occurs in the Separate Evaluation (SE) mode, where one good only (the one previously bought by the agent in the JE) is present. Once we take this distinction into account, we are not entitled to conclude that choice *reveals* the underlying preference structure of the agent. The simultaneous presence of multiple options to be compared ex ante generates a bias (Hsee and Zhang (2004) call it *distinction bias*) and may lead to choices which turn out to be unpleasant (or less pleasant than expected) ex post: evaluation mode may then drive a wedge between decision utility and experienced utility.

The choice-happiness wedge illustrated above is even more likely to occur in strategic interaction scenarios, i.e. in social contexts where agents interact strategically, constantly aiming at correctly predicting others' preferences and behaviors. Interestingly, Sen (1973) seems to note that the gap between individual preferences and welfare is greater the more social interaction is complex, i.e., we could say, the more it takes place in a sophisticated, strategic interaction situation – the subject matter of game theory. As we observed above, it is frequently the case that the actual degree of happiness people experience ex post (experienced utility) significantly differs from the one expected ex ante (predicted utility). In social contexts where agents interact strategically, that general proposition holds because 'disappointing' equilibria can occur even though every player would have preferred to end up in a different outcome (like inefficient, mutual defection equilibria in the PD). In such scenarios, it is worth specifying that, unlike what happens in non-strategic contexts, there is *nothing inherently paradoxical* in the fact that we cannot choose what we would prefer (e.g. a 'I defect-You cooperate' outcome in a PD for a selfish player) when our actions are *interdependent* the one with the other. This point is expressed very clearly by Binmore (1994): "Personally, I see no paradox at all in the fact that independent choice behavior by rational agents should sometimes lead to Pareto-inefficient outcomes. The rules of the Prisoners' Dilemma create an

environment that is inimical for rational cooperation and, just as one cannot reasonably expect someone to juggle successfully with his hands tied behind his back, so one cannot expect rational agents to succeed in cooperating when constrained by the rules of the Prisoners' Dilemma".

In social dilemmas and other strategic interaction environments, it is of interest to remark that selfish players could obtain better results if they behaved *as if* they had a preference for cooperation or through a principle of action other than utility maximization. In other words, in such contexts it is the case that if agents act on a non-maximizing principle of action, they may receive an individual benefit which is greater than the one they would obtain via explicit utility maximization. In the PD, for example, if we assume that players' principle of action is, say, (pseudo)'Kantian rationality', then the equilibrium they reach is the best possible outcome this interaction scenario may yield from two different points of view, i.e. from both (a) a Kantian and (b) a utility-maximizing perspective. As to the former perspective, it is easy to see that the equilibrium outcome of mutual cooperation is consistent with the moral prescription each agent Kantianly decides to comply with, i.e. with the universalizable law prescribed by the categorical imperative illustrated in Kant's *Groundwork*: "Act only on that maxim whereby thou canst at the same time will that it should become a universal law". As Sugden (1991) observes, "reasons may override desires: it may be rational to do what one does not desire to do. As the example of the Prisoner's Dilemma suggests, this line of thought threatens to undermine game theory". A very similar point was lucidly made by Rapoport (1987), arguing that developments of game theory "provide a rigorous rationale for Kant's Categorical Imperative; act in the way you wish others to act. Acting on this principle reflects more than altruism. It reflects *a form of rationality*" (italics added).

However, an even more interesting point is that such an equilibrium constitutes the best possible outcome even if agents were genuinely utility-maximizing players and only 'as if' Kantians, i.e. had adopted a Kantian principle of action for purely instrumental reasons. In other words, in social dilemmas scenarios, being 'englightened utility-maximizers' who consciously decide to opt for a Kantian moral law but, at the same time, keep on evaluating their own welfare in purely preferential terms, may turn out to pay off, i.e. to be a better comprehensive strategy with respect to actual utility maximization. In this regard, Sen (1973) affirms that "Even in the absence of a contract, the parties involved will be better off following

rules of behaviour that require abstention from the rational calculus which is precisely the basis of the revealed preference theory. People may be induced by social codes of behaviour to act *as if* they have different preferences from what they really have"[11].

In our view, this appears as one of the most interesting 'paradoxes of happiness in strategic interaction scenarios', i.e. a paradox arising in game-theoretic settings, where, by definition, rationality has a strategic and not a parametric nature. Anderson (2001), commenting on Sen's (1977) famous essay, observes that what is foolish in a PD is not the lack of a preference for cooperation: "it is hardly foolish to not prefer the act of cooperating in itself, apart from its consequences. What is foolish about non-cooperators is not their preferences, which are perfectly understandable, but their principle of rational choice. And what makes that principle foolish is its act-consequentialist structure. Any principle of rational choice that evaluates an individual's act solely according to its marginal causal impact on valued outcomes will meet the same difficulties. This is one powerful reason why many people are drawn away from act-consequentialism toward rule-consequentialism, or toward non-consequentialist frameworks".

On methodological grounds, such considerations suggest that introducing a *softer* link between preference and choice (i.e. a departure from the pure version of the revealed preferences approach) may allow for the achievement of important substantive results even within an ultimately utility-maximizing framework, as far as PD-like social interactions are concerned. In other words, what happens in social dilemmas (which are relatively special but rather frequent and relevant interaction scenarios) is somehow *the opposite* with respect to what is predicted by the 'as though' thesis defended by revealed preferences theorists. Their position is that any non-maximizing principle can, in fact, be conceptualized in maximizing terms, even when agents are not aware of this; by contrast, we claim here that, in social dilemmas, not only Kantian players preserve their identity at behavioral level and are neither actual nor 'as though' utility maximizers, but also that even utility-maximizers would

[11] Similarly, with regard to the PD, Sen (1985) observes that "if people are ready to act (individualistically) on the basis of some 'as if' – more cohesive – orderings, then they can do better than acting individualistically in direct pursuit of their real goals. And they do better judged in terms of the real goals themselves" (see also Menicucci and Sacco (1996) for interesting insights on this).

better adopt an enlightened reasoning and, without altering their actual, ultimately preference-centered rationality, act *as though* they were Kantian players[12].

6. Concluding remarks

In the light of the analysis developed in the previous sections, we can identify the following advantages in adopting the approach suggested here. In the first place, such an analytical framework allows us to discriminate the causal role of (suitably defined) 'preferences' and non-preferential moral principles – seen as distinct determinants of 'expected happiness' – on individual behavior within strategic interaction settings. As a consequence, it can account for the complex interplays taking place between such two dimensions, so that, in principle, even potential conflicts between them (occurring at intra-individual level) may be explored. Recognizing that the search for happiness in advanced societies increasingly depends on such interaction between preferential and non-preferential factors seems to be a significant reason for proceeding along the path indicated above.

Second, as we have seen, several important social scenarios exist where, insofar as we have recourse to non-utilitarian solution concepts such as, say, the Kantian principle of universalizability or team reasoning (see on this Sugden, 2000), individual players are capable of obtaining results which are Pareto-superior to the ones they would get within a classic, maximizing framework. In the light of the considerations developed in the last section, social dilemmas can be seen as one of the most interesting settings where, as far as the link between happiness, morality and game-theoretic solution concepts is concerned, paradoxical implications arise. The main thesis defended in Section 4 can be summarized as follows: in several social situations where agents interact strategically, even for utility-maximizers the achievement of the maximum degree of happiness may occur as a byproduct of non-preferential principles of action, that is of principles of action not aimed at generating such effect. Symmetrically, within such strategic interaction environments**,** purposely utility-maximizing patterns of behavior lead to disappointing results: with regard to the PD, Aumann (1987) points out that "The universal fascination with this game is

[12] It should be clear, at this stage, that while for expositional ease we are constantly referring to (pseudo)Kantian rationality, the line of reasoning developed here has a far wider reach and applies to any non-preferential principle of action which, in social dilemmas, is capable of making mutual cooperation a feasible equilibrium outcome.

due to its representing, in very stark and transparent form, the bitter fact that when individuals act for their own benefit, the result may well be disaster for all". A further and related paradox within such settings has to do with the fact that, provided that they decide to behave *as if* they were driven by the pursuit of ends other than preference satisfaction, utility maximizers can get happier than by acting individualistically. Happiness can then arise, in the contexts under study, either as the predictable consequence of enlightened utility-maximizing agents or as an unintended effect of non-preferential behavior, but not as the predictable consequence of choices made by standard utility-maximizing agents.

References

Anderson, E. (2001), 'Symposium on Amartya Sen's Philosophy: 2. Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Economics and Philosophy', Economics and Philosophy, 17, 1, 21-38.

Antoci, A., Sacco, P.L., Zarri, L. (2004), 'Coexistence of Strategies and Culturally-specific Common Knowledge: An Evolutionary Analysis', Journal of Bioeconomics, 6, 165-194.

Aumann, R. (1987), 'Game Theory', in John Eatwell, Murray Milgate and Peter Newman (eds.), The new Palgrave: A dictionary of economics, vol. 3., London: Macmillan.

Binmore, K. (1994), Game Theory and the Social Contract. Volume I: Playing Fair, Cambridge (Mass.): MIT Press.

Binmore, K. (1998), Game Theory and the Social Contract. Volume II: Just Playing, Cambridge (Mass.): MIT Press.

Camerer, C. (2003), Behavioral Game Theory, Princeton: Princeton University Press.

Easterlin, R.A. (2001), 'Income and Happiness: Towards a Unified Theory', Economic Journal, 111, 465-484.

Fehr, E., Fischbacher, U. (2002), 'Why social preferences matter – The impact of non-selfish motives on competition, cooperation and incentives', Economic Journal, 112, 1-33.

Fehr, E., Gächter, S. (1999), 'Reciprocal Fairness, Heterogeneity, and Institutions', Paper presented at the AEA Meeting in New York, Jan. 3-5.

Fehr, E., Gächter, S. (2002), 'Altruistic punishment in humans', Nature, 415, 137-140.

Frederic, S., Loewenstein G. (1999), 'Hedonic adaptation', in Daniel Kahneman, Ed Diener and Norbert Schwartz (eds.), Well-being: the foundations of hedonic psychology, New York: Russell Sage Foundation, 302-329.

Geanakoplos, J., Pearce, D., Stacchetti E. (1989), 'Psychological Games and Sequential Rationality', Games and Economic Behaviour, 1, 60-79.

Hausman, D. (2000), 'Revealed Preference, Belief, and Game Theory', Economics and Philosophy, 16, 99-115.

Hausman, D., McPherson, M. (1994), 'Economics, rationality, and ethics', in Daniel Hausman (ed.), The Philosophy of Economics, Cambridge: Cambridge University Press.

Hollis, M. (1994), The Philosophy of Social Science, Cambridge: Cambridge University Press.

Hsee, C.K. (2003), 'Inconsistency between choice and experience', Paper presented at the International Conference on the 'Paradoxes of Happiness in Economics', University of Milan-Bicocca, 21-23 March.

Hsee, C.K., Zhang, J. (2004), 'Distinction bias: Misprediction and mischoice due to joint evaluation', Journal of Personality and Social Psychology, 86, 5, 680-695.

Kahneman, D., Wakker P., Sarin R. (1997), 'Back to Bentham? Explorations of experienced utility', Quarterly Journal of Economics, 112, 2, 375-406.

Menicucci, D., Sacco, P.L. (1996), Rawlsian Altruism and Efficiency, Studi e Discussioni, 102, Department of Economics, University of Florence.

Ng, Y. (1997), 'A Case for Happiness, Cardinalism, and Interpersonal Comparability', Economic Journal, 107, 1848-1858.

Rabin, M. (1993), 'Incorporating Fairness into Game Theory and Economics', American Economic Review, 83, 1281-1302.

Rabin, M. (1997), 'Psychology and Economics', Berkeley Department of Economics, Working Paper, No. 97-251.

Rabin, M. (2002), 'A perspective on psychology and economics', European Economic Review, 46, 657-685.

Rapoport, A. (1987), 'Prisoner's Dilemma', in John Eatwell, Murray Milgate and Peter Newman (eds.), The new Palgrave: A dictionary of economics, vol. 3., London: Macmillan.

Sacco, P.L., Zamagni S. (1996), 'An Evolutionary Dynamic Approach to Altruism', in Francesco Farina, Frank Hahn and Stefano Vannucci (eds.), Ethics, Rationality and Economic Behaviour, Oxford: Clarendon Press, 265-300.

Sacco, P.L., Zarri, L. (2002), 'Collective Action Dilemmas and Norms of Social Reasonableness', Ars Interpretandi. Yearbook of Legal Hermeneutics, 7, 401-425.

Sen, A. (1973), 'Behaviour and the Concept of Preference', Economica, 40, 241-259.

Sen, A. (1974), 'Choice, Ordering and Morality', in S. Körner (ed.), Practical Reason, Oxford: Blackwell, 54-67.

Sen, A. (1977), 'Rational fools: a critique of the behavioral foundations of economic theory', Philosophy and Public Affairs, 6, 317-344.

Sen, A. (1982), Choice, welfare and measurement, Oxford: Basil Blackwell.

Sen, A., (1985), Goals, Commitment, and Identity, Journal of Law, Economics and Organization, 1, 2, 341-355.

Sen, A. (1994), 'The formulation of rational choice', American Economic Review, 84, 2, 385-390.

Sen, A., Williams, B. (1982), Utilitarianism and Beyond, Cambridge: Cambridge University Press.

Sugden, R. (1991), 'Rational Choice: A Survey of Contributions from Economics and Philosophy', Economic Journal, 101, 407, 751-785.

Sugden, R. (2000), 'Team Preferences', Economics and Philosophy, 16, 175-204.

Sugden, R. (2001), 'Ken Binmore's Evolutionary Social Theory', Economic Journal, 111, 213-243.

Sugden, R. (2002), 'Beyond Sympathy and Empathy: Adam Smith's Concept of Fellow Feeling', Economics and Philosophy, 18, pp. 63-88.

Van Hees, M. (2003), 'Acting autonomously versus not acting heteronomously', Theory and Decision, 54, 4, 337-355.

Zamagni, S. (2003), 'Happiness and Individualism: an Impossible Marriage', Paper presented at the International Conference on the 'Paradoxes of Happiness in Economics', University of Milan-Bicocca, 21-23 March.