View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Serveur académique lausannois

Studies in History and Philosophy of Biological and Biomedical Sciences 56 (2016) 24–31

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

The evolution of utility functions and psychological altruism

Christine Clavien*, Michel Chapuisat

Department of Ecology and Evolution, University of Lausanne, Unil-Sorge, Biophore, 1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Available online 18 November 2015

Keywords:
Psychological altruism
Utility function
Preference
Homo hamiltoniensis
Homo economicus
Model of social evolution

ABSTRACT

Numerous studies show that humans tend to be more cooperative than expected given the assumption that they are rational maximizers of personal gain. As a result, theoreticians have proposed elaborated formal representations of human decision-making, in which utility functions including "altruistic" or "moral" preferences replace the purely self-oriented *"Homo economicus"* function. Here we review mathematical approaches that provide insights into the mathematical stability of alternative utility functions. Candidate utility functions may be evaluated with help of game theory, classical modeling of social evolution that focuses on behavioral strategies, and modeling of social evolution that focuses directly on utility functions. We present the advantages of the latter form of investigation and discuss one surprisingly precise result: *"Homo economicus"* as well as "altruistic" utility functions are less stable than a function containing a preference for the common welfare that is only expressed in social contexts composed of individuals with similar preferences. We discuss the contribution of mathematical models to our understanding of human other-oriented behavior, with a focus on the classical debate over psychological altruism. We conclude that human can be psychologically altruistic, but that psychological altruism evolved because it was generally expressed towards individuals that contributed to the actor's fitness, such as own children, romantic partners and long term reciprocators.

© 2015 Elsevier Ltd. All rights reserved.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Biological and Biomedical Sciences*

1. Introduction

Neo-classical economics is often criticized for formalizing human decision-making as a purely self-oriented process, according to which humans act only upon preferences for their own welfare maximization—humans may choose actions that benefit others or the public welfare, but only when these actions also benefit them. Yet, there is room within the neo-classical theoretical framework for alternative and more other-oriented descriptions of human motivation. Which formal account of decision-making is the most fitting description or the best predictive tool will depend very much on the area of human activity that is investigated—e.g. stock market dynamic interactions, consumer–producer interactions, private social interactions (Kirchgässner, 2008).

An important difficulty faced by theoreticians is to find a reliable method for assessing alternatives formalizations of decision-making, and decide which one is the most fitting for the area of human activity that is investigated. Here we discuss the formalization of private social interactions with utility functions, we investigate various ways to assess the mathematical stability of these utility functions, and we discuss the impact of this area of research for understanding human other-oriented psychological mechanisms.

The paper is organized as follows. Section 2 provides some introductory notions of neo-classical economic theory and an overview of the sort of utility functions that can be elaborated to represent private social interactions. We then describe the main features of game theory (Section 3), classic models of social evolution (Section 4), and models of social evolution that take utility functions as evolving traits (Section 5). Along the way, we present important results obtained with these methods and explain why

* Corresponding author. Present address: iEH2, Centre Médical Universitaire, 1 rue Michel Servet, 1211 Genève 4, Switzerland.
E-mail address: christine.clavien@unige.ch (C. Clavien).

http://dx.doi.org/10.1016/j.shpsc.2015.10.008
1369-8486/© 2015 Elsevier Ltd. All rights reserved.

the latter method is the most promising mathematical tool to assess utility functions. In Section 5 we also present a fascinating result: Ingela Alger and Jörgen Weibull (2013) have found that "*Homo economicus*"—i.e. the purely self-oriented utility function—and all the "altruistic" functions—i.e. those that contain a preference for social partners welfare—are evolutionary unstable in the presence of a utility function containing a preference for the common welfare that is conditionally expressed in social contexts where other individuals have similar preferences. In Section 6 we provide psychological interpretations of utility functions and investigate the extent to which mathematical models inform us about human other-regarding and altruistic motivation. This investigation reveals the conditions under which humans are likely to care for collaborative interactions and to evolve psychological altruism.

2. A taxonomy of utility functions

An important goal in neo-classical economics is to find the best way to formalize human's choices of action. The aim is to understand and predict individual behavior in socio-economic contexts, such as situations of conflicting interests. Neo-classical economics assumes that, whenever humans have the choice between alternative actions—e.g. investing or not, collaborating or not, helping or not–, they choose to maximize their personal utility (Kirchgässner, 2008). Individual utility can be described mathematically as a function of hierarchically ranked preferences for objects of choice—e.g. goods, states of the world. Utility functions can take an infinite variety of forms (see Fig. 1) but their relevance depends on whether they capture real features of human decision-making in the particular area of activity that is investigated. Let us consider some utility function that may characterize private social interactions.

The simplest function, usually labeled *"Homo economicus"*, reduces human preferences to individuals' own welfare—or payoff—maximization, where welfare is defined as an objective and measurable currency such as material or economic profit, or number of offspring. Mathematically, for a two person interaction, it can be formalized with the following equation (Weibull, 1995):

$$\text{"}Homo\ economicus\text{"} \quad \mu_{HE,\ i}(x_i, x_j) = \omega_i(x_i, x_j) \quad , \quad i \neq j$$

where $\mu_i(x_i,x_j)$ describes the actor's *utility*, that is, how much she values—gives weight to—the outcome of the interaction (x_i,x_j) where the actor plays x_i and her social partner plays x_j, and $\omega_i(x_i,x_j)$ is the actor's *objective welfare* if interaction (x_i,x_j) is performed. The formula can be simplified to: $\mu_i = \omega_i$.

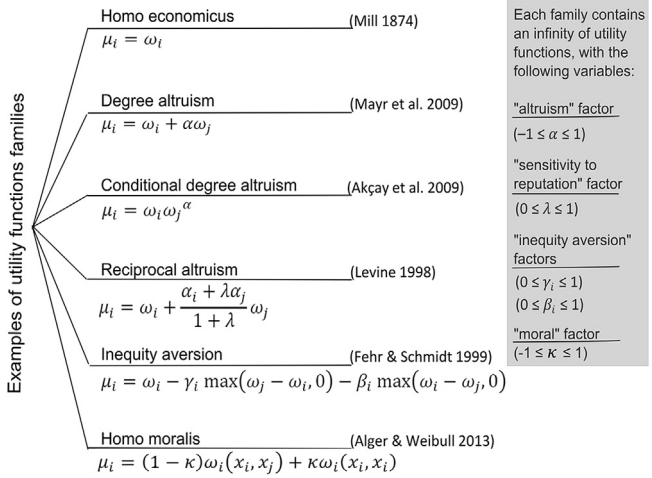

Fig. 1. Examples of utility functions families.

"Homo economicus" is a purely self-oriented utility function because it induces individuals to ignore other individuals' welfare, as well as the common good. Other utility functions combine self-directed and social or other-oriented preferences. Gary Becker (1976) for example defines a utility function for an actor who cares as much about her social partner's welfare as about her own welfare. We refer to this as:

$$\text{"}Egalitarian\ altruism\text{"} \quad \mu_{EA,\ i} = \omega_i + \omega_j$$

where μ_i describes the actor's utility, ω_i the actor's welfare, and ω_j the welfare for the social partner.

This model fails to capture the fact that humans usually care more for their own welfare than for others' welfare. To account for this phenomenon, several theoreticians have proposed a family of utility functions that integrate the sum of the actor's welfare and the welfare of the social partner weighted by an altruistic factor (e.g. Mayr, Harbaugh, & Tankersley, 2009). We label this specific form of altruism:

$$\text{"}Degree\ altruism\text{"} \quad \mu_{DA,i} = \omega_i + \alpha\omega_j$$

where α $(-1 \le \alpha \le 1)$ describes how much the actor cares for the welfare of her social partner. When $\alpha = 0$, she cares only for her own welfare; when $\alpha = 1$, she values her partner's welfare as much as her own. This formula captures the idea that some individuals in a population may be more altruistic than others. It can also represent spiteful preferences, since negative values of α mean that the actor is motivated to reduce the other's welfare. Note that *"Homo economicus"* and *"Egalitarian altruism"* are special cases of *"Degree altruism"*, where $\alpha = 0$ and 1 respectively.

An alternative utility function has been proposed by Akçay, Van Cleve, Feldman, & Roughgarden (2009). They represent how much an individual 'likes' a given outcome as the product of her own payoff and her partner's payoff weighted by an altruistic factor. We refer to this as:

$$\text{"}Conditional\ degree\ altruism\text{"} \quad \mu_{CDA,i} = \omega_i \omega_j^{\alpha}$$

Here, α also describes how much the actor cares for the welfare of her social partner but its weight depends on the welfare state of the actor. This model captures the fact that humans may be less likely to care for others when they are in a state of need—and reversely.

Another interesting family of utility functions has been developed by David Levine (1998). Here, the actor maximizes the addition of her personal welfare, and her social partners' welfare weighted by two factors: the extent to which the actor cares for her social partner (altruistic factor) *and* the extent to which her social partner cares for others (reciprocal factor). More precisely, the actor cares more for the welfare of social partners that are believed to be more altruistic. Levine defines this family of function as:

$$\text{"}Reciprocal\ altruism\text{"} \quad \mu_{RA,\ i} = \omega_i + \frac{\alpha_i + \lambda\alpha_j}{1+\lambda}\omega_j$$

where α represents the individual's coefficient of altruism (or spitefulness) and λ $(0 \le \lambda \le 1)$ represents how much the actor is sensitive to her partner's coefficients of altruism: $\lambda = 0$ means that she is not influenced by the other's character (in this case, *"Reciprocal altruism"* boils down to *"Degree altruism"*), and positive values imply that she cares for her partner welfare proportionally to the partner coefficient of altruism.

This formula captures the fact that "players care not only about other players' utility, but also that their attitudes toward other players depend on how they feel they are being treated" (Levine, 1998). Indeed, the equation implies that the actor is less altruistic towards less altruistic partners.

Our last example refers to Erst Fehr and Klaus Schmidt (1999). Their family of utility functions, labeled *"Inequity aversion"*, captures the fact that "people resist inequitable outcomes; i.e., they are willing to give up some material payoff to move in the direction of more equitable outcomes." Thus, the actor's utility is maximized if inequity is minimized:

$$\text{"}\textit{Inequity aversion}\text{"} \quad \mu_{IA,\ i} = \omega_i - \gamma_i \max(\omega_j - \omega_i, 0) - \beta_i \max(\omega_i - \omega_j, 0)$$

where $\gamma_i \max(\omega_j - \omega_i, 0)$ represents the utility loss from inequalities disadvantageous for the actor (this term becomes positive when $\gamma_i > 0$ and $\omega_j > \omega_i$), and $\beta_i \max(\omega_i - \omega_j, 0)$ represents the utility loss from inequalities advantageous for the actor (this term becomes positive when $\beta_i > 0$ and $\omega_i > \omega_j$). Note that whenever γ_i and $\beta_i > 0$, the actor's utility μ_i, is maximized when the welfare state of her partner equals her own welfare state, thus when $\omega_j = \omega_i$.

The sheer number of utility functions raises the difficulty of deciding which ones—if any—are accurate descriptions of human private social interactions. For this task, various assessment methods are available (see Fig. 2) such as: i) appraising the extent to which utility functions generate non-trivial predictions that can be tested in controlled experiments (Gigerenzer & Selten, 2001; Guala, 2005; Smith, 1962); ii) investigating the empirical validity of utility functions with field observations and qualitative data (Emirbayer & Mische, 1998); iii) examining utility functions' theoretical coherence with help of reasoning and logical argumentation (Kincaid & Ross, 2009; Nozick, 1969); iv) assessing utility function's mathematical stability. This can be done with game theory (Axelrod, 1984; Binmore, 2005), to investigate the extent to which a utility function leads to equilibrium situations—e.g. a Nash equilibrium, a Bayesian Nash equilibrium, an evolutionary stable equilibrium—in various strategic games or situations. Alternatively, information about evolutionary stability can be inferred from models of social evolution (Güth & Yaari, 1992). All these methods are relevant procedures for evaluating utility functions. Here, we present and discuss the importance of mathematical assessment tools.

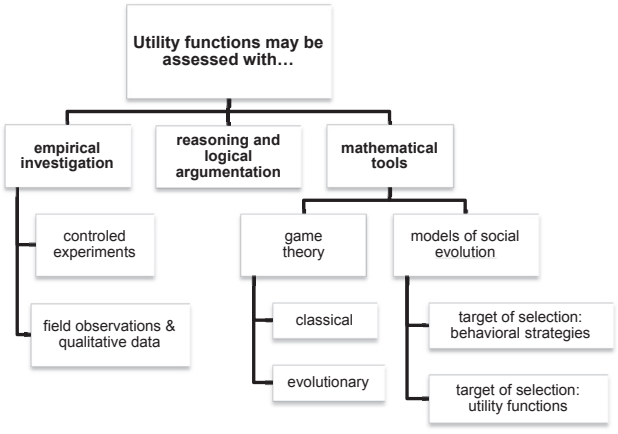

Fig. 2. Illustration of the various ways to assess the value of utility functions. See accompanying text for details.

3. Game theory

Game theory is the study of strategic interactions among social partners (Nash, 1950; Schelling, 1960). It provides mathematical tools for assessing if behavioral strategies reach stable equilibria. Games are composed of general interaction rules—e.g. players interact pairwise, they can play cooperation or defection —, a set of behavioral strategies which are instructions about which action to take in each possible situation in the game—e.g. 'Always play defection', 'Choose randomly', 'Defect if your partner has a cooperation-reputation below x and cooperate otherwise' —, and an outcome matrix which defines the payoff of players for each possible interaction. Scholars search for behavioral equilibria in these games, such as the Nash solution that arises when each strategy is a "best reply" to the other strategy; this means that given her partner's choice of action, none of the players would have earned more in choosing an alternative action. These equilibria are considered as indices for stability and efficiency in real situations (Binmore, 2005).

In the subfield of *evolutionary game theory*, payoff is a surrogate for fitness and strategic interactions are modeled as collections of repeated games—e.g. series of pairwise interactions between randomly matched individuals —, where strategies "reproduce" and transmit clones of themselves to the next series of repeated games. In these models, the desired equilibrium is the evolutionary stable situation where one strategy—or a well-defined mixture of strategies—cannot be invaded by another strategy.

Game theory may be used as an indirect way to test the evolutionary stability of utility functions because there is a causal link between behavioral strategies and utility functions. Let us specify this causal link before examining how game theory can be used for assessing utility functions.

Behavioral strategies are immutable action rules for specific social contexts that humans may encounter during their life. *Utility functions* provide general instructions for deciding which behavioral strategies to use in a range of social contexts. These general instructions may include some level of preference for own welfare—ω_i—some level of preference for the social partner's welfare—$\alpha\omega_j$ —, some level of aversion against inequitable outcomes—γ_i and β_i —, etc. A utility function can generate different behavioral strategies in different social settings. For instance, in a two players iterated prisoner's dilemma[1] situation with no knowledge about social partners' behavioral propensities, the "*Homo economicus*" utility function induces the strategy "always defect". In contrast, when more information is provided—e.g. knowledge about social partners' past behavior —, "*Homo economicus*" induces cooperative behavior towards partners that are likely to cooperate with cooperators and to punish non cooperators—assuming that the cost for being punished is higher than the benefit from defecting. As second example, *"Reciprocal altruism"* with positive α and λ values leads to an other-oriented strategy when the actor *believes* that she interacts with an 'altruistic' partner, and to a self-interested strategy otherwise. This rule holds independently of whether the social partner has cooperated or defected in the preceding interaction. From this example, it should be clear that *"Reciprocal altruism"* as utility function should not be confused with *Reciprocal altruism* as behavioral strategy—i.e. *Tit For Tat*.

[1] A prisoner's dilemma is a two player game. Each player decides either to cooperate or to defect without knowing the other player's choice. The payoff matrix of the game is such that mutual cooperation provides better benefit than mutual defection, while unilateral defection provides the largest benefit. In such a situation, both players have an incentive to defect, although they would be better off if they both decided to cooperate.

Assuming that utility functions are inherited, those that induce behavioral strategies that are overall advantageous for their bearers will be selected, and vice versa. A well-known result in evolutionary game theory (Axelrod, 1984; Binmore, 2005) is that discriminative cooperation with other cooperators but not with defectors usually reaches equilibrium in iterated prisoner's dilemma situations. Thus, one may expect that stable utility functions induce such discriminative cooperation. Reversely, utility functions that fail to do so are likely to be unstable. This is for instance the case of *"Egalitarian altruism"* because it always induces its bearers to choose a cooperative strategy that promotes both partners' payoff irrespective of whether the partner is a cooperator or a defector.

This way of assessing utility functions is limited for two reasons. First, different utility functions may give rise to the same behavior in a specific social context. For instance if individuals cannot obtain information about their partners' utility functions, *"Degree altruism"* and *"Reciprocal altruism"* induce the same behavioral strategies. As second example, the various altruistic utility functions with a positive but not too large α value usually induce cooperative behavior—the α factor leads the actor to do so—except when own welfare is threatened too much—the ω_i factor leads the actor to avoid being fully exploited. Therefore, whenever two utility functions generate similar outcomes in a given game, they cannot be differentiated. Second, since utility functions can generate different behavioral strategies in different social contexts, their evolutionary stability should be assessed across social contexts. Indeed, evolution is always a trade-off between different fitness costs and benefits obtained in the course of organisms' life (Nettle, 2006). If a utility function is successful in most fitness-relevant contexts—i.e. it generates a behavior that is generally advantageous across contexts —, evolution is likely to select it. In light of this knowledge, results obtained with models that contain specific assumptions over social interactions—e.g. prisoner's dilemma situations with random assortment between individuals and memory of past interactions—and take into account a limited number of competing strategies cannot easily be generalized.

4. Models of social evolution I—focus on behavioral strategies

Population genetics provides the foundation for models of social evolution (Frank, 1998). These models have helped understanding a vast number of animal and human behavior such as cooperative interactions, parental investment, or mate choice (Cavalli-Sforza & Feldman, 1981; Davies, Krebs, & West, 2012; MacElreath & Boyd, 2007). They compare the evolutionary success of competing types—i.e. behavioral rules endowed by individuals—in well-defined environmental settings. A population, composed of same-types resident individuals, faces the invasion of a mutant. If the mutant has a fitness advantage over the residents, it invades the population.

Models of social evolution that focus on behavioral strategies may also be used as assessment tools for evaluating the stability of utility functions. For instance, one group selection model developed by Gintis (2000) provides evidence for the evolutionary stability of a behavioral strategy called *Strong reciprocity*.[2] The setting is a repeated public good situation[3] where individuals interact in groups that emerge, last for a period of time, and dissolve in a cyclic way. Two strategies compete against each other: *Strong reciprocity* that always induces its bearers to contribute to the public good, and punish non-contributors at some personal cost; and the *Self-interest* strategy which induces no costly punishment of non-contributors and cooperation only when individually advantageous. Gintis found that *Strong reciprocity* cannot drive *Self-interest* to extinction, but stabilizes at some fraction in the population under a large range of parameter values of his model. This result is about behavioral strategies rather than utility functions. However, it indicates that a pure *Homo economicus* view of human preferences is not satisfactory because in the social environment formalized in Gintis' model, *"Homo economicus"* would never induce the *Strong reciprocity* evolutionary stable strategy. Some authors have suggested instead that the *Strong reciprocity* strategy is induced by an "Inequity aversion" preference (Fehr & Gintis, 2007).

This way of assessing utility functions is limited for the same reasons mentioned in the previous section. As for game theory, these models of social evolution focus on behavioral strategies and reflect specific social interactions. For instance, Gintis' model includes non-trivial constraints, such as a public good situation, only two strategies available, and a complex cycle of group formation and dissolution. These specificities make it difficult to generalize the model's results to other social environments and to make strong claims for or against particular utility functions.

5. Models of social evolution II—focus on utility functions

In order to overcome the limitations of traditional evolutionary methods, some theoreticians have developed new models of social evolution to test the evolutionary stability of utility functions.[4] These models consider utility functions instead of behavioral strategies as evolving traits of interest (Akçay et al., 2009; Alger & Weibull, 2012, 2013; Grund, Waloszek, & Helbing, 2013; Güth & Yaari, 1992; Huck & Oechssler, 1999). They rely on the assumption that utility functions represent genetically or culturally inherited psychological mechanisms shaped by selection.[5] Utility functions define what individuals value, thus provide general instructions for deciding which behavioral strategies to use in different social contexts. Utility functions thus guide individual behavior and are selected on the basis of the welfare payoff obtained by the social partners during their interactions, across social contexts and over their lifetime. A utility function promoting behavioral strategies that are overall advantageous for its bearers will be selected.

More precisely, in models of social evolution focusing on utility functions, the social partners inherit a utility function which they cannot modify during their life. Across social contexts, they always choose the behavioral strategies that lead them to maximize their utility function. Since a utility function can prescribe different behavioral strategies in different social settings, partners may change their behavioral strategies and adjust them to their knowledge of the relevant features of the social context, such as which game they are playing or which type of social partner they are interacting with. In models of social evolution, this behavioral plasticity is formalized in the following way. When social partners

[2] In the literature, *Strong reciprocity* may take various significations (Clavien & Chapuisat, 2013) but in the present context, it refers to a behavioral strategy by which the actors "increase the fitness of unrelated individuals at a cost to themselves." (Gintis, 2000, p. 173).

[3] A public good game is a prisoner's dilemma game (see footnote 1) which counts more than two players.

[4] In the literature, terminology such as "models of preference evolution", or "indirect evolutionary approach" refer to these models for the selection of utility functions.

[5] Models of social evolution do not specify how these traits are inherited from one generation to the next. In theory, inheritance can be genetic or cultural (memetic transmission). However the cultural interpretation is less realistic because transmission of cultural traits entails a high error rate which hinders the selection process.

encounter one another, they are instantaneously capable of seeing the equilibrium situation—e.g. a Nash equilibrium or a Bayesian Nash equilibrium—from which they are not likely to deviate if they repeatedly encounter the same situation. They thus always "rationally" choose the best strategy given their utility function and the features of the social situation. The underlying conjecture is that in the real world, behavioral dynamics operates quickly and individuals are usually fast at finding an equilibrium and behave accordingly during long series of social interactions.

This procedure can be represented with a model of social evolution for two players interactions. A resident population composed of individuals of one type θ competes against mutants of an alternative type τ. θ and τ represent utility functions such as *"Homo economicus"*, *"Degree altruism"* or *"Inequity aversion"*. When two individuals encounter each other, they play a Nash equilibrium, that is, they both choose the behavior that is the best reply to their partner's choice of action, given their own type and their knowledge of the situation. Each pair of choice results in individual fitness defined by the payoff matrix of the game.

After each round of pairwise interactions, utility functions' payoffs can be calculated and compared. This calculus takes into account the payoff—i.e. fitness—received by individuals in the three possible interactions—i.e. when θ meets θ, when θ meets τ, and when τ meets τ—and the probabilities of occurrence of these three forms of interactions. These probabilities usually depend on the proportion of θ and τ individuals in the population, but they can also be influenced by other factors such as a biased probability to meet partners of one's own type. For a utility function to be evolutionary stable, it needs to withstand the invasion of all alternative mutant types, by providing a better fitness payoff to its bearers in all Nash equilibria, and by being the best response to itself.[6] This procedure allows for testing the evolutionary stability of any utility function, θ, against any alternative utility function, τ, in any kind of two players strategic interaction—prisoner's dilemma, hawk-dove game, etc.

In an important paper, Alger and Weibull (2013) analyzed the evolutionary stability of utility functions in two players interactions in an infinite population. One particular feature of their model is the matching process between individuals: it includes an index of assortativity which defines how likely individuals with a given utility function are matched with individuals sharing the same function. Thus, the probability that an individual with utility function θ meets an individual with utility function τ depend on the proportion of individuals with utility functions θ and τ in the population and on the index of assortativity. Another feature of the model is that individuals have limited knowledge about their social partners' utility functions: utility functions are not directly observable, but individuals know the index of assortativity, thus have statistical information about their social partner type. When two individuals encounter each other, they play a Bayesian Nash equilibrium, that is, they both choose the behavior that is the best reply to their partner's choice of action, given their own type and their limited knowledge of the situation.[7] In their study, Alger & Weibull showed the importance of one new family of utility functions, which they label *"Homo moralis"*.

"*Homo moralis*" $\mu_{HM,i}(x_i,x_j) = (1-\kappa)\omega_i(x_i,x_j) + \kappa\omega_i(x_i,x_i)$

This function partitions the actor's utility in two separate terms whose respective weight depends on the κ ($0 \leq \kappa \leq 1$) value: $\omega_i(x_i,x_j)$ represents the payoff obtained by the actor, *i*, assuming that her interaction partner is of a different type, *j*, and $\omega_i(x_i,x_i)$ represents the payoff obtained by the actor, *i*, assuming that her interaction partner is of her own type. Note that this partition in two components only makes sense in social situations where the actors have an imperfect knowledge about their partners' utility functions—this is an important feature of Alger & Weibull's model. Moreover, since individuals behave according to a Bayesian Nash equilibrium, at evolutionary equilibrium—i.e. when one utility function has invaded the population—they maximize their own welfare in the first term of the equation, and contribute to the common welfare in the second term of the equation. This explains why Alger & Weibull label κ the "degree of morality". One may describe it as a degree of interest for the common welfare although the authors describe it as a motivation to follow the Kantian imperative (see below for details).

Interestingly, Alger and Weibull do not emphasize an alternative interpretation of κ. We think that this factor also captures the actor's propensity to choose the action that would produce the best outcome *for herself if everybody—including herself—would behave the same*. Indeed, the second part of the equation contains the assumption that $\mu_j = \mu_i$, and $\omega_j = \omega_i$. Thus, κ may also be summarized as the actor's belief that her social partner adopts the same behavior. Under this reading, high values of κ indicate that the actor believes there is high probability of interacting with a partner of her type.

The *"Homo moralis"* family of functions varies on a single dimension, κ. At one extreme, if $\kappa = 0$, it becomes:

"*Homo economicus*" $\mu_{HE,i}(x_i,x_j) = \omega_i(x_i,x_j)$

At the other extreme, if $\kappa = 1$, the actor always choses the action that maximizes the common welfare assuming that her interaction partner behaves the same. Alger & Weibull link this utility function to a Kantian form of behavior because the actor's choice reflects the famous principle: "act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction." They consequently label the following utility function:

"*Homo kantiansis*" $\mu_{HK,i}(x_i,x_j) = \omega_i(x_i,x_i)$

After exploring the evolutionary stability of a number of functions—including *"Egalitarian altruism"*, *"Degree altruism"*, *"Reciprocal altruism"* and *"Inequity aversion"*—in social contexts with different payoff matrixes and degrees of assortativity, Alger & Weibull found that evolution selects *"Homo moralis"* whenever κ equals the index of assortativity of the matching process. This result leads to the most stable utility function, which the authors label:

"*Homo hamiltoniensis*" $\mu_{HH,i}(x_i,x_j) = (1-r)\omega_i(x_i,x_j) + r\omega_i(x_i,x_i)$

Where *r* is the index of assortativity between two players—i.e. the probability that the actor encounters a social partner of her type.[8]

"Homo hamiltoniensis" drives to extinction competing utility functions in most social contexts—form of game, index of assortativity—*and* cannot be invaded by an alternative utility function

[6] There is one technical difficulty linked to the fact that two or more utility functions may output the same evolutionary stable behavioral strategy in a given social situation, leaving room for drift to select between them. This difficulty can be addressed with a complementary test for *in*stability—how competing utility functions resist to invasion (see Alger & Weibull for details).

[7] In case of multiple possible equilibria, the analysis takes into account the results of the interactions in all equilibria.

[8] Note that *r* is calculated in the same way for *"Homo hamiltoniensis"* as for Hamilton's rule (Hamilton, 1964, 1970).

that does not induce the same behavior than the one generated by *"Homo hamiltoniensis"*. Conversely, Alger & Weibull show analytically that all utility functions that induce other behaviors than those generated by *"Homo hamiltoniensis"* are evolutionary unstable.

It follows that *"Homo economicus"* is evolutionary stable only in social contexts characterized by fully random matching, which means zero assortativity between individuals of the same type. Selection of *"Homo kantiansis"*, on the other hand requires full assortativity. This result is particularly interesting in light of the fact—well-documented in the social sciences—that humans rarely encounter these extreme situations of zero or full assortativity. Many factors, including social class, family bounds, or working environments, raise the probability of interactions between similar individuals, but meeting with "outsiders" is not rare either.

To sum up, a general conclusion from Alger & Weibull's model is that it is advantageous to have a preference for the common welfare ($\kappa > 0$), but only to the extent that we interact with individuals of the same type. Thus, *"Homo hamiltoniensis"* could also be described as "*Homo parochialis"* or "*Homo homophilus*". An alternative—equally acceptable—interpretation is that it is advantageous to have some preference for personal welfare when we are capable of assessing the probability of interacting with individuals of our own type. In light of the latter interpretation, it may seem odd to use the moralistic terminology for describing κ.

Models of social evolution that focus directly on utility functions are particularly informative because they provide general results which help to interpret specific outcomes of models that focus on *behavioral* strategies. For instance, Alger and Weibull's analysis solves the tension between individual and collective interest by making benevolent strategic behavior proportional to the probability of interacting with benevolent partners. This general result explains why Gintis (2000) found that the behavioral strategy *Strong reciprocity* can be stabilized: it is because individuals interact in small groups during a successive number of time periods, which is one way—among others—to introduce assortativity in a system. It should be easy to show that *"Homo hamiltoniensis"* would induce *Strong reciprocity* in the particular setting described by Gintis.

6. Impact on our understanding of human other-oriented psychology

Let us assume that knowledge about the stability—or instability—of utility functions in mathematical models conveys relevant general messages about human social decision-making: that is, humans have evolved decision mechanisms that tend to maximize stable utility functions. Under this assumption, it makes sense to investigate what psychological mechanisms could underlie the preferences contained in utility functions (Section 6.1). Such investigation provides information about which psychological mechanisms may—or may not—have evolved for regulating human social interactions. This investigation will also lead us to argue that psychological altruism evolved towards individuals that generally contributed to the actor's individual fitness (Section 6.2).

6.1. Psychological interpretation of utility functions

Utility functions represent general and stable patterns of individuals' decisions-making. Moreover, when used in models of social evolution, utility functions also correspond to bundles of heritable traits subject to natural selection. Applied to humans, utility functions thereby represent observable expressions—i.e. phenotypes—of inherited psychological mechanisms that guide everyday social choices. The particular features of the proximate psychological mechanisms underlying utility functions remain unclear. To illustrate this point, let us consider how "*Homo hamiltoniensis*" may operate in the real world.

One important feature of this utility function is $\omega_i(x_i,x_j)$, a preference for own welfare. Welfare is defined as an objective and measurable currency. In evolutionary models, this currency correlates with individual fitness—i.e. number of offspring. Adapted to human everyday interactions, objective welfare would correspond to various states of the world that correlate with an increase in individual fitness: gathering economic goods, having a romantic partner, healthy children or long term collaborative relationships would count among these states of the world. Therefore $\omega_i(x_i,x_j)$ may be generated by a web of psychological mechanisms that induce individuals to seek these various states of the world: for example, an attraction towards economic goods, or a capacity to experience love and caring feelings towards romantic partners or towards one's own children.

Another important feature of "*Homo hamiltoniensis*", $\omega_i(x_i,x_i)$, is a preference for the common welfare assuming that the interaction partners share one's own type—i.e. assuming that they are similarly motivated to follow one's own social rules. The weight given to $\omega_i(x_i,x_i)$ depends on the magnitude of r, the index of assortativity between partners. Here, various interpretations are possible and Alger & Weibull's model does not help to discriminate between them.

One line of interpretation would be that humans have developed particular skills for assessing r, the general probability of interacting with individuals having similar social preferences. For this, humans may be particularly receptive to some observable features of their social environments, such as which rules are supported and followed by the surrounding citizens, and whether they match with one's own rules. In addition, humans may have developed the tendency to adjust the strength of their preferences $\omega_i(x_i,x_j)$ and $\omega_i(x_i,x_i)$ to their estimation of r. In particular, the proximate mechanisms underlying $\omega_i(x_i,x_i)$ would be selectively elicited when the actors believe that they interact with individuals of their own type. Such a context-sensitive behavioral choice may be performed in different ways. It could be caused by a rational calculus: human beings understand that they can increase their own welfare by contributing to the common good when r is high. Alternatively, contribution to common welfare may be caused by non-reflexive homophilous mechanisms, including a tendency to trust individuals that share our own social norms, a drive to engage in collective projects with these individuals, or a genuine desire to contribute to the common good or to follow the golden rule—which is a proxy for the Kantian imperative—when interacting with these individuals. Note that these other-oriented mechanisms are homophilous in the sense that their magnitude is proportional to some rough assessment of the probability of interacting with individuals having similar preferences.

Another line of interpretation is that r plays a major role during human ontogenesis but not so much at the adult stage. During ontogenesis, individuals develop psychological character-traits that will influence their decisions during their whole life. In social environments predominantly composed of individuals sharing a homogenous cooperative preference, growing infants may develop character-traits that induce them to engage in collective projects, to act in favor of the common welfare, or to follow the golden rule. Once developed, these character-traits are not context-sensitive, but they will tend to be directed towards individuals sharing the same preferences if migration among social groups is restricted.

It is often difficult to discriminate between alternative potential proximate mechanisms that may generate the same preference. In some cases however, it may be possible to identify the underlying psychological mechanism. For instance, the utility function "*Degree altruism*" with $\alpha > 0$ involves a restriction of $\omega_i(x_i,x_i)$, the actor's

preference for her objective welfare in favor of $\omega_j(x_i,x_i)$, the preference for the partner's welfare. Few psychological mechanisms are likely to produce this pattern of behavior *in the whole range of possible social interactions*; non-discriminative empathic feelings towards social partners or genuine desire to help others are most likely involved.

6.2. Impact on the debate over psychological altruism

Philosophers and psychologists have long debated over humans' capacity to produce altruistic actions (Batson & Shaw, 1991; Butler, 1991; Hutcheson, 2004). The most important condition for psychological—as opposed to biological[9]—altruism is a genuine concern for others as necessary cause for the action.[10] Whether psychological altruism exists, depends on the profound motives—i.e. desires, emotions—underlying humans' actions. It is a question about human psychological mechanisms. The mechanisms that most likely generate altruistic actions are genuine desires or drives to help (Sober & Wilson, 1998), but also basic loving, caring, or empathic feelings (Clavien, 2012; Kitcher, 2011).

Mathematical models designed for testing the stability of utility functions are relevant for investigating the scope of psychological altruism because utility functions inform us about what psychological mechanisms may—or may not—have evolved for regulating human social interactions. One important mathematical result is that some level of preference for own welfare is needed for the stability of a utility function. Indeed utility functions lacking this $\omega_i(x_i,x_j)$ component induce purely cooperative strategies that are easily exploited by less cooperative strategies. This is why "*Homo hamiltoniensis*", the most stable utility function reviewed in this paper, includes a preference for own welfare.

In evolutionary models, objective welfare includes states of the world such as having a romantic partner, or healthy children. Therefore, one might expect humans to have developed specific loving, caring and empathic feelings towards romantic partners and own children—or towards any individual that reliably contributes to the agent's welfare. The evolutionary reason is that the welfare of their children or romantic partners is crucial for the actors' own welfare—i.e. fitness. At the proximate level, actors need not be aware of it[11] and need not calculate the effect of their helping and caring behavior on their own objective welfare. Sober and Wilson (1998, chap 10) make this point very clear with the example of parental care. Self-directed calculus seems to be a less reliable system than an unmediated drive to help one's own children. Since the latter mechanism is more reliable it is more likely to have been selected in the course of evolution.

Another mathematical result of Alger & Weibull's model is that *"Homo hamiltoniensis"* outcompetes all alternative utility functions, including *"Homo economicus"*, *"Inequity aversion"* and all those containing an "altruistic" factor. This result suggests that humans may have genuine concerns for the common good, or for the golden rule, or drives to engage in collective projects. However, these concerns likely evolved in conditions where the actors were amongst other *"Homo hamiltoniensis"* individuals. These concerns may also be homophilous and have coevolved with the capacity to roughly detect the probability to interact with individuals having similar preferences. Note that concerns for the golden rule, the common good, or drives to engage in collective projects do not count as psychologically altruistic because they are not directed towards particular individuals.

To conclude, the evolutionary stability of the *"Homo hamiltoniensis"* utility function suggests that psychological altruism probably evolved because it contributed to the actor's own fitness. It was originally triggered mainly towards individuals increasing the fitness of the actor. Hence, own children, romantic partners and long-term social partners were the most likely beneficiary of altruistic actions. This would explain why the most convincing cases for psychological altruism are to be found in situations such as parental care (Sober & Wilson, 1998), life in kin groups (Kitcher, 2011), and long term reciprocal exchanges (Trivers, 1971). This does not exclude however that humans have recently extended the domain of their altruism towards a broader range of individuals (Kitcher, 2011), as is the case for many other psychological mechanisms (Sperber & Hirschfeld, 2004).

[9] For more on the distinction between psychological and biological altruism, see (Clavien & Chapuisat, 2012; Sober & Wilson, 1998).

[10] Although self-directed motives—including quest for pleasure, power, honor, or avoidance of pain—may coexist with the altruistic motive, the latter is a necessary condition for an altruistic action to come about.

[11] They do not even need to be aware of what counts as their objective welfare.

Acknowledgments

We thank Laurent Lehmann, Justin Garson, Danielle Mersch, Séverine Vuilleumier, Justin Bruner, Grant Ramsey, the Geneva "lgBIG" group, and two anonymous referees for their helpful comments.

References

Akçay, E., Van Cleve, J., Feldman, M. W., & Roughgarden, J. (2009). A theory for the evolution of other-regard integrating proximate and ultimate perspectives. *Proceedings of the National Academy of Sciences, 106*, 19061–19066 (Also available at http://dx.doi.org/10.1073/pnas.0904357106)

Alger, I., & Weibull, J. W. (2012). A generalization of Hamilton's rule—love others how much? *Journal of Theoretical Biology, 299*, 42–54 (Also available at http://dx.doi.org/10.1016/j.jtbi.2011.05.008)

Alger, I., & Weibull, J. W. (2013). Homo Moralis; preference evolution under incomplete information and assortative matching. *Econometrica, 81*, 2269–2302 (Also available at http://dx.doi.org/10.3982/ECTA10637)

Axelrod, R. M. (1984). *The evolution of cooperation*. New York: Basic Books.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory, 2*, 107–122.

Becker, G. S. (1976). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature, 14*, 817–826 (Also available at http://dx.doi.org/10.2307/2722629)

Binmore, K. G. (2005). *Natural justice*. New York: Oxford University Press.

Butler, J. (1991). Fifteen sermons. In D. D. Raphael (Ed.), *British moralists, 1650–1800: Selected and edited with comparative notes and analytical index* (pp. 325–377). Oxford: Clarendon Press.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton, NJ: Princeton University Press.

Clavien, C. (2012). Altruistic emotional motivation: An argument in favour of psychological altruism. In K. Plaisance, & T. Reydon (Eds.), *Philosophy of behavioral biology* (pp. 275–296). Dordrecht: Boston Studies in Philosophy of Science, Springer Press (Also available at http://dx.doi.org/10.1007/978-94-007-1951-4_13)

Clavien, C., & Chapuisat, M. (2012). *Altruism – A philosophical analysis* (pp. 1–6). Chichester: eLS JohnWiley & Sons, Ltd (Also available at http://dx.doi.org/10.1002/9780470015902.a0003442.pub2)

Clavien, C., & Chapuisat, M. (2013). Altruism across disciplines: One word, multiple meanings. *Biology and Philosophy, 28*, 125–140 (Also available at http://dx.doi.org/10.1007/s10539-012-9317-3)

Davies, N., Krebs, J. R., & West, S. (2012). *An introduction to behavioral ecology*. Oxford: Wiley-Blackwell.

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology, 103*, 962–1023 (Also available at http://dx.doi.org/10.1086/231294)

Fehr, E., & Gintis, H. (2007). Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology, 33*, 43–64. Also available at http://dx.doi.org/10.1146/annurev.soc.33.040406.131812.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 817–868 (Also available at http://dx.doi.org/10.1162/003355399556151)

Frank, S. A. (1998). *Foundations of social evolution*. Princeton, NJ: Princeton University Press.

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox*. Cambridge, Mass: MIT Press.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*, 169–179 (Also available at http://dx.doi.org/10.1006/jtbi.2000.2111)

Grund, T., Waloszek, C., & Helbing, D. (2013). How natural selection can create both self- and other-regarding preferences, and networked minds. *Scientific Reports, 3* (Also available at http://dx.doi.org/10.1038/srep01480)

Guala, F. (2005). *The methodology of experimental economics*. Cambridge; New York: Cambridge University Press.

Güth, W., & Yaari, M. (1992). An evolutionary approach to explain reciprocal behavior in a simple strategic game. In U. Witt (Ed.), *Explaining process and change — approaches to evolutionary economics*. Ann Arbor: University of Michigan Press.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology, 7*, 1–52.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature, 288*, 1218–1220.

Huck, S., & Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior, 28*, 13–24 (Also available at http://dx.doi.org/10.1006/game.1998.0691)

Hutcheson, F. (2004, 1st ed. 1725). *An inquiry into the original of our ideas of beauty and virtue*. Indianapolis, Ind.: Liberty Fund.

Kincaid, H., & Ross, D. (2009). *The Oxford handbook of philosophy of economics*. Oxford; New York: Oxford University Press.

Kirchgässner, G. (2008). *Homo oeconomicus: The economic model of behaviour and its applications in economics and other social sciences*. New York: Springer.

Kitcher, P. (2011). *The ethical project*. Cambridge, Mass: Harvard University Press.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics, 1*, 593–622 (Also available at http://dx.doi.org/10.1006/redy.1998.0023)

MacElreath, R., & Boyd, R. (2007). *Mathematical models of social evolution: A guide for the perplexed*. Chicago: The Univ. of Chicago Press.

Mayr, U., Harbaugh, W. T., & Tankersley, D. (2009). Neuroeconomics of charitable giving and philantropy. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (1st ed.). (pp. 303–320) Amsterdam; Boston: Elsevier Academic Press.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America, 36*, 48–49 (Also available at http://dx.doi.org/10.1073/Pnas.36.1.48)

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *American Psychologist, 61*(6), 622–631 (Also available at http://dx.doi.org/10.1037/0003-066X.61.6.622)

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel* (pp. 114–146). Dordrecht: Reidel.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge: Harvard University Press.

Smith, V. L. (1962). An experimental-study of competitive market behavior. *Journal of Political Economy, 70*, 111–137 (Also available at http://dx.doi.org/10.1086/258609)

Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, Mass: Harvard University Press.

Sperber, D., & Hirschfeld, L. A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences, 8*(1), 40–46 (Also available at http://dx.doi.org/10.1016/j.tics.2003.11.002)

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

Weibull, J. W. (1995). *Evolutionary game theory*. Cambridge, Mass: MIT Press.