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.

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/

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: $\mu_i = \omega_i$.

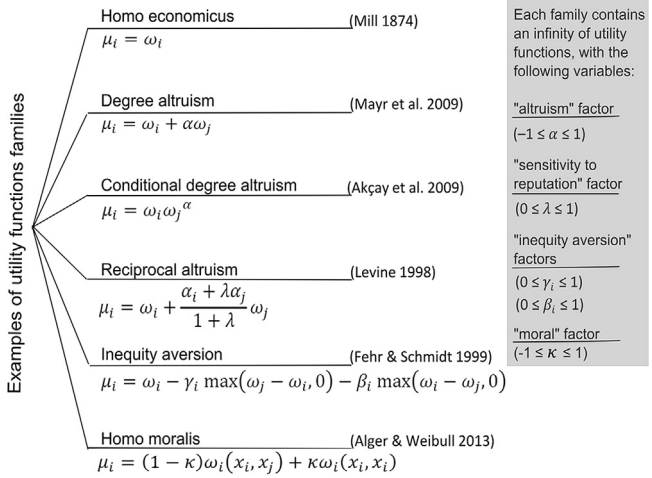

Fig. 1. Examples of utility functions families.

"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.