MORALITY
- evolutionary foundations and policy implications -

Ingela Alger* and Jörgen W. Weibull†

June 3, 2016‡

Abstract. Since the publication of Adam Smith's *Wealth of Nations*, it has been customary among economists to presume that economic agents are purely self-interested. However, research in experimental and behavioral economics has shown that human motivation is more complex and that observed behavior is better explained by additional motivational factors such as a concern for fairness, social welfare etc. Instead of hypothesizing such utility functions and testing them experimentally, we have carried out theoretical investigations into the evolutionary foundations of human motivation (Alger and Weibull 2013, 2016). We found that natural selection, in starkly simplified but mathematically well-structured environments, favors preferences that combine self-interest with morality in line with Immanuel Kant's categorical imperative. Roughly speaking the moral component evaluates one's own action in terms of what would happen, if, hypothetically, this action were adopted by others. Such moral preferences have important implications for economic behavior. They motivate individuals to contribute to public goods, to give fair offers when they could get away with cheap offers, and to contribute to social institutions and act in environmentally friendly ways even if their individual impact is negligible.

"Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." [Immanuel Kant, *Groundwork of the Metaphysics of Morals*, 1785]

"One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die." [Charles Darwin, *On the Origin of Species*, 1859]

*Toulouse School of Economics, CNRS, and Institute for Advanced Study in Toulouse

†Stockholm School of Economics, Institute for Advanced Study in Toulouse, KTH Royal Institute of Technology, and Toulouse School of Economics

‡This manuscript was prepared for the conference "The State of Economics, The State of the World", to be held at the World Bank in Washington D.C. on June 8-9, 2016. The authors thank Daniel Chen, Jean-François Laslier, Assar Lindbeck, Erik Mohlin, Paul Seabright, Jean Tirole, Nicolas Treich, Yu Wen, and Peter Wikman for valuable comments and suggestions.

1. Introduction

The academic discipline of economics has over many years provided policy makers all over the world with a powerful toolbox. Conceptual, philosophical and methodological disagreements are relatively rare and the discipline is not torn by fights between disparate schools of thought. Whether this monolithic character of the field is a sign of strength or weakness is not easy to say, but this methodological unity and power has, arguably, given the discipline great influence on policy. The strong methodological core of economics, in the 1950s-1980s epitomized by general equilibrium theory, and later incorporating game theory, has enabled positive and normative analysis of a wide range of economic and social issues.

So what, more exactly, does this core consist of? In a nutshell, it has two main components. The first is that it views economic agents—who may be individuals, households, firms, or organizations—as goal-oriented; as if they each had some goal function that they strive to maximize under the constraints they face, the information they have, and given their beliefs about relevant aspects of the world they live in. The second component is that interactions between these economic agents are taken to meet certain consistency requirements, formalized as equilibria, that is collections of action plans, one for each agent, such that no agent can unilaterally improve the expected value of her goal function (usually profit or utility).

Both components can, and have been, contested. Individuals may not be so systematic and consistent, and interactions may be chaotic and volatile. Having a theoretically well-founded and empirically accurate understanding of human motivation is, arguably, in any case of utmost relevance for policy.

Among the more noticeable new methodological developments in economics is the emergence of behavioral and experimental economics, where the first strand endows economic agents with richer motivations than what is traditional in economics, usually in the form of pro-social or other-regarding preferences. The second strand of the literature tests such models, old and new, in controlled laboratory experiments and in randomized field experiments. The external validity of laboratory experiments can be questioned, and field experiments may depend on local and historical factors with little generality, but this development of the discipline of economics, towards an empirically founded science, appears as essentially very healthy. It was not long ago that economics was thought of as similar to meteorology and astronomy in that all it could do was to observe what is happening, without possibility to experiment. Moving away from mere observation of data that happen to come about to carefully designed controlled experiments, reminds of the way Galileo Galilei once lead the way from Aristotelean scholastic discourse to modern science.

While behavioral and experimental economics no doubt will improve the predictive power and the usefulness of economics, further improvements could certainly be made if the underlying factors that shape human motivation were better understood. The literature on the evolutionary foundations of human motivation aims at providing such understanding, by asking: What preferences should humans be expected to have if preferences and moral values are transmitted in society, from generation to generation? If certain pro-social or anti-social preferences, or moral values, give their carriers on average better material outcomes than other preferences or values (all else being equal), then one would expect the former to spread more in the population (be it by biological or cultural mechanisms). Our goal in this essay is to discuss a recent theoretical result concerning such evolutionary selection of preferences and moral values, and to examine its implications for a range of social and economic issues.

Milton Friedman (1953) claimed that "unless the behavior of businessmen in some way or other approximated behavior consistent with the maximization of returns, it seems unlikely that they would remain in business for long". In a similar vein, one may claim that unless the behavior of an individual is consistent with the maximization of own material payoffs, other, materially more successful behaviors will take over in the interacting population.[1] Game theorists have shown that this claim is theoretically valid when *(i)* the population at hand is very large, *(ii)* interacting individuals do not know each other's goal functions, and *(iii)* interactions are perfectly random in the sense that each encounter is just as likely (Ok and Vega-Redondo, 2001; Dekel, Ely, and Yilankaya, 2007).

In reality, however, populations are not always large, and interacting individuals sometimes know or learn about each other's preferences—think, for instance, of the great number of interactions that take place within families or small communities. It has been shown that in such settings, preferences or goal functions can usually serve as effective commitment devices and evolution will almost always favor goal functions that differ from own material payoffs (Banerjee and Weibull, 1995, and Heifetz, Shannon, and Spiegel, 2007). Furthermore—and this is what we will focus on here—encounters are only rarely perfectly random; geographic location, language, culture and religion often have an impact on the likelihood of specific encounters. For example, business partners may know each other from college, and neighbors may have chosen to live in the same place because they share socioeconomic or cultural background and/or location preferences etc. In such structured populations, some encounters are more likely than others, even if the overall population is large. In two recent theoretical studies (Alger and Weibull, 2013, 2016), we show that such *assortative matching* makes evolution favor individuals who are not purely self-interested but who attach some value to "doing the right thing", even though the population is large and interacting individuals do not know each other's preferences. This, for us initially surprising finding suggests an evolutionary founda-

[1] The seminal articles in this literature are Frank (1987) and Güth and Yaari (1992).

tion for a psychologically plausible form of morality, in line with Immanuel Kant's categorical imperative.

In the next section we describe this novel class of preferences and their evolutionary foundations. In Section 3, we discuss the implications of such preferences for a number of much studied social and economic behavior and policy issues, ranging from public goods provision, concern for the environment, as well as informal lending. Section 4 concludes.

2. Evolution and Kantian morality

Imagine a population that has evolved for many generations in a stationary environment, and that in each generation in this population, individuals engage in some interaction, the same one in each generation. For instance, in a population of self-subsistence farmers, the interaction could be team-work in the fields, the extraction of resources from a commonly owned lake or piece of land, lending activities, or the maintenance of institutions. In Alger and Weibull (2013, 2016), we propose a theoretical model of precisely such populations. We then formalize the interaction by assuming that individuals are now and then randomly matched into groups of arbitrary (but fixed and given) size n to interact with each other within the group. (There are no interactions between groups and hence no group selection takes place.). The interaction may involve elements of cooperation and/or conflict, asymmetric information, repetition or interaction of arbitrary duration, possibility of helping, rewarding and/or punishing others etc. There are essentially only two restrictions imposed on the interaction. First, the material payoff consequences to a participant depend only on the participant's own actions and on some aggregate of other group members' actions (not on who of them does what). In game theory such interactions are called *aggregative* games. Examples are market competition where only competitors' aggregate output or lowest price matter, contributions to public goods where only the sum of others' contributions matter, some environmental externalities etc. Second, the material payoff function is the same for all individuals.

In our evolutionary stability analysis we assume that each individual has some utility function that he or she seeks to maximize. We then ask what kind of utility function, if any, would be favored by natural selection if each individual's preferences are his or her private information, and groups are formed according to some given random matching process. Given a symmetric aggregative interaction, defined in terms of the material payoff consequences for the interacting individuals, and given a random matching process, we analyze which utility functions, if any, are evolutionarily stable in the sense that, if almost all individuals in the population have such preferences, these individuals would materially outperform individuals with other preferences? Thus, the material payoffs are taken to be the drivers of evolution.

We define *Homo oeconomicus* as individuals who always seek to maximize their own material payoff.[2]

This approach is a generalization of the work of Maynard Smith and Price (1973), from the notion of an evolutionarily stable strategy, or ESS, to that of an *evolutionarily stable utility function.* A major challenge arises with this generalization. In any *population state*—the preference distribution in the population—there may be multiple equilibrium behaviors, and hence several possible material payoff allocations. We define a utility function to be *evolutionarily stable against another utility function* if in every population state where the latter utility function is rare, individuals equipped with the former utility function outperform those with the latter in terms of the resulting material payoffs in *all* equilibria.[3] Conversely, a utility function is *evolutionarily unstable* if there exists another utility function such that, no matter how small its population share, there is some equilibrium in which the latter utility function materially outperforms the former. In both definitions, the test scenario is to let in a small population share of "mutants", who may be migrants or carriers of spontaneously and randomly arising alternative utility functions, into the population of *incumbents* or *residents.* We impose minimal constraints on the nature of potential utility functions. They are not required to take any particular parametric form or even to depend on the material payoffs. Hence, individuals may be selfish, altruistic, spiteful, fairness-minded, inequity averse, environmentalists or moralists, etc. Our only assumption is that each individual's utility function is continuous in all group members' courses of action.

A second key feature of our approach is that it allows the random matching, when groups are formed, to be *assortative.* First, while distance is not explicitly modeled here, geographic, cultural, linguistic and socioeconomic distance imposes (literal or metaphoric) transportation costs, which imply that (1) individuals tend to interact more with individuals in their (geographic, cultural, linguistic or socioeconomic) vicinity,[4] and (2) cultural or genetic transmission of types (say, behavior patterns, preferences or moral values) from one generation to the next also has a tendency to take place in the vicinity of where the type originated.[5]

[2] Some writers define *Homo oeconomicus*, or "economic man" more generally as an individual who always acts in accordance with some goal function, whether this be pure self-interest or not. All agents in the present study are varieties of *Homo oeconomicus* in this broad sense.

[3] By "equilibrium" we mean Bayesian Nash equilibrium under incomplete information.

[4] Homophily has been documented by sociologists (e.g., McPherson, Smith-Lovin, and Cook, 2001, and Ruef, Aldrich, and Carter, 2003) and economists (e.g., Currarini, Jackson, and Pin, 2009, 2010).

[5] In biology, the concept of assortativity is known as *relatedness*, and the propensity to interact with individuals locally is nicely captured in the infinite island model, originally due to Wright (1931). Hamilton (1964) provided a first formalization of what is now known as Hamilton's rule: that evolution will select for behaviors whereby the external effects on others are internalized at a rate provided by the relatedness (see also Dawkins, 1976, for a popular account of this idea, as well as Rousset, 2004, for a comprehensive treatment).

Taken together, these two tendencies may generate the assortativity that we here allow for. We formalize the assortativity of the random matching process in terms of a vector we call the *assortativity profile*. This is the probability vector for the events that none, some, or all the individuals in a (vanishingly rare) mutant's group also are mutants, i.e., that the number of other mutants is k, for $k = 0, \ldots, n-1$.[6]

Our analysis delivers two main results. First, although we impose virtually no restrictions on permissible utility functions, evolution favors a particular class of utility functions that we call *Homo moralis*. Individuals with preferences in this class attach some weight to their own material payoff but also to what can be interpreted as a probabilistically generalized version of Kantian morality. In his *Grundlegung zur Metaphysik der Sitten* (1785), Immanuel Kant wrote "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." Similarly, *Homo moralis* attaches some weight to the goal of "acting according to that maxim whereby you can, at the same time, will that others should do likewise with some probability." More precisely, a *Homo moralis* individual in a group of arbitrary size n maximizes a weighted average of equally many terms, indexed $k = 0, \ldots, n-1$, where each term is the material payoff that she would obtain if, hypothetically, she could replace the strategies of so many other individuals in the group by her strategy. We call the vector of these probability weights the individual's *morality profile*.

The class of *Homo moralis* preferences has two familiar extremes: *Homo oeconomicus*, who considers only her own material payoff, and *Homo kantiensis*, who considers only the material payoff that she would obtain if all others were to act like she does. In between these two extremes there is a whole range of *Homo moralis* preferences with different morality profiles whereby an individual examines what would happen if some but not all the others were to act like him- or herself. *Homo moralis* partly evaluates her own actions in this (hypothetical) probabilistic Kantian sense. In other words, she is to some extent concerned with the morality of her own acting, irrespective of what others do. She asks herself, before taking her action, what action would she prefer if, hypothetically, also others would (probabilistically) choose the same action in her situation? We show that *Homo moralis* with morality profile identical with the assortativity profile is evolutionarily stable.

Our second main result is that any preferences that are behaviorally distinct from those of *Homo moralis* (with the stable morality profile) are evolutionarily unstable. Hence, although we made no parametric assumption about utility functions, it appears that natural selection

In an article on the evolution of behaviors in interactions between siblings, Bergstrom (1995) was probably the first to bring Hamilton's rule into the economics literature.

[6] This generalizes Bergstrom's (2003) definition of the *index of assortativity* for pairwise encounters. See also Bergstrom (2012) and Alger and Weibull (2013) for further discussions of assortativity under pairwise matchings.

strongly favors the utility function of *Homo moralis.* In particular, our results imply that *Homo oeconomicus*—pure material self-interest—is evolutionarily unstable under any random matching process with positive assortativity. Rare mutants may indeed garner a higher material payoff than *Homo oeconomicus*, on average, by behaving somewhat pro-socially, because when there is positive assortativity the benefits of this pro-social behavior is sometimes bestowed on other mutants, whereas the residents almost never benefit from it.

The intuition behind our first result, the evolutionary stability of *Homo moralis*, is not based on group selection, an old argument (appearing already in Charles Darwin's writings; see also Alexander, 1987) that essentially says that evolution will lead to behaviors that enhance the survival and reproduction of the group. Quite on the contrary; the intuition behind our result is that natural selection will lead to utility functions that *preempt entry* into the population, in the sense that the best a potential mutant can do, if striving for material payoff, is to mimic the residents.

In sum, our analysis provides an evolutionary foundation for people to have an innate concern for what would happen if others acted in the same way as they do. We interpret this concern as being of a moral nature, for it involves an examination of the virtue of alternative behaviors, should those behaviors be undertaken by others, rather than an examination of the actual consequences of one's behavior on the others in the situation at hand. As such, it may be that our work provides an evolutionary foundation for what social psychologist Jonathan Haidt calls "elevation" (Haidt, 2003, p.275): "Elevation is elicited by acts of virtue or moral beauty; it causes warm, open feelings ('dilation?') in the chest; and it motivates people to behave more virtuously themselves (to 'covenant to copy the fair example')".[7]

Homo moralis is easily defined for pairwise interactions, $n = 2$. Let $\pi(x, y)$ denote the material payoff to an individual who plays strategy x when the opponent plays strategy y. Then the utility function of *Homo moralis* is

$$U_\kappa(x, y) = (1 - \kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x),$$

where $0 \leq \kappa \leq 1$ is the individual's *degree of morality.* The two extreme degrees of morality represent *Homo oeconomicus* ($\kappa = 0$) and *Homo kantiensis* ($\kappa = 1$), respectively, and intermediate degrees of morality corresponding to individuals who attach *some* weight to own material payoff, $\pi(x, y)$, and *some* weight to "the right thing to do if everyone were to choose the same behavior", $\pi(x, x)$.

For $n > 2$ the precise definition of *Homo moralis* is fairly involved,[8] but it is analytically

[7] See also Greene et al. (2001) and Nichols (2004) for further evidence and discussion of the links between emotions and moral norms.

[8] The general definition of *Homo moralis* is given in an appendix at the end of this essay.

straight-forward in the special case when the random matching is such that the types of any *other* two group members are statistically independent given the member's own type. The morality profile is then a binomial distribution, and the utility function of a *Homo moralis* individual i is the expected value of i's material payoff if, hypothetically, other members of the group would randomly and statistically independently switch to use i's strategy with probability κ, which is then i's degree of morality. At one end of the interval of such *Homo moralis*, $\kappa = 0$, we find *Homo oeconomicus* while at the other end, $\kappa = 1$, we find *Homo kantiensis.* Moreover, in large groups, the share of mutants in a mutant's group is, by the deMoivre-Laplace Theorem, approximately normally distributed with mean value κ and variance $\kappa(1-\kappa)/(n-1)$. Hence, the share of other mutants is then almost deterministic and equal to κ. An evolutionarily stable *Homo moralis* then acts as if she hypothetically assumed that her behavior were to become, if not a "universal law", a "random law" applying to a randomly sampled share of size κ out of her group's other members.

It is worth noting that the utility function of *Homo moralis* differ sharply from any utility function that only depends on the payoffs to all participants, such as altruism, inequity aversion, or a concern for social efficiency. We will discuss this in detail at the end of next section.

While morality and ethics have been discussed at great length by many economists, including Smith (1759), Edgeworth (1881), Arrow (1973), Sen (1977), and Harsanyi (1992), to mention a few, the family of *Homo moralis* preferences has, to the best of our knowledge, not been studied, or even known, before.[9] So how do individuals of this "new species" behave? What are the policy implications if economists' models are populated not by the familiar but extreme variant *Homo oeconomicus* but by *Homo moralis* of different morality profiles?

3. Kantian morality and economics

Economists' policy advice usually relies on models in which individuals have *Homo oeconomicus* preferences. Here we examine how individuals with *Homo moralis* preferences may behave and make a comparison with the behavior of *Homo oeconomicus.* Such an exploration is a major undertaking and we will here merely scratch the surface at some places, by studying but a few examples.

3.1. Personal loans. In many developing countries large fractions of the populations are still shut out form formal credit markets, see e.g. Kendall et al. (2010). Then informal lending,

[9] An exception is Bergstrom (1995) who shows that evolutionary stability of strategies in interactions between siblings induces behavior which he calls "semi-Kantian." For a discussion of several ethical principles in relation to strategy evolution, see Bergstrom (2009).

in the form of not legally binding loans between individuals, can sometimes be enforced by the threat of future non-renewal of lending (Ghosh and Ray, 2016) and/or social disapproval. Moreover, evidence from laboratory experiments suggests that such lending may even take place in one-shot interactions (Charness and Dufwenberg, 2006). Here we study the behavior of *Homo moralis* in a one-shot personal loan.

Consider two individuals, A and B, where B needs a temporary loan from A, who has the money needed. They cannot write a legally binding contract, so the loan has to be given on trust. Will A lend the money, and, if given the loan, will B pay it back? What is the morally "right thing to do" in each of the two roles? Should A lend? Should B pay back?

To represent this situation in the simplest possible way, let the material payoff (or personal utility) to B be normalized to zero if he is not given the loan, $b > 0$ if he is given the loan and pays it back, and $b + 1$ if he is given the loan and does not pay it back. Likewise, A's material payoff (or personal utility) if not lending is zero, $-a$ if lending and the loan is paid back, and $-a - 1$ if lending and the loan is not paid back. Assume that $a < b$, so A's opportunity cost of lending is smaller than the benefit to B from getting the loan. In this simple interaction, A has two pure *strategies*, to lend (L) or not lend (N). Likewise, a pure strategy for B is to either honor the loan and pay it back (H), or defect by not paying it back (D). The material payoffs to both parties are summarized in the table below, with A choosing row and B column.

	H	D
L	$-a, b$	$-a - 1, b + 1$
N	$0, 0$	$0, 0$

If the opportunity cost a of lending is positive and A is a *Homo oeconomicus*, caring (only) about her own material payoffs, she will not give the loan, so in this case both parties will end up with material payoff zero.[10] If also B is a *Homo oeconomicus*, then he would not pay back the loan if given, so in this case no loan would be given, even if it involved some interest (that is, even if a were negative). Thus, in a society consisting of *Homo oeconomicus*, informal lending cannot arise spontaneously between two parties, unless they are in a long-term relationship in which lending needs and opportunities to lend arise repeatedly, in which case there may exist equilibria with lending if the parties are sufficiently patient (along with equilibria with no lending). A policy implication of this observation is that for one-shot informal lending opportunities to be seized in a society consisting of *Homo oeconomicus*, a formal enforcement institution, such as legally binding contracts, is necessary, and this may be needed even in repeated interactions (to avoid the socially inefficient equilibria).

[10] For *Homo oeconomicus*, strategy L is (weakly) dominated by strategy N when $a > 0$ ($a = 0$).

What would a *Homo moralis* do in the above one-shot lending situation? In order to apply the *Homo moralis* machinery, we need to embed the above situation in a scenario that renders the interaction *a priori* symmetric. A canonical way to do this is to initially cover the interaction under a veil of ignorance as to who will be in what role, that of A or that of B.[11]

Consider, thus, two individuals who will meet in a situation as just described, where each individual is just as likely to be in either role A or role B. Behind the veil of ignorance, a strategy for an individual is now what to do in each of the two roles. Each individual now has four (pure) strategies, for example, to lend (L) in role A and to honor the loan (H) in role B. The resulting material payoff matrix is given below. Each entry is the expected material payoff (personal utility) to an individual who chooses the associated row strategy when the other individual chooses the associated column strategy.

	LH	LD	NH	ND
LH	$(b-a)/2$	$(b-a-1)/2$	$-a/2$	$-(a+1)/2$
LD	$(b-a+1)/2$	$(b-a)/2$	$-a/2$	$-(a+1)/2$
NH	$b/2$	$b/2$	0	0
ND	$(b+1)/2$	$(b+1)/2$	0	0

In terms of material payoffs, the strategy ND, to not lend and not pay back, strictly dominates the two lending strategies, and weakly dominates strategy NH, to not lend but honor a loan. And this is precisely what *Homo oeconomicus* would do. Both parties would then obtain zero material payoff. By contrast, in a population of *Homo kantiensis*, who attach weight only to what would happen if everybody behaved as they do, we would either see everybody use strategy LH or everybody use strategy LD. In the first case, the individual in role A always lends and the individual in role B always pays back.

In the other case, the individual in role A always lends but individual B never pays back. At first sight, this may seem immoral. However, we are here mislead by the wording, because if this was the behavior pattern established in a society, then it could just as well be interpreted as giving to the needy. This is just another convention, no less moral than the first (to lend and pay back). And indeed, this is what we see in everyday life when the stakes are small. If you go for coffee with a friend or acquaintance who has no cash on hand, you usually do not lend money for the coffee, you treat the other person.

By contrast, when stakes are high, say, money needed for the purchase of an expensive item, lending and paying back is more common. And, indeed, this would also be the case in the present example if the individuals were risk averse and equally rich or poor. In such a case, the lender's loss if the loan is not paid back, would be larger in utility terms, than the

[11] The veil of ignorance is due to Vickrey (1945), Harsanyi (1953), and Rawls (1957).

lender's gain from not paying back. Instead of writing the material payoff to a party under strategy profile (LH,LH) as $(b-a)/2$ we would have to write it as $[u(\omega+b)+u(\omega-a)]/2$. Likewise, the material payoff under strategy profile (LD,LD) would have to be written as $[u(\omega+b+1)+u(\omega-a-1)]/2$. Clearly the latter is smaller than the first, for any concave utility function u, so in this case the convention LH would Pareto dominate the convention LD, and *Homo kantiensis* would act accordingly.

In sum, *Homo moralis* preferences may sustain materially beneficial behaviors even in the absence of formal institutions or repeated interactions.

3.2. Trust. There is variation across countries in the extent to which people are trusting, and trust is correlated with economic growth (Algan and Cahuc, 2010). In economics, the so-called *trust game* has been used extensively in controlled laboratory experiments as a way to measure trust and trustworthiness in different countries and cultures. This literature was pioneered by Berg et al. (1995) and has received a lot of attention among behavioral economists and experimentalists. The trust game is succinctly described by Cesarini et al. (2008):

> "Many mutually beneficial transactions involve an element of interpersonal trust and may fail to materialize in the absence of an expectation that trust will be reciprocated. The prevalence of trust in a society has therefore been assigned primacy in a number of domains, for instance empirical and theoretical studies of economic growth. In recent years, the trust game has emerged as a favorite instrument to elicit an individual's interpersonal trust and willingness to reciprocate trust. More generally, the game has been widely used to study cooperative behavior. In a trust game, an individual (the investor) decides how much money out of an initial endowment to send to another subject (the trustee). The sent amount is then multiplied by some factor, usually three, and the trustee decides how much of the money received to send back to the investor. The standard game-theoretic prediction for a single anonymous interaction between two purely self-interested individuals is for the investor to send nothing, rationally anticipating that the trustee will not reciprocate. Yet, experiments consistently show that cooperation flourishes in the trust game; the average investor sends a significant share of her endowment, and most trustees reciprocate. A voluminous body of theoretical and experimental work examines the mechanisms through which natural selection can favor cooperation, and proposed mechanisms include kin selection, reciprocity, indirect reciprocity, and group selection. These models offer different accounts for the ultimate explanation for the existence of cooperation and also

> generate different predictions about genotypic variation in equilibrium." (op. cit., p. 3721)

What will *Homo moralis* do in such an interaction? Consider a two-player game, a simple strategic interaction between two individuals. With equal chance, one of them is offered an endowment and an investment opportunity, in which trusting the other and honoring such trust is beneficial for both. A *strategy* for any one of the two parties then has two components. First, if given the endowment and investment opportunity, what share s of it to invest in the other party. Second, if not given the investment opportunity, what "payback rule" p to use, where such a payback rule prescribes for any invested share t chosen by the other party what share, p, of the (multiplied and received) amount received to pay back. In the standard formulation of the trust game (see citation above), the expected material payoff to strategy $x = (s, p)$ when used against strategy $y = (t, q)$ is

$$\pi(x, y) = \frac{1}{2} v(1 - s + 3qs) + \frac{1}{2} v(3(1 - p)t),$$

where v is their hedonic utility from own wealth (the marginal hedonic utility is taken to be positive but decreasing). In an interaction between two *Homo oeconomicus*, no party is trustworthy (they will both choose $p = q = 0$). Thus, if each party know the other's type, no investment is made in equilibrium ($t = s = 0$). The resulting expected material payoff to each party is $v(1)/2$, the probability of being given the initial endowment times the utility from keeping it. If instead both parties were *Homo kantiensis*, then they would each invest all the money if given the opportunity ($t = s = 1$), and return *half* the (received and multiplied) amount ($p = q = 0.5$). The resulting expected material payoff to each party is then $v(1.5)$, much higher than what *Homo oeconomicus* obtains.

Full morality is not necessary in order to induce full investment. In a pair of equally moral *Homo moralis*, full investment ($s = t = 1$) obtains in equilibrium for any sufficiently high degree of morality, although as soon as morality is less than full ($\kappa < 1$), the trustee pays back less than half the gross returns from investment, in which case the trustee ends up being better off than the investor. As the degree of morality κ falls, the amount paid back decreases, and it eventually falls short of the amount originally invested, in which case the investor makes a material loss; nonetheless, morality makes the investor accept this loss and invest anyway, up to some point.[12] Indeed, for sufficiently low degrees of morality the investor invests less than his full endowment, and eventually, when morality drops below a certain level, he invests nothing.

[12] To see this, note that the derivative of $U_\kappa(x, y)$, where $x = (s, p)$, $y = (t, q)$, with respect to s, and evalutated when $t = s = 1$, is positive even for $p < 1/3$ for $\kappa < 1$ large enough.

The situation captured by the trust game is in fact similar to the one studied above; it can be interpreted as a situation of informal lending. The present model of trust is more fine-grained than the above model of personal loans, as it allows the loan/investment as well as the amount paid back to vary on a continuous scale. However, the policy implications are the same. Morality may be sufficient for efficiency-enhancing one-shot informal lending to take place, even in situations where the lender/truster ends up making a loss.

3.3. Public goods. A host of situations that are important for economic growth may be represented as situations in which people can make voluntary contributions towards a public good, including the generation and dissemination of knowledge, and institution building. We examine the behavior of *homo moralis* in a community of n members, each of whom is in a position to make a voluntary contribution to a public good (the contribution may be monetary or in kind). Suppose, then, that i obtains material payoff

$$\pi\left(x_i,\boldsymbol{y}\right)=B\left(x_i+\sum\nolimits_j y_j\right)-C\left(x_i\right)$$

if she makes the contribution x_i and the sum of the contributions from the other community members is $\sum y_j$. Here B is a production function for the public good and C a cost function for a contributing individual—representing foregone private consumption, income, or leisure. We take the marginal cost of making a contribution to be increasing and the marginal benefit of the aggregate contribution to be decreasing.

In a community of *Homo oeconomicus*, the first-order condition for the unique Nash equilibrium contribution, $\hat{x}_0$, writes

$$B'\left(n\hat{x}_0\right)=C'\left(\hat{x}_0\right). \tag{1}$$

It follows that in communities with more members, each individual contributes less. The intuition is that if all contributions were to remain unchanged then the marginal benefit from each contribution would fall. Thus, each individual will have a weaker incentive to contribute. However, the total contribution, $\hat{X}_0=n\hat{x}_0$, increases with the size of the community.[13]

By contrast, the socially optimal individual contribution, x^*, may well be increasing in n. To see this, first note that maximization of the sum of all members' material payoffs requires that the marginal social benefit equals the marginal social cost,

$$nB'\left(nx^*\right)=C'\left(x^*\right). \tag{2}$$

Second, consider a conventional production function of the power form $B\left(X\right)=X^a$, where $0<a<1$. Then it is easily verified that the left-hand side in (2) is increasing in n (given

[13] To see this, note that (1) can equivalently be written as $B'(\hat{X}_0)=C'(\hat{X}_0/n)$, where B' is decreasing and C' increasing.

x^*), from which it follows that the socially optimal individual contribution x^* is increasing in n. As a consequence, free-riding—the tendency for people to under-provide public goods—is exacerbated when group size increases.

Suppose now instead that everyone in the community is a *Homo moralis* with the same degree of morality $\kappa \in [0,1]$. Then their unique individual equilibrium contribution, $\hat{x}_\kappa$, can be shown to satisfy

$$[1 + (n-1)\kappa] \cdot B'(n\hat{x}_\kappa) = C'(\hat{x}_\kappa).$$

For any positive degree of morality, group size has two counter-acting effects on the individual contribution. The negative effect is, as before, due to the decreasing marginal productivity. The positive effect is that in larger groups each individual's contribution benefits a larger number of individuals. The "right thing to do", as the group increases, is thus to increase one's contribution. The positive effect may outweigh the negative.

To see this, consider again the conventional production function used above, and note that for purely Kantian individuals ($\kappa = 1$) the individual contribution always increases with n. For intermediate values of κ, the individual contribution decreases with n, when this is small, but increases with n when this is large. See Figure 1 below which the equilibrium contribution of Homo moralis with degree of morality κ as a function of community size n, with higher curves for higher degrees of morality (for $\kappa = 0$, 0.25, 0.5, 0.75, and 1).

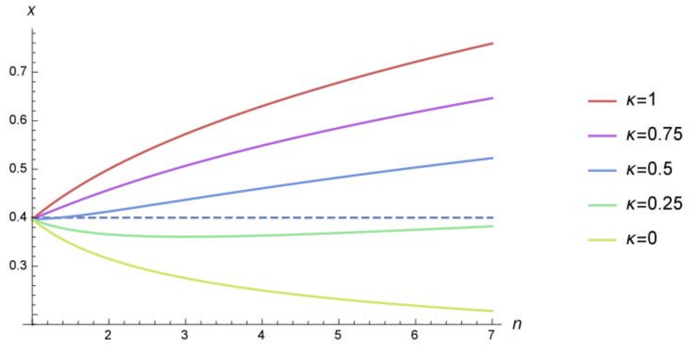

Figure 1: The unique Nash equilibrium contribution in the public-goods game for different degrees of morality

These predictions may potentially help explain observations made in laboratory experiments, in which group size sometimes has a positive effect and sometimes a negative effect on individual contributions (see Nosenzo, Quercia, and Sefton, 2015, for a review).

For a sufficiently high degree of morality, group size has a positive impact on individual contributions. One may thus get the impression that the free-rider effect is diminishing with group size. However, this is not always true. In the parametric specification used in Figure 1 ($B(X) \equiv \sqrt{X}$ and $C(x) \equiv x^2$), the individual contribution relative to the first-best contribution is

$$\frac{\hat{x}_\kappa}{x^*} = \left(\kappa + \frac{1-\kappa}{n}\right)^{2/3},$$

a ratio that decreases as group size n increases (for any given degree of morality $\kappa < 1$).[14] A smaller ratio indicates more free riding, so this equation shows that as morality (κ) increases, the effect of group size (n) on the extent of free riding declines. Moreover, the extent of free riding is bounded; the ratio $\hat{x}_\kappa / x^*$ exceeds $\kappa^{2/3}$ for all n.[15] Hence, compared to the outcome under *Homo oeconomicus*, an important policy implication is that, when κ is high, group size does not matter so much when it comes to free riding.

3.4. Environmental economics. According to World Bank president Kim Jim, “If we don’t confront climate change, we won’t end poverty”.[16] A number of instruments have been proposed to help mitigate climate change, such as a carbon tax, regulation of production technologies, subsidies to public transportation, and support to R&D concerning environmentally friendly technologies for different forms of green energy, etc. Determining the “right” carbon tax requires knowing how it will affect behavior and welfare. Here we briefly analyze the behavior of *Homo oeconomicus* and, more generally, *Homo moralis*, in a standard model of consumption that has an external effect on the environment (Musgrave, 1959, Arrow, 1970). In this model, the group is taken to be so large that each individual’s impact on the group’s environment is negligible.

More specifically, there is a continuum of consumers, indexed $i \in I = [0,1]$, and there are two consumption goods, goods 1 and 2, where good 1 is environmentally neutral (that is, its consumption has no effect on the environment) and good 2 is environmentally harmful. Aggregate consumption of these goods are

$$X_1 = \int_I x_1(i)\, d\mu \quad \text{and} \quad X_2 = \int_I x_2(i)\, d\mu,$$

where $x(i) = (x_1(i), x_2(i))$ is the consumption bundle of individual i, and μ is a density on I. Since all consumers are infinitesimally small, aggregate demand is not affected by any individual’s personal consumption.

We take the material payoff to each individual i to be that individual’s hedonic utility from own consumption, $x(i)$, and from the quality of the environment, which in turn depends on ag-

[14] Formally, $d(\hat{x}_\kappa / x^*)/dn < 0$ when $0 < \kappa \leq 1$.

[15] Formally, $\frac{d^2(\hat{x}_\kappa / x^*)}{dn d\kappa} > 0$, and $\lim_{n\to\infty}(\hat{x}_\kappa / x^*) = \kappa^{2/3}$.

[16] See http://www.worldbank.org/en/news/feature/2014/03/03/climate-change-affects-poorest-developing-countries

gregate consumption, X_2, of the environmentally harmful good. We write $v\left(x_1\left(i\right), x_2\left(i\right), X_2\right)$ for this hedonic utility and assume that it is increasing in consumption of each good and decreasing in aggregate consumption of the environmentally harmful good. Using good 1 as numeraire, writing p for the price of good 2, and assuming that all individuals have the same income, a socially efficient consumption bundle, x^*, the same for all individuals i, satisfies

$$\frac{v_2\left(x_1^*, x_2^*, X_2^*\right)}{v_1\left(x_1^*, x_2^*, X_2^*\right)} = p - \frac{v_3\left(x_1^*, x_2^*, X_2^*\right)}{v_1\left(x_1^*, x_2^*, X_2^*\right)}, \tag{3}$$

where subscripts on the personal utility function denote partial derivatives. If good 2 would have no effect on the environment ($v_3 = 0$), this equation would simply say that the marginal rate of substitution between goods 2 and 1 must equal the relative price of good 2. However, when good 2 has a negative effect on the quality of the environment ($v_3 < 0$), the marginal rate of substitution must deviate from the relative price. More precisely, the marginal rate of substitution between the environmentally harmful and environmentally neutral goods should equal the relative price of the harmful good net of the marginal rate of substitution between the utility from the quality of the environment and the neutral good. In other words, social efficiency requires that, at given prices, consumers consume less of a good the more harmful it is to the environment.

By contrast, in a population consisting entirely of *Homo oeconomicus*, an (interior) equilibrium allocation in which everybody consumes the same bundle x^0 necessarily satisfies the first-order condition

$$\frac{v_2\left(x_1^0, x_2^0, X_2^0\right)}{v_1\left(x_1^0, x_2^0, X_2^0\right)} = p. \tag{4}$$

Under decreasing marginal utility of consumption, this means that *Homo oeconomicus*, not surprisingly, consumes more of the environmentally harmful good than required by socially efficiency.

As observed above, for interactions in infinitely large groups the utility function of *Homo moralis* with degree of morality $\kappa \in [0,1]$ is the material payoff that would obtain if a share κ of the group would behave in the same way as the individual herself or himself. In the present context, if an individual consumes the bundle $x = (x_1, x_2)$ and all the others consume some bundle $y = (y_1, y_2)$, then the utility-function value, or utility, to a *Homo moralis* with degree of morality κ would be

$$U_\kappa\left(x, y\right) = v\left(x_1, x_2, \left(1-\kappa\right) y_2 + \kappa x_2\right),$$

where, in this expression, we have normalized the total mass of individuals in the group (which could be a village, region, country, continent, or the whole world) to unity. In a group consisting entirely of *Homo moralis* with the same degree of morality κ, an (interior) equilibrium allocation, in which everybody consumes the same bundle x^κ satisfies the first-order condition

$$\frac{v_2\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)}{v_1\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)} \quad = \quad p - \kappa \cdot \frac{v_3\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)}{v_1\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)}. \tag{5}$$

We first note that if good 2 would not cause any externality ($v_3 = 0$), then *Homo moralis* would "automatically" behave precisely as the classical *Homo oeconomicus* (equation (5) would boil down to equation (4)): for such goods there is no "right thing to do," and hence, morality has no bite. By contrast, if the production of good 2 is harmful for the environment ($v_3 < 0$), then for any positive degree of morality κ each individual refrains somewhat from consuming this good, compared to *Homo oeconomicus*, although each individual—knowing that she is negligible—is fully aware that her own consumption has *no* effect on the overall quality of the environment! Hence, if people are in fact somewhat moral, then policy advice based on models inhabited by *Homo oeconomicus* may exaggerate the need for pecuniary incentives such as carbon taxes. If people are more like *Homo moralis* with some positive degree of morality, then, in addition to some carbon taxes it may be effective to provide individuals with information about how carbon dioxide affects the climate.[17]

3.5. Voting. Another class of situations in which *Homo moralis* may make a difference is collective decision-making by voting. By and large, countries with more developed economies tend to have more democratic political systems. Recent research suggests that this may not merely be a correlation, but that democracy in fact may have a positive impact on growth (see, e.g., Persson and Tabellini, 2006, and Acemoglu et al., 2014). In order for democracy to work, it is important that citizens participate in elections, committee work etc., and it is still much debated in economics and political science why and how people vote. As has been pointed out by economists, high participation rates in large elections appear incompatible with rational *Homo oeconomicus* behavior. The reason being that the act of voting usually has some personal cost, say lost income or leisure, and this cost easily outweighs the expected benefit to the individual of participating in the election, since the probability of being pivotal is virtually nil. This is the well-known *voters' paradox.* Despite this, the turn-out in general and local elections in many countries is many times impressive. So what then motivates people to participate in elections? Can *Homo moralis* provide an explanation?

A closely related, and arguably equally important issue is participation and voting in committees, such as parliamentary bodies, company boards, court juries, central bank boards etc. As shown by Austen-Smith and Banks (1996), when committee members have private information and are *Homo oeconomicus,* then voting may fail to aggregate information effi-

[17] We note that equations (3) and (4) are the special cases of (5) when $\kappa = 0$ (Homo oeconomicus) and $\kappa = 1$ (Homo kantiensis). Laffont (1975) considers these two extreme cases.of self-interested individuals (our *Homo oeconomicus*) and "Kantian individuals" (our *Homo kantiensis*).

ciently even when they have the same preferences. This observation challenges the so-called Condorcet Jury Theorem (Condorcet, 1785), which states that democracy in the form of majority rule in such situations is a great institution since it implies that the right decision is almost always taken if the electorate is large enough. How would *Homo moralis* vote in such committees?

3.6. Morality, altruism and social norms. Theoretical work on the evolutionary foundations of human motivation provides insights about potential *ultimate* causes of human behavior. This line of research is complementary to behavioral economics, the branch of economics that investigates the explanatory power of richer motivations than mere self-interest. In the language of evolutionary biology, the focus in behavioral economics is on the *proximate* causes of observed human behaviors. Here we briefly discuss how *Homo moralis* preferences compare with those considered in this literature.

In the 1970's and 80's, altruistic preferences were proposed to explain intra-family transfers, transfers to the poor, and contributions to public goods (Becker 1974, 1976, Lindbeck and Weibull, 1988, Andreoni, 1988). However, altruism turned out to be insufficient to explain the data, and "warm glow" was then proposed to enhance the understanding of voluntary contributions to public goods (Andreoni, 1990). In the 1990's, inequity aversion, or a preference for fairness, was introduced by Fehr and Schmidt (1999) as an explanation for why people have a tendency to turn down low offers in the ultimatum bargaining game (Güth, Schmittberger and Schwarze, 1982). Still other forms of human motivation that have been proposed, and sometimes tested, include conformity (Bernheim, 1994), conditional altruism (Levine, 1998), identity (Akerlof and Kranton, 2000), and honesty and truth-telling (Alger and Ma, 2003, Alger and Renault, 2007, Demichelis and Weibull, 2008).

Although conceptually very different from *Homo moralis*, these preferences would be compatible with evolutionary stability if they gave rise to the same equilibrium behaviors as those of *Homo moralis*.[18] For what class of material payoff functions such behavioral equivalence obtains remains to be analyzed. Here we will limit ourselves to pointing out that *Homo moralis* preferences sometimes give rise to radically different behaviors compared to preferences that may appear to be similar. For this purpose consider altruistic preferences. An altruistic individual's preferences are usually represented as a utility function that attaches unit weight to the individual's own material payoff and a positive weight, less than one, to other individuals' material payoffs. An altruist hence internalizes some of the external effects of his or her behavior on others. Let the latter weight be denoted α, the individual's *degree of*

[18] However, *Homo moralis* are the only preferences that are evolutionarily stable in the whole class of interactions analyzed in Alger and Weibull (2013, 2016).

altruism towards the other party.[19] For some material payoff functions an altruist with degree of altruism α behaves exactly like a *Homo moralis* with degree of morality $\kappa = \alpha$. Hence, in some interactions one cannot discriminate between moralism and altruism as explanations for observed behavior. However, the two classes of preferences are conceptually quite distinct, and induce radically different behaviors in some interactions. This is particularly striking in two situations: first, in small groups facing coordination problems; second, in interactions with arbitrary many participants.

Let us first briefly consider an example from Alger and Weibull (2013), a simple 2×2-coordination game in terms of material payoffs.

	A	B
A	$2,2$	$0,0$
B	$0,0$	$1,1$

There are two alternative potential societal "conventions" when individuals pair up to play this game, namely, that either both parties take action A or both parties take action B. Clearly the first convention is Pareto superior to the second. However, under each convention, *Homo oeconomicus* individuals have no incentive to unilaterally deviate. Granted a sufficiently large population share act according to the going convention, an individual deviator would looses material payoff, and, in addition, incur a payoff loss on the unfortunate opponent.[20] Therefore also an altruist would stick to the going convention, even if this happened to be the socially inferior convention to always take action B. But not so a *Homo moralis* of high enough degree of morality. For suppose a *Homo kantiensis* were to visit an country where every citizen takes action B in every encounter, and suppose that the visitor is indistinguishable from a citizen. Then *Homo kantiensis* would take action A in each encounter, since this would be "the right thing to do" if upheld as a universal law" of conduct.[21] This moralistic visitor will earn material payoff zero in each encounter and so will the unfortunate citizens who meet him. The citizens would very much wish that the visitor instead had been a *Homo oeconomicus* or an altruist.

[19] For $n = 2$, an altruist's utility writes $u_\alpha(x,y) = \pi(x,y) + \alpha\pi(y,x)$. We note that this function may also be interpreted as the individual having a concern for efficiency, since it is a monotone transformation of $v_\alpha(x,y) = \pi(x,y) + \frac{\alpha}{1-\alpha}[\pi(x,y) + \pi(y,x)]$.

[20] These are strict Nash equilibria in terms of material payoffs. The game also has a mixed equilibrium, in which each individual plays A with probability 1/3. However, this equilibrium is unstable in all plausible population dynamics. See Young (1993) and Myerson and Weibull (2015) for formal models of stable conventions in large populations.

[21] Indeed, to take action A is optimal for all *Homo moralis* with dgree of morality $\kappa \geq 1/3$.

Turning now to the second situation in which *Homo moralis* preferences give rise to radically different behaviors compared to altruism, consider again the environmental-economics and the public goods examples. In the former example morality led consumers to reduce their consumption even though each individual's consumption was negligible. In the public goods example, as n tends to infinity the donation to the public good tends to a positive amount for any positive degree of morality. By contrast, Andreoni (1988) has shown that in a population of altruists the proportion of individuals who make positive donations shrinks to zero as the number of individuals grows infinitely large; for each individual donation then has a negligible effect on the total value of the public good. There is thus a sharp distinction between morality and altruism when group size is very large. Even if an individual is highly altruistic and cares a lot about the consequences of her behavior for others, she will behave very much like *Homo oeconomicus* if her environmental impact is marginal. By pondering what would happen if others were to behave as she does, a *Homo moralis* comes to care directly about her behavior, beyond the effects that this behavior has on her own material payoff.

This observation may have important implications for other policy issues, such as tax evasion. It has been noted by some economists (see Sandmo, 2005), that there appears to be less tax evasion in certain countries than would be compatible with *Homo oeconomicus*'s behavior. The risk of being caught is often small and the penalties mild, so maximization of expected personal utility would suggest much tax evasion. So why do people, in those countries, and perhaps many in other countries, not evade taxes more? Since the marginal effect of any change in an individual's tax payment is, with few exceptions, negligible, pro-social preferences such as altruism or inequity aversion may fail to explain why individuals evade taxes less than *Homo oeconomicus* would. However, as suggested by the analysis above, *Homo moralis* may supply an explanation, since a *Homo moralis* may, to a certain extent, prefer to pay their taxes, since she cares about the moral quality of her actions.

A final point before concluding. Some researchers have developed models in which individuals care about norms, and/or have a concern for their image (in the eyes of others and perhaps also in their own eyes) or a desire to avoid social stigma (Lindbeck, Nyberg, and Weibull, 1999, Brekke, Kverndokk and Nyborg, 2003, Bénabou and Tirole, 2006, Ellingsen and Johannesson, 2008, Huck, Kübler, and Weibull, 2012). In some of these models, individuals have a baseline intrinsic wish to behave well, and their image concern may strengthen this wish (Falk and Tirole, 2016). Our theory did not include image concerns in the set of possible preferences. Nonetheless, it provides an explanation of the ultimate reasons for why individuals would have a baseline intrinsic wish to behave well, even absent such image concerns.

4. Conclusion

In this essay, we have discussed (a) evolutionary foundations for human motivation, (b) how evolution favors the class of *Homo moralis* preferences, and (c) implications for economics and policy of such preferences. We have sought to convey the following main points:

1. Economics possesses powerful analytical tools that enable positive and normative analyses of a very wide range of social and economic phenomena. These tools should not be abandoned but brought to more general use.
2. The conventional assumption among economists, since the days of Adam Smith's (1776) *Wealth of Nations*, is that economic agents are purely self-interested and focused on their own consumption. Yet other social and behavioral sciences, experimental work, everyday observation, and introspection suggest that human motivation is much more complex, sometimes systematically deviating from narrow self-interest.
3. First principles in evolutionary biology, formalized in terms of evolutionary stability along the lines of Maynard Smith and Price (1973), suggest that natural selection favors human motivation in the form of *Homo moralis*, a generalization of *Homo oeconomicus* that allows for varying degrees of morality alongside self-interest.
4. By applying the powerful analytical tools of economics to the more general *Homo moralis*, rather than only to the special case of *Homo oeconomicus*, new predictions and policy recommendations follow. In particular, since *Homo moralis* is not only motivated by her material gains and losses, policy based on *Homo oeconomicus* may lead to exaggerated use of pecuniary incentives, such as distortionary taxes. If people do have a natural inclination for moral concerns, it may be more effective to provide them with information about the consequences of their actions, for themselves and others.

Our results being purely theoretical, empirical and experimental work will be necessary to determine the empirical validity of *Homo moralis.* To this end, also further theoretical analysis is needed, for although we have here examined the behavior of *Homo moralis* in some common situations, we have but scratched the surface, and, moreover, many fundamental questions have not been addressed at all. In particular, one fundamental issue that we have not (yet) addressed is welfare. For economic and social policy, this is a most important, and yet philosophically non-trivial issue, especially when individuals have "social" preferences. More specifically, if individuals have *Homo moralis* preferences, perhaps idiosyncratic degrees of morality, should then welfare be defined in terms of the material payoffs or in terms of individuals' utility functions? More generally, if individuals have pro-social or other-regarding preferences, perhaps involving a concern for social norms and/or their image in the eyes of others (and perhaps also themselves), perhaps even involving some spite etc., how then define welfare?

This philosophically and methodologically difficult issue was addressed by John Harsanyi in two wonderful essays that deal with game theory, utilitarianism and ethics, see Harsanyi (1979, 1992). In these essays he advocates what he calls "rule utilitarianism", an approach we find appealing also for *Homo moralis*. Harsanyi distinguishes between an individual's "personal preferences" and his or her "moral preferences", and advocates that, when defining welfare in a society, one should only consider the personal preferences.[22] In cases when individuals' preferences can be represented by an additive utility function, where one term can be taken to represent "personal utility", Harsanyi argues that welfare should be defined as the sum of all individuals' expected personal utilities, behind the veil of ignorance as to what societal position each individual will end up in. This is in line with *Homo moralis*. If we take the material payoff function π used above to represent personal utility, then welfare in a society consisting of *Homo moralis* individuals (each with his or her degree of morality) should be defined simply as the sum of their expected material payoffs, just as in ordinary utilitarian welfare theory.

A final point we would like to make concerns the status of economics as a discipline, in the general public and among the other behavioral and social sciences. Conventional economics textbooks may give the false impression that economic rationality incorporates selfishness (see discussion in Rubinstein, 2006, and the references therein). This misreading of conventional economics probably hurts the reputation of economists. If economists would use the more general *Homo moralis* instead of the special case of *Homo oeconomicus*, then such misunderstandings and critique would fall flat to the ground. Then the economist's analysis would not be prejudiced in favor of neither selfishness nor morality, but would allow for the whole spectrum of intermediate degrees of morality, spanning from pure self-interest to pure Kantian morality.

5. Appendix

In order to give an exact definition of *Homo moralis* some notation and technicality will be needed, here kept to a minimum (readers interested in more detail are suggested to consult Alger and Weibull, 2016). First, let $\pi(x, \boldsymbol{y})$ denote the material payoff to an individual who

[22] The following example is due to Peter Diamond (in conversation with one of the authors). Suppose a parent has one selfish and one altruistic child, and has a cake to divide between them. Should the parent give a bigger slice to the selfish child, thus maximizing the sum of their altruistic and selfish utilities, or should the parent given them equally large slices, thus maximizing the sum of only their hedonic utilitites?

takes course of action x, or, to use game-theoretic jargon, uses strategy x, in a situation when the other $n-1$ group members use strategies $\boldsymbol{y} = (y_1, y_2, ..., y_{n-1})$.[23] Our assumption that the interaction is aggregative can now be expressed precisely as follows: the material payoff $\pi(x, \boldsymbol{y})$ is invariant under permutation of the components of the strategy profile $\boldsymbol{y}$ of the other group members.

So much about the interaction that takes place within each group. We next need to briefly consider the (exogenous) random matching process whereby groups are formed. Consider any "population state" in which only two types of individual are present, those with some utility function U, a type we take to be more frequent, and those with another utility function V, a type we take to be less frequent. Let the population share of the latter type be denoted $\varepsilon > 0$. We call individuals of the first type *incumbents* or *residents* and individuals of the second type *mutants*. In any group that is about to interact, the number of mutants is a random variable the probability distribution of which depends on the matching process (which we here take to be exogenous). For any given *mutant* group member, let $q_m(\varepsilon)$ be the probability that the number of *other* mutants in his or her group is m (for $m = 0, 1, ..., n-1$) and write $\boldsymbol{q}(\varepsilon) = (q_0(\varepsilon), ..., q_{n-1}(\varepsilon))$ for the so defined probability distribution. Let $\boldsymbol{q}^*$ be its limit as $\varepsilon \to 0$.

For example, under uniform random matching (what biologists refer to as a well-mixed population), there is almost surely no other mutant in a mutant's group, in the limit as the mutant type becomes vanishingly rare, so then $\boldsymbol{q}^* = (1, 0, 0, ..., 0)$. By contrast, if groups are formed exclusively among siblings, who each inherited their type from one of their parents (with equal probability for both parents), the number of other mutants in a mutant's group, is binomially distributed, with probability parameter $p = 1/2$.[24]

We are now in a position to define *Homo moralis.*

Definition 1. *An individual is a* ***Homo moralis*** *if his or her utility function U satisfies $U(x, \boldsymbol{y}) \equiv \mathbb{E}[\pi(x, \tilde{\boldsymbol{y}})]$ where $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_{n-1})$ is a random strategy profile for the other group members, with each component $\tilde{y}_i$ being either y_i or x, and where the probability distribution for $\tilde{\boldsymbol{y}}$ is such that each component of $\boldsymbol{y}$ is equally likely to be replaced by x.*

Members of this new *Homo moralis* "species" may differ in their probability distributions for the random vector $\tilde{\boldsymbol{y}}$. For any given *Homo moralis*, let μ_m denote the probability that exactly m of the $n-1$ components of $\boldsymbol{y}$ are replaced by x (by definition with equal probability

[23] All participants have access to the same set of strategies.

[24] In the limit as mutants become vanishingly rare, a given mutant almost surely has exactly one mutant parent (the probability of no mutant parent is approximately ε and the probability of two mutant parents is approximately ε^2). Hence, the probability that any given other sibling is also a mutant is approximately $1/2$.

for each subset of size m) while the remaining components of $\boldsymbol{y}$ keep their original values. We will call the so defined probability vector $\boldsymbol{\mu}$ the *morality profile* of that member of *Homo moralis*.[25] Clearly, *Homo oeconomicus* is a special member of *Homo moralis*, namely, the member with morality profile $\boldsymbol{\mu} = (1, 0, ..., 0)$. Then $\Pr[\tilde{\boldsymbol{y}} = \boldsymbol{y}] = 1$ and so its utility is its own material payoff, $U_E(x, \boldsymbol{y}) \equiv \pi(x, \boldsymbol{y})$. At the opposite extreme of the spectrum of *Homo moralis* we find what we call *Homo kantiensis*, those members of the *Homo moralis* family that have the opposite morality profile $\boldsymbol{\mu} = (0, ..., 0, 1)$. Then $\Pr[\tilde{\boldsymbol{y}} = (x, x, ..., x)] = 1$ and thus their utility function is $U_K(x, \boldsymbol{y}) \equiv \pi(x, (x, x, ..., x))$. Individuals of this "pure Kantian" variety of *Homo moralis* always choose a strategy x that, if hypothetically adopted by everyone else in the group would maximize each group member's material payoff.

The behavior of all other varieties of *Homo moralis* (that is, with arbitrary morality profile $\boldsymbol{\mu}$) lies between these two extremes; they attach some weight to the consequences for their own material payoff and some weight to what would be "the right thing to do" if their own behavior became a probabilistically followed "universal law". This is most easily seen in the case of pairwise interactions. For $n = 2$, the identity $U(x, \boldsymbol{y}) \equiv \mathbb{E}[\pi(x, \tilde{\boldsymbol{y}})]$ boils down to $U(x, y) \equiv \mu_0 \cdot \pi(x, y) + \mu_1 \cdot \pi(x, x)$. This utility function is a convex combination of pure selfishness (U_E) and pure Kantian morality (U_K), with weight μ_0 attached to the first goal and the complementary probability weight $\mu_1 = 1 - \mu_0$ to the second.[26]

For interactions between more than two parties, the utility function of *Homo moralis* is mathematically fairly involved. However, this is not always the case. In particular, suppose that, for any given mutant the types of any two *other* group members are statistically independent.[27] Then the evolutionarily stable variety of *Homo moralis*, that is, the variety with morality profile equal to the assortativity profile of the matching process, is binomial:

$$\mu_m = q_m^* = \binom{n-1}{m} \sigma^m (1 - \sigma)^{n-m-1}$$

(for any $n > 1$ and $m = 0, 1, .., n - 1$), where $0 \le \sigma \le 1$ is the probability that a randomly drawn other group member in a mutant's group is also a mutant.

The utility of a member of this "subspecies" of *Homo moralis* is to maximize his or her expected material payoff if, hypothetically, other members of his or her group would randomly and statistically independently switch to use her strategy with probability σ. This *Homo*

[25] We note that all μ_m lie between zero and one and that they sum to one.

[26] In Alger and Weibull (2013) we focus exclusively on the case of pairwise interactions and call μ_1 the *degree of morality.*

[27] This restriction on the nature of the matching process is vacuous in the case of pairwise matching and is always met for siblings (and other relatives in the same generation) under haploid inheritance and sexual reproduction.

moralis "subspecies" is thus one-dimensional—parametrized by a single number σ in the unit interval—and spans from pure selfishness (*Homo oeconomicus*), at $\sigma = 0$, to pure Kantian morality (*Homo kantiensis*), at $\sigma = 1$. Moreover, in large groups with such conditional independence, the *share* of mutants in a mutant's group is, by the deMoivre-Laplace Theorem approximately normally distributed with mean value σ and variance $\sigma(1-\sigma)/(n-1)$. Hence, in large groups the share of mutants is almost deterministic and equal to σ. A *Homo moralis* in such large groups acts as if she hypothetically assumed that her behavior were to become, if not a "universal law" ($\sigma = 1$), a "random law" applying to a randomly sampled share of size σ out of her group's other members.

6. References

Acemoglu, D., S. Naidu, P. Restrepo, and J.A. Robinson (2014): "Democracy Does Cause Growth," NBER Working Paper 20004.

Akerlof, G. and R. Kranton (2000): "Economics and Identity," *Quarterly Journal of Economics,* 115, 715-753.

Alexander, R. D. (1987): *The Biology of Moral Systems.* New York: Aldine De Gruyter.

Algan, Y., and P. Cahuc (2010): "Inherited Trust and Growth," *American Economic Review*, 100, 2060-2092.

Alger, I., and A. Ma (2003): "Moral Hazard, Insurance, and Some Collusion," *Journal of Economic Behavior and Organization*, 50, 225-247.

Alger, I. and R. Renault (2007): "Screening Ethics when Honest Agents Care about Fairness," *International Economic Review*, 47, 59-85.

Alger, I., and J. Weibull (2013): "Homo Moralis – Preference Evolution under Incomplete Information and Assortativity," *Econometrica*, 81, 2269-2302.

Alger, I., and J. Weibull (2016): "Evolution and Kantian Morality," forthcoming, *Games and Economic Behavior.*

Andreoni, J. (1988): "Privately Provided Public Goods in a Large Economy: The Limites of Altruism," *Journal of Public Economics*, 35, 57-73.

Andreoni, J. (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 100, 464-477.

Arrow, K. (1970): "Political and Economic Evaluation of Social Effects and Externalities," in J. Margolis (ed.), *The Analysis of Public Output.* New York: Columbia University Press.

Arrow, K. (1973): "Social Responsibility and Economic Efficiency," *Public Policy*, 21, 303-317.

Austen-Smith, D., and J.S. Banks (1996): "Information Aggregation, Rationality, and the Condorcet Jury Theorem," *American Political Science Review*, 90, 34–45.

Banerjee, A., and J. Weibull (1995): "Evolutionary Selection and Rational Behavior," in Alan Kirman and Mark Salmon (eds.), *Learning and Rationality in Economics*, Oxford: Basil Blackwell.

Becker, G. (1974): "A Theory of Social Interaction", *Journal of Political Economy*, 82, 1063-1093.

Becker, G. (1976): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology", *Journal of Economic Literature*, 14, 817-826.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652-1678.

Berg, J., J. Dickhaut, and K. McCabe (1995): "Trust, Reciprocity, and Social History," *Games and Economic Behavior,* 10, 122-142.

Bergstrom, T. (1995): "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review,* 85, 58-81.

Bergstrom, T. (2003): "The Algebra of Assortative Encounters and the Evolution of Cooperation," *International Game Theory Review,* 5, 211-228.

Bergstrom, T. (2009): "Ethics, Evolution, and Games among Neighbors," Working Paper UCSB.

Bergstrom, T. (2012): "Models of Assortative Matching," Working Paper UCSB.

Bernheim, B.D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102:841–877.

Brekke, K.A., S. Kverndokk, and K. Nyborg (2003): "An Economic Model of Moral Motivation," *Journal of Public Economics*, 87, 1967–1983.

Cesarini, D., C. Dawes, J. Fowler, M. Johannesson, P. Lichtenstein, and B. Wallace (2008): "Heritability of Cooperative Behavior in the Trust Game," *Proceedings of the National Academy of Sciences*, 105, 3721-3726.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74, 1579–1601.

Condorcet (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris: L'Imprimerie Royale.

Currarini, S., M.O. Jackson, and P. Pin (2009): "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica,* 77, 1003–1045.

Currarini, S., M.O. Jackson, and P. Pin (2010): "Identifying the Roles of Race-Based Choice and Chance in High School Friendship Network Formation," *Proceedings of the National Academy of Sciences,* 107, 4857–4861.

Darwin, C. (1859): *The Origin of Species, by Means of Natural Selection.* London: John Murray.

Dawkins, R. (1976): *The Selfish Gene.* Oxford: Oxford University Press.

Dekel, E., J.C. Ely, and O. Yilankaya (2007): "Evolution of Preferences," *Review of Economic Studies*, 74, 685-704.

Demichelis, S., and J. Weibull (2008): "Language, Meaning, and Games: A Model of Communication, Coordination, and Equilibrium", *American Economic Review*, 98, 1292-1311.

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

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98, 990-1008.

Falk, A., and J. Tirole (2016): "Narratives, Imperatives and Moral Reasoning", mimeo Toulouse School of Economics.

Feddersen, T., and A. Sandroni (2006): "A Theory of Participation in Elections," *American Economic Review* 96, 1271-1282.

Fehr, E., and K. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817-868.

Frank, R.H. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, 77, 593-604.

Friedman, M. (1953): *Essays in Positive Economics.* Chicago: University of Chicago Press.

Ghosh, P., and D. Ray (2016): "Information and Enforcement in Informal Credit Markets," *Economica*, 83, 59-90.

Greene, J., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D. Cohen (2001): "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science*, 293, 2105-2108.

Güth, W., R. Schmittberger and B. Schwarze (1982), "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior & Organizatio*n, 3, 367-388.

Güth, W., and M. Yaari (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.

Haidt, J. (2003): "The Moral Emotions," in R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (eds.), *Handbook of Affective Sciences.* Oxford: Oxford University Press.

Hamilton, W.D. (1964): "The Genetical Evolution of Social Behaviour," *Journal of Theoretical Biology*, 7, 1-52.

Harsanyi, J..C. (1953): "Cardinal Utility in Welfare Economics and in the Theory of Risk Taking," *Journal of Political Economy* 61, 434-435.

Harsanyi, J.C. (1979): "Rule Utilitarianism, Rights, Obligations and the Theory of Rational Behavior," *Journal of Philosophy*, 76, 218-236.

Harsanyi, J.C. (1992): "Game and Decision Theoretic Models in Economics," in R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory.* Amsterdam: North-Holland.

Heifetz, A., C. Shannon, and Y. Spiegel (2007): "What to Maximize if You Must," *Journal of Economic Theory*, 133, 31-57.

Huck, S., D. Kübler, and J.W. Weibull (2012): "Social Norms and Economic Incentives in Firms," *Journal of Economic Behavior & Organization*, 83, 173-185.

Kant, I. (1785): *Grundlegung zur Metaphysik der Sitten.* [In English: *Groundwork of the Metaphysics of Morals.* 1964. New York: Harper Torch books.]

Kendall, J., N. Mylenko, and A. Ponce (2010): "Measuring Financial Access around the World." World Bank Policy Research Working Paper 5253.

Laffont, J.-J. (1975): "Macroeconomic Constraints, Economic Efficiency and Ethics: an Introduction to Kantian Economics," *Economica*, 42, 430-437.

Levine, D. (1998): "Modelling Altruism and Spite in Experiments," *Review of Economic Dynamics*, 1, 593-622.

Lindbeck, A., S. Nyberg, and J. Weibull (1999): "Social Norms and Economic Incentives in the Welfare State," *Quarterly Journal of Economics*, 114, 1-33.

Lindbeck, A., and J. Weibull (1988): "Altruism and Time Consistency - the Economics of Fait Accompli," *Journal of Political Economy*, 96, 1165-1182.

Maynard Smith, J., and G.R. Price (1973): "The Logic of Animal Conflict," *Nature,* 246, 15-18.

McPherson, M., L. Smith-Lovin, and J.M. Cook (2001): "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415-444.

Musgrave, R.A. (1959): *The Theory of Public Finance; A Study in Public Economy.* New

York: McGraw Hill.

Myerson, R. and J. Weibull (2015): "Tenable Strategy Blocks and Settled Equilibria", *Econometrica*, 83, 943-976.

Nichols, S. (2004): *Sentimental Rules.* Oxford: Oxford University Press.

Nosenzo, D., S. Quercia, and M. Sefton (2015): "Cooperation in Small Groups: The Effect of Group Size," *Experimental Economics*, 18, 4-14.

Ok, E.A., and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory,* 97, 231-254.

Persson, T. and G. Tabellini (2006): "Democracy and Development: The Devil in the Details," *American Economic Review*, *Papers and Proceedings*, 96, 319-324.

Rawls, J. (1957): "Justice as Fairness," *Journal of Philosophy*, 54, 653-662.

Rousset, F. (2004): *Genetic Structure and Selection in Subdivided Populations.* Princeton: Princeton University Press.

Rubinstein, A. (2006): "A Sceptic's Comment on the Study of Economics," *Economic Journal*, 116, C1-9.

Ruef, M., H.E. Aldrich, and N.M. Carter (2003): "The Structure of Founding Teams: Homophily, Strong Ties, and Isolation among U.S. Entrepreneurs," *American Sociological Review*, 68, 195-222.

Sandmo, A. (2005): "The Theory of Tax Evasion: A Retrospective View," *National Tax Journal*, 58, 643-663.

Sen, A.K. (1977): "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy & Public Affairs*, 6, 317-344.

Smith, A. (1759): *The Theory of Moral Sentiments.* Reedited (1976), Oxford: Oxford University Press.

Smith, A. (1776): *An Inquiry into the Nature and Causes of the Wealth of Nations.* Reedited (1976), Oxford: Oxford University Press.

Vickrey, W. (1945): "Measuring Marginal Utility by Reactions to Risk," *Econometrica*, 4, 319-333.

Wright, S. (1931): "Evolution in Mendelian Populations," *Genetics*, 16, 97–159.

Young, P. (1993): "Conventions," *Econometrica*, 61, 57-84.