Journal of Business Economics (2020) 90:401–425
https://doi.org/10.1007/s11573-019-00964-z

ORIGINAL PAPER

"There's Life in the Old Dog Yet": The *Homo economicus* model and its value for behavioral ethics

Philipp Schreck[1] · **Dominik van Aaken**[2] · **Karl Homann**[3]

Published online: 19 November 2019

Abstract
The *Homo economicus* model (HEM) is widely used in the social sciences in general and in business ethics in particular. Despite its success, the model is frequently criticized for being empirically flawed and normatively dangerous, and its critics argue that it should be abandoned and replaced by more realistic models of human behavior. In response to the HEM's critics, this paper develops a precise methodological approach that makes it possible to integrate within the HEM seemingly contradictory empirical evidence. Using the methodology we develop, we will integrate recent findings in behavioral economics and show how a rational-choice approach to behavioral ethics can illuminate the emergence, salience and persistence of morality.

Keywords Behavioral economics · Behavioral ethics · Business ethics · Dilemma structure · Cooperation · *Homo economicus* · Methodology · Rational-choice

JEL Classifications A12 · B41 · D01 · D9

1 Introduction

The *Homo economicus* model is widely used in the social sciences in general and in the field of business ethics in particular. Its main underlying assumption is that decision-makers act as if they were purely rational and cared solely about their self-interest (Kirchgässner 2008; Vanberg 2002). In its basic form, the model has also been used to study morality, even though morality implies that an individual takes

✉ Philipp Schreck
philipp.schreck@wiwi.uni-halle.de

[1] School of Law and Economics, Martin-Luther-University Halle-Wittenberg, Große Steinstr. 73, 06108 Halle (Saale), Germany

[2] Department of Business, Economics and Social Theory, University of Salzburg, Kapitelgasse 5-7, 5020 Salzburg, Austria

[3] School of Philosophy, Philosophy of Science, and Religious Studies, Ludwig-Maximilians-University Munich, Ludwigstr. 28, 80539 Munich, Germany

into account another individual's interests (Donaldson and Dunfee 1999; Gauthier 1986; Schreck et al. 2013).

Despite the prominence of the *Homo economicus* model (HEM) in the social sciences, its validity and usefulness have been challenged on both normative and empirical grounds. The normative criticism of the model has its roots in a long tradition of Western philosophy that starts with Plato and Aristotle and extends to some of the most prominent nineteenth century philosophers, who were critical of economic activities and profit per se. Martin Luther's "On Trade and Usury" (1524) is an example of this tradition. Luther explicitly cites the need to stand against disreputable merchants who make too much profit, referring to psalm 37:16: "Better is the little that the righteous hath than the great possessions of the godless." One of the most influential approaches to ethics—the work of Immanuel Kant—also belongs to this tradition. Strictly separating prudence from morality, Kant denies the moral quality of a merely prudent act.[1]

More recent instances of normative criticism focus on how the *Homo economicus* model has been applied in the social sciences, notably in business and economics (Anderson 2000; Dierksmeier 2011; Hirschman 1985; Néron 2015; Sen 1977). Some critics have claimed that business schools are responsible for the business elite's irresponsible behavior and are guilty of "propagating ideologically inspired amoral theories" (Ghoshal 2005: 76) among students. If, so the argument goes, students are exposed extensively to the "entirely inhuman *Homo Oeconomicus*" (Hühn 2014: 537), they themselves will begin behaving as selfishly and opportunistically as the model assumes.[2]

The *Homo economicus* model has also been challenged on empirical grounds. Critics of the model have put forward the findings of experimental research in various disciplines to argue that the model is empirically flawed. The experimental evidence, notably in the domain of behavioral economics, that these critics cite suggests that human beings simply do not behave as the assumptions underlying the model seem to imply. More specifically, this evidence contradicts the assumptions of rationality, self-interested utility maximization, information processing, and consistency of choices (e.g., Gintis et al. 2005). Critics have used this evidence to argue that the assumptions of the HEM are "obviously false" (Thaler and Sunstein 2008: 9), and hence cast doubt on the usefulness of this model (e.g., Kluver et al. 2014; Levitt and List 2008; Mueller 1995).

In sum, the *Homo economicus* model has been criticized for being empirically flawed and normatively dangerous. As a consequence, some of the model's critics

[1] Kant famously illustrates his position with an example from business: if a merchant who could easily cheat charges a fair price to everybody because he is worried that, if he did not, his reputation would suffer, there is no moral worth in his honesty, as his behavior is just a matter of prudence (Kant 1785/2013: 487). Another prominent critic of economic acts is Karl Marx who denies the possibility of morality, solidarity and humanity in capitalist market systems (Marx 1959/2007). On egoist motives in Marxism, see Churchich (1994: 145–169).

[2] Some researchers describe a similar "indoctrination effect" to explain why in laboratory experiments economics students tend to behave less cooperatively than students of other social sciences (Bauman and Rose 2011; Frank et al. 1993; Ostrom 1998).

conclude that it should be abandoned and replaced by more realistic models of human behavior (e.g., Friedland and Cole 2017; Laville 2000). Against the background of this criticism, this paper will seek to evaluate the usefulness of the *Homo economicus* model for studying morality. In particular, we will reflect upon the role that the model might play in a positive, as opposed to a normative, theory of moral behavior. What we propose, in effect, is a rational-choice approach to behavioral ethics. To support our approach, we will focus on the empirical charges against the concept of *Homo economicus*. Although we do not challenge the empirical results that give rise to such criticism, we believe that they should not be interpreted as a falsification of the *Homo economicus* model. The gist of our paper is that these results should rather be taken as the starting point for exploring the structural conditions that facilitate moral behavior, and that a rational-choice approach to behavioral ethics, with the HEM at its methodological core, facilitates the rigorous analysis of these conditions.

To develop our arguments, we will, first, propose and defend a precise methodological conception of the *Homo economicus* model in the next section. In particular, we will draw on Popper's concept of the "logic of the situation" and its role in explanatory theories in the social sciences (Popper 1964/1994). These clarifications will help us specify the methodological status of the *Homo economicus* model in Sect. 3. There we will show that the main reason for using this model is not that it is empirically valid, but that it draws the researcher's attention to situational constraints as the crucial determinant of behavior. In Sect. 4 we will explain why the *Homo economicus* model is a useful heuristic for the study of morality. Our main argument will be that *Homo economicus* is an adequate analytical tool to model amoral behavior in competitive markets or, more generally, in social dilemmas. Inspired by recent work in social psychology and anthropology (e.g., Greene 2013; Haidt 2008; Tomasello 2016), in Sect. 5 we will interpret morality as a cultural achievement that allows humans to cooperate successfully. From this perspective, we will argue that the experimental findings of behavioral economics should not be interpreted as falsifications of the HEM, but as refinements of the model that specify the conditions under which certain forms of morality may emerge and persist. Finally, in Sect. 6 we will outline the normative implications of our descriptive approach for research.

2 The rationality principle in the social sciences

2.1 *Homo economicus*: assumptions and empirical evidence

The *Homo economicus* model, which is also known as the "rational-choice theory" or "the economic model of man", remains one of the methodological cornerstones of economics (Sugden 1991; Vanberg 2004). Two assumptions lie at the center of this decision-making model: rational maximization and material self-interest. In a nutshell, these assumptions imply that when decision makers have multiple options, they choose the alternative that maximizes their own material pay-off.

Rational-choice theory has been widely used in the social sciences, including the area of business studies (Combs et al. 1999; Kirchgässner 2008; Di Stefano et al.

2015). Most importantly for our purposes, the rational-choice approach has been used for the study of morality in general (Gauthier 1986; Greene 2013; Hosmer and Chen 2001), and in business ethics in particular (Donaldson and Dunfee 1999; McWilliams and Siegel 2001; Pies et al. 2009; Schreck et al. 2013).

The success of the HEM would not have been possible without the groundbreaking work of economist and Nobel Prize laureate Gary Becker. Becker (1976, 1993) proposed that economics should be understood as a general approach to analyzing human behavior. In his view, economics as a method of *analysis* is not restricted to the economy, but can be applied to a wide range of social phenomena, including marriage, reproductive behavior or drug consumption. This extension of the phenomenological domain of economics to an unlimited range of topics has been dubbed "economic imperialism" (Lazear 2000). In effect, the rational-choice approach has been used anywhere, from economics to sociology, psychology, law and ethics (Kirchgässner 2008).

The model's success is somewhat surprising, given that it is plainly unrealistic. Even without being familiar with research in the modern behavioral sciences (Kahneman and Tversky 1979; Loewenstein and Thaler 1989; Simon 1957), it is not difficult to see that the HEM offers a very poor account of what human beings *actually* are. To explain this argument, let us consider the model's core assumptions, rationality and self-interest. Human beings of flesh and blood typically do not act rationally in a strict sense: they do not know every alternative, they are not capable of comparing and evaluating the consequences of every alternative, and they often fail to identify the alternative that maximizes their profit or happiness, among other things.

Similarly, in the context of ethics, the assumption of strict self-interest contradicts daily experience. For example, many people donate money to help people they do not even know, refrain from lying as much as possible in their tax statements, and do not betray others whenever it is possible to benefit from betrayal without being punished. Hundreds of experiments in the fields of experimental economics and behavioral ethics provide scientific evidence that, by and large, many, if not most, of us *do* follow social norms to some extent, even when this behavior does not pay off in monetary terms (Cooper and Kagel 2016; Fehr and Schmidt 2003).

On these grounds it has been argued that the HEM is empirically flawed and should thus be replaced by more realistic models (Etzioni 2010; Laville 2000; Levitt and List 2007; Lindenberg 1990; Meckling 1976). And for those who believe in scientific progress through falsification, isn't it all too obvious that a lack of empirical support for a theory is a reason to reject it (Popper 1959/2005)? Interestingly, it is exactly Karl Popper who provides arguments that can be used to justify the *Homo economicus* model's validity despite its apparent empirical flaws.

2.2 Explanation and falsification in the social sciences

Popper's methodology is of particular relevance to our purposes because it clarifies how scientists may proceed in order to explain social phenomena. Two caveats are important at this point. First, the main goal of the approach we propose here is to

explain, rather than justify, moral behavior. Second, by referring to Popper's writings, we build on a particular philosophy of science. Although his critical-rationalist view plays a major role in economic methodology (Blaug 1992; Dow 1997), it is—as any philosophy of science—contingent, i.e., not without alternatives.[3]

According to Popper (1963/1985, 1964/1994), explanations in the social sciences include two classes of elements: first, the analysis of the "logic of the situation" (first mentioned in Popper 1945/2011: 308, 324) as an acting agent sees it and, second, the agent's reaction to this situation. This reaction is assumed to be rational in the sense of "adequate" or "appropriate"; this is, in essence, Popper's rationality principle. Interestingly, Popper characterized this principle as an "almost empty […] methodological postulate" (Popper 1964/1994: 169) rather than a psychological proposition and concluded that this principle is false in the sense that it is not "universally true", and thus "does not play the role of […] a testable hypothesis" (p. 169). In contrast, Popper argued, the logic of the situation can be approximated by means of "models". Situational models try to capture empirical conditions and predict, on that basis, hypothesized behavioral reactions. These models can be "empirically more or less adequate" (Popper 1964/1994: 169) and are thus testable. When such models are falsified, they should be amended and improved.

For Popper, the two main elements of an explanatory theory in the social sciences are the non-testable principle of rationality and empirical models of a given situation. Taken together, these two classes of elements form a theory that allows for explaining and predicting observed behavior. When a theory is tested and empirical evidence contradicts the theory's predictions, the researcher has to decide which part of the theory is responsible for its predictive failure. In Popper's view it is not the principle of rationality that should be questioned when empirical evidence proves the theory's predictions wrong, but the model (Popper 1964/1994).

This is not because the rationality principle was true in the sense that people always behave according to this principle. Because the principle of rationality has "little or nothing to do with the empirical or psychological assertion that man always, or in the main, or in most cases, acts rationally" (Popper 1964/1994: 169). Popper's reason for supporting this methodological strategy is that it gives "rise to better testable explanatory hypotheses—that is, conjectural situational models—than other methods" (Popper 1964/1994: 171). This is because situational models are more informative with respect to the social world than the rationality principle is. In Popper's view, there is not much to learn when we learn that the rationality assumption is "wrong"—we knew this before (Popper 1964/1994: 178).

When observed behavior is interpreted as a rational response to a specific "situation," the researcher's attention is drawn to the situation as the explanatory element for that behavior. Holding the first element of a theory, the rationality principle, constant and adjusting situational models to empirical observation is Popper's proposed

[3] Popper himself did not think of his methodology as 'contingent,' but understood it as the only way of guaranteeing progress in science. In this respect, we do not follow Popper. His position is prescriptive and hence not falsifiable in itself (see e.g., Küpper 2011). Methodological choices are normative stipulations and therefore cannot be true in an objective sense.

way of developing workable models of the respective situations and thus improving the underlying theory's explanatory and predictive quality. The following proposition captures this methodological strategy[4]:

Proposition 1 *From a critical rationalist perspective, the principles of a theory constitute invariant and untestable elements of that theory. As such, they are not empirically true or false but represent a methodological stipulation.*

We are now ready to apply this proposition to our key problem: the methodological status of the *Homo economicus* model in a rational-choice approach to behavioral ethics.

3 The methodological principles of a rational choice approach to behavioral ethics

Now that we have clarified our epistemological background, we are ready to specify the methodological principles of the rational-choice approach to behavioral ethics that we propose. Given that our approach is rooted primarily in economics, our discussion will draw heavily on methodological works in that domain.

3.1 *Homo economicus* as a pre-empirical model

The basic argument we have put forward is that, generally, the principles of a theory constitute a methodological stipulation, rather than an empirical statement, and can therefore not be falsified empirically. We based this argument on Popper's conception of explanations in the social sciences (Popper 1959/2005). Applying our argument to the HEM, we further argued that this model's core assumptions represent *principles* of the rational-choice paradigm and should not be seen as empirical propositions. In that sense they are *pre-empirical*: they are not a statement on social reality per se, but provide guidance on how to look at reality (Homann 1994).

The interpretation of the HEM as a set of non-empirical principles has a long tradition in economics, where the concept of rational-choice is widely used (e.g., Keynes 1917; McKenzie 2009; Mill 1836/1967; Pareto 1907/1971). Two prominent advocates of rational-choice are Friedman and Becker. In his conceptualization of positive economics, Friedman (1953) assumes that people respond to different situations "as if" they were rational decision makers. Similarly, Becker (1976: 7) in his conception of economics as a general approach to human behavior, maintains stable preferences and rationality, and insists on holding on to these assumptions even when empirical observation contradicts theoretical predictions.

[4] For the sake of expositional clarity, we summarize our results in the form of a proposition. In a recent *Academy of Management Review* editorial, Cornelissen (2017) referred to this style of theorizing as the "proposition-based style." In contrast to that, our propositions do not introduce "cause–effect relationships" but outline the cornerstones of our proposed methodology.

In more recent discussions some economists have also suggested that the HEM's core assumptions are not testable propositions about human psychology. For instance, Boland (1981: 1031) characterized the theoretical core assumptions of neoclassical economics as metaphysical and argued that "no criticism of [the neoclassical] hypothesis will ever be successful." Another example comes from Vanberg (2004), who proposed that rationality can be interpreted in two distinct ways, as a (non-refutable) principle and as a testable hypothesis. Interpreted as a principle, Vanberg (2004: 3) argued, rationality "does not qualify as an empirically contentful, refutable conjecture." As these references demonstrate, the HEM should *not* be understood as a "psychological model" (Tomasello 2016: 158; similarly: Donaldson 1990) of human nature, but as a methodological heuristic. On that basis we conclude that the HEM is not an empirical concept and, thus, should not be viewed as an ontological statement or a model of human nature.

Proposition 2 *The HEM is a basic principle of the rational-choice approach. As such, it is not empirically refutable. Consequently, the HEM is not an either positive or normative model of human nature.*

3.2 Variant and invariant elements in scientific explanations

There is a second parallel between Popper's methodology and the methodology of standard economics: Popper (1964/1994) distinguished between the invariant and variant elements of an explanatory theory. Similarly, several economists distinguish between invariant preferences (as those assumed in the HEM) and variable constraints (Blaug 1992; Glass and Johnson 1988; Hahn and Hollis 1979).

Economics explains human behavior through a model that is based on specific assumptions: first, that people have constant preferences[5] and, second, that their preferences are subject to certain behavioral constraints, such as monetary and psychological costs. This model interprets behavior as an agent's rational response to a given situation. Figure 1 illustrates the interplay between the invariant assumptions of the HEM, a contingent model of the situation, and behavior (for an application of the same methodology in the management sciences, see Mackenzie and House 1978). Figure 1 also mentions the analogous elements in Popper's concept as described in the previous section.

Specifying how scientific explanations rest on irrefutable principles and of variable situational constraints on behavior allows us to conceptualize HEM's methodological status with precision. The assumptions underlying the HEM correspond to a methodological heuristic that helps explain human behavior.

[5] We should note that Becker's concept of 'preferences' does not refer to mere tastes, but to *fundamental* preferences, as reflected in the following quote: "The preferences that are assumed to be stable do not refer to market goods and services, like oranges, automobiles, or medical care, but to underlying objects of choice that are produced by each household using market goods and services, their own time, and other inputs. These underlying preferences are defined over fundamental aspects of life, such as health, prestige, sensual pleasure, benevolence, or envy, that do not always bear a stable relation to market goods and services" (Becker 1976: 5).

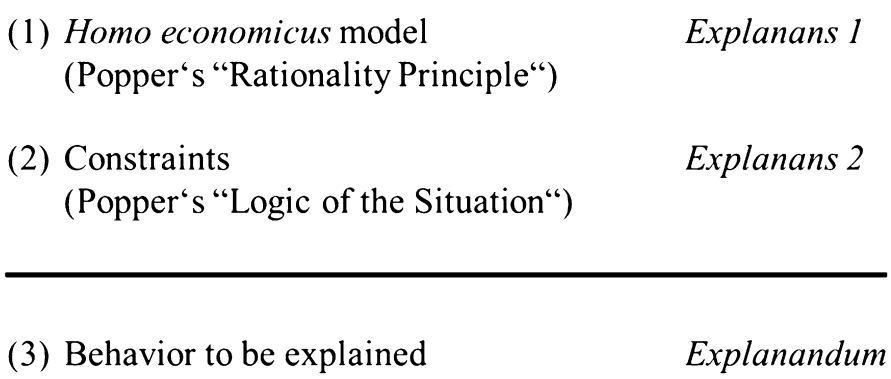

Fig. 1 Scientific explanation within a rational-choice approach

If the observed behavior is markedly different from the predicted behavior, the researcher is instructed to concentrate on stating precisely the empirical constraints that may explain why the theory failed to predict behavior correctly. In sum, Fig. 1 shows that in order to offer rigorous explanations of empirical phenomena (*Explanandum*), it is necessary to specify the empirical situation at hand (*Explanans 2*) and to apply the HEM (*Explanans 1*). As Vanberg (2004: 3) puts it, Popper's rationality principle is "a heuristic principle that tells us what we should look for when we seek to explain human action."

In the same vein, Becker explicitly advises against abandoning the rationality assumption when observations contradict theoretical predictions—a methodological strategy which, Becker laments, economists often use: "if some Broadway theater owners charge prices that result in long delays before seats are available, the owners are alleged to be ignorant of the profit-maximizing price structure rather than the analyst ignorant of why actual prices do maximize profits" (Becker 1976: 12). Such an approach, however, would not be rigorous. Economic explanations of such behavior, Becker (1976: 7) argued, should "not take refuge in assertions about irrationality," but should seek to model the constraints in a specific situation until the model's predictions are in line with empirical observations.

To illustrate the implications of the methodology we propose for interpreting moral behavior, we would like to draw on an analogy that the physicist and philosopher Carl Friedrich von Weizsäcker made. This analogy illustrates aptly the difference between the actual experience and the scientific reconstruction of the same phenomenon by means of a historical example: Aristotle's and Galileo's ways of explaining why objects fall.

> Aristotle says that heavy bodies fall fast, light bodies fall slowly, very light bodies will even rise. This is exactly what everyday experience teaches us; a stone will fall fast, a sheet of paper more slowly, a flame will even rise. Galileo says that all bodies fall with equal acceleration and will therefore after equal time have acquired equal velocity. In everyday experience this is just wrong. Galileo goes on to tell us that in a vacuum bodies would

> really behave like that. Here he states the hypothesis that there is a vacuum, an empty space, again contradicting not only Aristotle's philosophy but every-day experience. He was not able to produce a vacuum himself. But he greatly encouraged later seventeenth century physicists, like his pupil Torricelli, to make a vacuum; and in fact, when a sufficiently empty space was there, Galileo's prediction proved true. Further, his assertion opened the way for a mathematical analysis of buoyancy and friction, the two forces responsible for the different behaviour of falling bodies of different specific weights, sizes, and shapes. Only if you know how a body would fall without these forces will you be able to measure them by their impeding effect (Weizsäcker 1964: 104–5).

Galileo's laws of motion are counterintuitive because they contradict daily experience. Regardless of that, however, they constitute a general theory that instructs researchers where to look for the reasons that explain why objects do not fall as predicted. Using the HEM to build a rational-choice approach to behavioral ethics is very similar to this example. Interpreting observed moral behavior as a "falsification" of the HEM is like interpreting the slow fall of a feather as a falsification of Galileo's laws of motion. But this is not how physicists interpret apparent contradictions between observed and predicted phenomena. Instead of adjusting the laws of physics to explain why objects fall in ways that contradict the relevant laws, they explain these discrepancies on the basis of the situational conditions that affect the speed with which objects fall (in this example: air resistance).

We propose that this methodological strategy can be transferred to the HEM: when observed behavior seems to contradict the assumptions and predictions of the model, the discrepancy should not be interpreted as a falsification of the underlying theory's principles. Rather, discrepancies between predicted and observed behavior should prompt us to investigate which constraints may have been responsible for the behavior we observe. As Boland (1981: 1035) notes in the context of economics:

> The research program of neoclassical economics is the challenge of finding a neoclassical explanation for any given phenomenon—that is, whether it is possible to show that the phenomenon can be seen as a logical consequence of maximizing behavior—thus, maximization is beyond question for the purpose of accepting the challenge.

Consequently, if the participants in laboratory experiments behave morally and if this behavior seems to contradict the assumptions of the HEM, we are well advised not to abandon the model's assumptions but to examine the empirical constraints to find out whether these can explain the agent's actual behavior.

We can synthesize the points we have made thus far to formulate a central thesis of our paper as follows: taking a rational-choice approach to behavioral ethics implies that variances in moral behavior should not be regarded as evidence of variances in preferences (e.g., for honesty, trust, or reciprocity), but as a reason for asking which situational constraints may have led to the observed behavior.

Proposition 3 *A rational-choice approach to behavioral ethics implies that situational constraints, rather than faulty assumptions, explain discrepancies between the predictions of the Homo economicus model and observed behavior.*

4 A central problem in behavioral ethics: cooperation, social dilemmas and the limits of morality

Now that we have specified the methodological principles of our rational-choice approach, we can turn our attention to another central question: what makes this approach useful for studying morality? Our answer, which we will expand on below, is that a rational-choice approach to behavioral ethics, with HEM at its methodological core, facilitates the rigorous analysis of human cooperation and cooperation, in turn, plays a fundamental role in ethics.

Since Aristotle's (1925/1998) *Nicomachean Ethics*, the prevalent view has been that the ultimate goal of ethics is to contribute to *Eudaimonia*, or human flourishing. In modern terms, the goal of ethics is to improve everybody's prospects for peace and fulfilled lives, whatever that may mean from each individual's perspective. Generally, individuals cannot pursue their goals without at least some degree of cooperation from other people (Buchanan 1995). At the same time, any form of morality involves the consideration of others. Consequently, taking a behavioral approach to ethics involves developing a theoretically consistent explanation for the reasons that lead humans to cooperate, the reasons for which cooperation may fail and the conditions under which cooperation occurs or fails (Homann 2014).

In order to specify these conditions, we will begin with the most basic case—that of non-cooperation.[6] Once we have explained theoretically why agents fail to cooperate, we can go on to identify the factors that help overcome this failure and enable cooperation. Reconstructing the emergence of cooperation and morality from a non-cooperative situation has a long tradition in economics-based treatises on cooperation and morality (e.g., Brennan and Buchanan 1985; Gauthier 1986; Greif 2000; Ostrom 2000).

So why do people fail to cooperate? From a rational-choice perspective, the most general answer is, because of *competition*. Human interaction is constrained by scarcity. As a result, people compete with each other for scarce resources. This competitive structure of interaction is particularly pronounced in market economies. Various mathematical models of social dilemmas have been developed in game theory to illustrate how competitive interactions affect potential cooperation. In these models, agents have an interest to cooperate but fail to do so because of the underlying social dilemma.

At this point, it is important to note that cooperation need not be morally good or socially desirable. Often communities deliberately install dilemma structures to prevent cooperation among competitors. Collusion is a case in point: in

[6] Note that this methodological choice does not deny that evolutionary cooperation may have developed first (Tomasello 2009).

modern market economies, the participants are not allowed to coordinate their prices, because price competition is believed to yield socially desirable results. More drastic examples include criminal organizations: to pursue their illegal and immoral goals, members of the Mafia need to cooperate. Such examples show that not every form of cooperation is ethically valuable.

In some cases, however, dilemma structures prevent cooperation that would be desirable. In economics, models of public goods or the "tragedy of the commons" are classic examples. In any case, cooperation mainly fails because of social dilemmas. From a rational-choice perspective, to understand cooperation it is necessary to understand in depth how social dilemmas function. One model that has often been used to analyze human cooperation is the so-called "prisoners' dilemma" (Axelrod 1984; Gauthier 1986; Greene 2013). For the sake of simplicity, we will concentrate on the prisoners' dilemma, although our arguments also hold for other classes of social dilemma models (such as public goods or common pool resources). Poundstone (1992: 21) provides a concise description of the situation:

> Two members of a criminal gang are arrested and imprisoned. Each prisoner is in solitary confinement with no means of communicating with the other. The prosecutors lack sufficient evidence to convict the pair on the principal charge. They hope to get both sentenced to a year in prison on a lesser charge. Simultaneously, the prosecutors offer each prisoner a bargain. Each prisoner is given the opportunity either to: betray the other by testifying that the other committed the crime, or to cooperate with the other by remaining silent. The offer is:
>
> - If A and B each betray the other, each of them serves 2 years in prison
> - If A betrays B but B remains silent, A will be set free and B will serve 3 years in prison (and vice versa)
> - If A and B both remain silent, both of them will only serve 1 year in prison (on the lesser charge)

We note three points that are most important for our purposes. First, although each agent decides individually, their decisions are interdependent because each agent's decision affects everybody else's pay-off. Second, the agents have both shared and conflicting interests simultaneously. By means of cooperation, all agents could improve their situation and serve their common interests. At the same time, because of conflicting interests, the incentives are such that if one agent cooperates, the other is tempted *not* to cooperate but, instead, to improve his or her own pay-off. Third, the existence of conflicting interests implies that there is a risk of competitive exploitation. Agents who choose to cooperate make themselves vulnerable to the competitive behaviors of others. It is exactly an agent's willingness to cooperate that allows others to free-ride and exploit cooperation. Consequently, the only way to protect oneself from competitive exploitation is defection.

The well-known theoretical result of social dilemmas is non-cooperation. Hobbes (1651/2005) summarized this state of affairs as a "war of all against all." This state of affairs forms our methodological point of departure.

Proposition 4 *Methodologically, in a rational-choice approach to behavioral ethics, social dilemmas represent the basic form of human interaction.*

The critical reader may counter that social reality includes many examples of successful cooperation, so a methodology based on the assumption that social dilemmas are the fundamental form of human interaction must be flawed. To address this potential objection we refer to Popper's critical rationalist methodology. Again, our objectives in this paper are theoretical, not phenomenological. That is, we do not deny that there are many situations in the real world where interacting agents cooperate successfully without becoming trapped in social dilemmas. However, instances of successful cooperation are exactly what we aim to explain. Cooperation is our *explanandum*, not the *explanans*. From a rational-choice perspective then, our starting point is non-cooperation. The heuristic use of social dilemmas (as the explanation for failed cooperation) is pre-empirical in that it guides our scientific perspective on reality. As such, it is a methodological stipulation and it should not be confused with an ontological statement on social reality per se.

As the above explanation shows, the assumption that social dilemmas are the basic form of human interaction is a contingent methodological decision—other theories may be based on different assumptions. For the rational-choice approach we propose, we chose the non-cooperation that results from social dilemmas as the basic form of interaction because it is *useful*, not because it is *true* in an empirical sense. For our purposes, this choice is useful because it directs attention to the factors that make cooperation possible despite social dilemmas (see, e.g., McKenzie 2009). As we will discuss in more detail below, various strategies exist to overcome social dilemmas and thus enable cooperation. Such strategies include the establishment of institutions such as coded law as well as various kinds of formal and informal agreements that facilitate cooperation in spite of social dilemmas (Greene 2013; Haidt 2008; Heath 2014). In sum, the assumption of omnipresent social dilemmas is the adequate analytical heuristic for our purposes because it illuminates how social dilemmas can be overcome for the sake of cooperation.

Galileo's laws guided physicists to ask why different objects fall at different speeds and to look for situational factors that determine the speed at which objects fall, rather than adjusting the laws to fit their observations. Analogously, the methodology we propose stipulates that instances of observed cooperation should not be attributed to the absence of social dilemmas, but to the presence of institutions that successfully overcome social dilemmas. In the absence of such institutions, the social dilemmas will hinder cooperation.

The assumption of omnipresent social dilemmas is important for the aims of this article because it justifies the use of the *Homo economicus* model. Our argument is that *Homo economicus* is the preeminent, indispensable analytical tool for predicting the *aggregate* behavioral consequences of social dilemmas. In this sense, what we propose is a "micro-funded macro theory" (Zintl 1989): although *Homo economicus* is a model of individual behavior, its explanatory and predictive power unfolds at the macro-level.

Proposition 5 *The use of the Homo economicus model is justified not because the model is empirically valid at the individual level, but because it helps predict reliably the behavioral consequences of social dilemmas at the macro level.*

In the context of the prisoners' dilemma, the interacting agents are assumed to maximize their individual pay-offs and the standard prediction is that they will choose the defection strategy (non-cooperation). Indeed, experiments have shown that when a lab setup of the prisoners' dilemma approximates sufficiently the theoretical model, the vast majority of participants defect (Andreoni and Miller 1993 for just one example).

Each agent may choose to defect (i.e., not to cooperate) for one of two reasons. Thomas Hobbes was the first to identify these reasons. The first is "glory,"[7] or the "pleasure of superior power with respect to others" (Slomp 1990: 76); the second is "diffidence." The latter defective strategy is a response to the social dilemma that this situation entails, because if an agent chooses to cooperate, he or she risks being exploited by a competing agent. In other words, if one agent is worried that the other agent will defect, the only way to defend himself or herself is to defect preemptively. It is surely no coincidence that Hobbes used the term "defensio" in his own Latin translation of *Leviathan* to describe the strategy of "diffidence." The term Hobbes chose shows that defection need not be an offensive strategy, but can also be a defensive and preemptive strategy in social dilemmas. In that sense, our rational-choice approach interprets diffidence as an incentive-induced imperative, rather than a genuine human motive; agents do not *act* but *react* like *homines oeconomici* to the (anticipated) behavior of others whose decisions may be to their disadvantage.

Several behavioral economists also take this view. Even though these economists posit that the participants in laboratory experiments have social preferences, they acknowledge that the latter are sometimes unable to act in line with their preferences. As Fehr and Schmidt (1999: 834) state in their analysis of the preference for equity: "It is, thus, the impossibility of preventing inequitable outcomes by individual players that renders inequity aversion unimportant in equilibrium." Another example comes from the seminal study of Fehr and Gächter (2000) who showed in the context of an experiment how the possibility to peer-punish free-riders can stabilize cooperation. Most importantly, the authors showed that the *same* participants in the experiment varied their cooperating behavior dramatically, depending on whether there was an option to punish free-riders or not.

In our view, these results suggest that the participants may have been *willing* to cooperate, but, given how risky cooperative behavior is in dilemma situations, they were *unable* to ensure collective cooperation. When the option of punishing free-riders was not available, the participants were aware that anyone who cooperated might be exploited by free-riders who took advantage of less competitive peers. As a result, almost everyone defected. Once the participants were presented with the option to punish free-riders, they could protect themselves from competitive

[7] In *Elements of Law* Hobbes (1650/1994: 50) states: "GLORY […] is that passion which proceedeth from the imagination or conception of our own power, above the power of him that contendeth with us."

exploitation and enforce the cooperative norm in their group (similarly, Gächter et al. 2008).

We take such results of experimental economics to conclude that, ultimately, the question of moral behavior is not a matter of individual dispositions (i.e., an agent's *willingness*), but one of structural conditions (i.e., an agent's *ability*). This is why the HEM is the adequate tool for analyzing behavior in social dilemmas, although this model does not describe how humans actually are. From this perspective, immoral behavior does not reflect moral deficiencies—but it is a rational reaction to prevalent dilemma structures. Thus, the methodological focus of our rational-choice approach to behavioral ethics is not on individual motives, but on the situational constraints that are responsible for the failure of cooperation.

Proposition 6 *A rational-choice approach to behavioral ethics interprets failed cooperation (immoral behavior) as the result of social dilemmas, not of moral deficiencies.*

Although the failure of cooperation served as our methodological point of departure, our rational-choice approach to behavioral ethics also needs to explain the various forms of cooperation we observe both in the real world and in controlled experiments. Explaining deviations from the basic state of affairs—the failure of cooperation—is instrumental to one of the most important goals in ethics: the facilitation of cooperation.

5 Explaining the existence of morality in the world of homo economicus

Unsurprisingly, experimental research has shown that non-cooperation is not, in fact, the rule. Human beings do act morally in various ways even when this behavior does not maximize their monetary returns in the lab. Many participants in behavioral experiments cooperate and are willing to incur costs so that free-riders are punished in exchange; many behave fairly, altruistically, reciprocally and honestly (Ariely 2011; Chaudhuri 2011; Güth and Kocher 2014). Given that *Homo economicus* has no ex ante morality, how can we explain these behaviors without abandoning our model's assumptions?

First, it is important to note that the behaviors we observe in the lab rest on an array of presuppositions. Experiments do not take place out of context. Participants are human beings with their own morality, some of it innate, some of it learned. They bring these "homemade" dispositions into the lab (Andreoni and Miller 1993; Camerer and Weigelt 1988) and many refuse to abandon their long-practiced behaviors just because a laboratory experiment offers minor short-term benefits (Andreoni and Samuelson 2006; Axelrod 1980; Nowak et al. 2000; Sterelny et al. 2013).

A rational-choice approach to behavioral ethics can illuminate moral behavior in two important ways. First, it can help explain the evolution of human moral dispositions and formally describe their various forms; second, it can elucidate which

conditions are responsible for the stability and erosion of moral dispositions and how these conditions produce such effects. In the following sections we will elaborate on both of these points.

5.1 Morality as individual disposition

As we explained, the participants in behavioral experiments bring into the lab their "homemade" moral dispositions which seemingly contradict the assumptions of HEM. But where do these dispositions arise? If we apply the HEM framework to the evolution of morality, it emerges that it may simply have been rational for humans to develop and adopt social norms and thus to cooperate, share payoffs and tell the truth. This interpretation is rooted in the works of philosophy and economics that interpret morality as a useful human evolutionary *adaption*. According to David Gauthier, for example, human beings commit themselves to moral standards because this allows them to cooperate usefully with other market participants: "rational constraints on the pursuit of interest have themselves a foundation in the interest they constrain. Duty overrides advantage, but the acceptance of duty is truly advantageous" (Gauthier 1986: 2).

Recent works in anthropology and moral psychology support this perspective, arguing that "morality is a set of psychological adaptions that allow otherwise selfish individuals to reap the benefits of cooperation" (Greene 2013: 23; similarly: Haidt 2008: 70). Evolutionary explanations for the existence of morality include the works of Boehm (2012), De Waal (1996), Hauser (2006), Joyce (2007), Sober and Wilson (1999) and Tomasello (2016). The respective theories vary in their exact explanations, but they share the conviction that the evolution of human morality can be seen as an adaption that was advantageous to the species' survival. In this sense, human morality is a cultural achievement as a result of which our species is better off.

Discussing each of these theories in detail is beyond the scope of this study. For our purposes, it suffices to sketch the arguments of Tomasello (2016) as one representative example of this body of theories. On the basis of a comprehensive review of experimental research, Tomasello (2016) developed a theory that seeks to explain why humans—but not chimpanzees and bonobos, the two closest relatives of our species—developed a morality of sympathy and fairness. As captured in his "interdependence hypothesis" (Tomasello et al. 2012), his theory proposes that the evolution of morality was a uniquely human response to an increasing need for cooperation, first on the level of dyadic interaction and for the sake of foraging, and later on the level of cultural groups and in all domains of life. These forms of cooperation were necessary for survival and led to an unprecedented level of interdependence.

According to what Tomasello (2016: 5) calls a "hypothesized natural history" of human morality, humans, in response to the high level of interdependence that resulted from (mostly dyadic) cooperation some 400,000 years ago, developed a set of cognitive skills; notably, joint intentionality, self-domestication and second-personal agency (Tomasello 2014, 2016). These skills had evolutionary advantages as they reduced the risks involved in collaborative hunting—such as the risk that the

partner abandons the hunting early, or that he seeks an uneven split of the spoils. The newly developed cognitive skills allowed individuals to construct a sense of "we" with hunting partners, which enabled them to consider the partner's interests as equally important as one's own. At later stages of human evolution, when humans started living in larger groups some 150,000 years ago, cultural practices and shared social norms emerged. These practices developed into an objective morality, in the sense that notions of right and wrong were not confined to the social group, but were seen as objectively right (Tomasello 2016). In sum, the theory's central claim is "that the skills and motivation to construct with others an independent, plural-agent "we" (…) are what propelled the human species from strategic cooperation to genuine morality" (Tomasello 2016: 4).

The case for an evolutionary account of morality is far from obvious. There is considerable evidence that humans often act cooperatively and even altruistically, although this behavior does not offer any sort of monetary, reputational or other reward at all (e.g., Ariely 2011; Chaudhuri 2011; Güth and Kocher 2014). At first sight, it may appear like a paradox to assume that it was for selfish reasons that human beings developed the capacity for such unselfish behavior. As Frans de Waal writes in the introduction to his book on the evolution of human morality: "We are facing the profound paradox that genetic self-advancement at the expense of others—which is the basic thrust of evolution—has given rise to remarkable capacities for caring and sympathy" (De Waal 1996: 5). So how can we dissolve the apparent tension between the assumption of selfishness and the empirical observation of genuinely moral behavior?

To see why it is no contradiction between the HEM's assumption that humans act as they do out of self-interest and observing altruism in the lab, it is useful to draw a distinction between two types of mechanisms that shape behavior: ultimate and proximate mechanisms. This distinction goes back to biologist and Nobel Prize Laureate Nikolaas Tinbergen (1963) and is now a well-established concept in contemporary behavioral sciences (Scott-Phillips et al. 2011).[8] While *ultimate* mechanisms describe evolutionary explanations for certain adaptions, *proximate* mechanisms explain why in a certain situation a given organism behaves in a certain way. Ultimate mechanisms generally refer to the evolutionary advantages that a certain behavior has for a species in terms of Darwinian fitness. On the individual level, however, actors are motivated by proximate mechanisms and are mostly unaware of their behavior's ultimate evolutionary background. In the actor's mind, proximate reasons become a goal in themselves (Tomasello 2016: 47).[9]

[8] We are grateful to Alicia Melis for drawing our attention to this distinction.

[9] Similarly, economics draws a difference between an explanatory model of behavior and the agent's perception of that same behavior (his *Lebenswelt*): "The critics of rational choice invariably—and I mean invariably—misrepresent the theory. In particular, it does not imply that rational actors are egoists, or that they maximize pleasure, or in fact, that they maximize anything. It is useful to keep in mind at all times that the rational choice model is a key tool of animal behavior theory (…). It is difficult to consider a creature lacking nociceptors (e.g., most insects) as a happiness maximizer, and yet the rational actor model is very illuminating even for such creatures. They maximize fitness" (Gintis 2016: vii).

The same is true for morality: even though the adaption of, say, altruistic attitudes has self-serving evolutionary origins, morality is not limited to selfishness and strategic reasoning. On the proximate level, morals can not be reduced to strategic considerations: in the eyes of decision-makers, their altruistic behavior is *genuinely*, as opposed to instrumentally, moral. So when participants enter the lab, their proximate moral motivations may compete with proximate selfish motivations.

Applying the distinction between ultimate and proximate reasons to the case of morality demonstrates that a rational-choice approach to behavioral ethics need not be reductionist. Suggesting that the evolution of moral disposition in humans is a self-serving adaption does not imply that individual agents perceive their moral acts as advantageous. "As moral beings, we may have values that are opposed to the forces that gave rise to morality. To borrow Wittgenstein's famous metaphor, morality can climb the ladder of evolution and then kick it away" (Greene 2013: 25). Hence, there is no contradiction between assuming that morality is a self-serving evolutionary adaption and accepting that an altruistic act is genuinely moral which can run counter to one's self-interest in a given situation (Tomasello 2016: 149).

The three main points of the discussion so far are the following: First, a rational-choice approach to behavioral ethics is in line with evolutionary explanations for the existence of morality. Second, the distinction between ultimate and proximate reasons shows why the empirical finding that many participants in laboratory experiments behave in altruistic ways does not contradict the HEM's assumptions. Third, our approach is not reductionist: morality may be an adaption that proved beneficial in the course of human evolution, but, once adopted, it does not boil down to strategic reasoning.

Proposition 7 *With regard to the ultimate mechanisms, a rational-choice approach to behavioral ethics interprets the evolution of morality as a beneficial disposition that facilitated human cooperation. With regard to the proximate mechanisms, this view does not imply that morality is reduced to strategic considerations.*

5.2 The stability and erosion of morals: the role of institutions

In the preceding section we argued that a rational-choice approach is fully in line with the view that humans developed a set of useful moral dispositions (Greene 2013; Haidt 2012; Tomasello 2016) which, in effect, lead to behavior that does not correspond to the HEM. Does this mean we should abandon the HEM in favor of new, more realistic models of 'social preferences' as some experimental economists seem to suggest (Fehr and Schmidt 2003; Korth 2009)? As we will argue in this subsection, we believe that the answer should be 'no.' Models of moral behavior that are based on social preferences are not alternatives to but refinements of the HEM. In the following paragraphs we will develop this argument step by step.

In the 1990s, behavioral economists began to develop formal models of 'social preferences.' Their starting point is the observation that the standard HEM fails to account for moral behaviors observed in the lab. These models aim to incorporate various moral dispositions such as fairness, altruism and reciprocity (for an early

review, cf. Fehr and Schmidt 2003). Their main purpose is to explain better behavior that is observed in the lab. For example, Fehr and Schmidt (1999) suggested a formal theory of inequity aversion that explains actual behavior in bilateral bargaining situations such as the ultimatum game better than standard economic theory based on the HEM.

In contrast to the HEM, models of social preferences formally describe moral motivations on the proximate level of human behavior. These models are very useful in that they—almost like psychological theories—aim to present certain moral dispositions and the consequences they should have, e.g., for designing incentives in markets and organizations (e.g., Fehr and Fischbacher 2002). Note that when researchers use such models to analyze the consequences of human morality, they treat morality exogenously, as an *explanans*.

The crucial point is, however, that the moral behaviors we may observe in the lab do not persist unconditionally. In the context of standard dilemma experiments, for example, even if levels of cooperation are initially high, they tend to go down within just a couple of periods (for just two examples, cf. Andreoni and Miller 1993; Fehr and Gächter 2000). Our interpretation of such experimental evidence is this: When coming into the lab with their pre-existing moral convictions, people routinely apply the judgments that have proven useful in their experience (Greene et al. 2004; Haidt 2001). So, initially, they may act morally for the sake of cooperation, and they may even be willing to favor moral choices over personal advantages. However, if they experience repeatedly the trade-offs between morality and personal advantage that are typically involved in social dilemmas, most people begin to reflect on their initial behavior. That is, they adapt their convictions of what is adequate and inadequate behavior given the situation at hand.

Earlier, we hypothesized that morality rests on a rational foundation; that is, that its existence hinges upon an evolutionary advantage. More specifically, we suggested that on the proximate level agents may not be aware of the evolutionary background of their moral minds and their moral motivation is genuine in that it goes far beyond strategic considerations. However, the links between advantages on the ultimate level and behavioral motivations on the proximate level become apparent when the logics of these levels clash. This occurs when, in social dilemmas, acting morally is exploitable and systematically disadvantageous. If it is true that moral dispositions evolved because of their evolutionary advantages, they will likely cease to exist if their beneficial foundation erodes (Pies and Hielscher 2014).

What we hence need is a contingency theory of moral behavior—a theory which, along with the description of moral dispositions, specifies the conditions under which agents are likely to act in accordance with these dispositions. Behavioral economics is well equipped to accomplish exactly that, because it has succeeded in identifying various conditions under which patterns of moral behavior are likely to persist—or erode.

Most important for the purposes of the present paper are the works of experimental economics that shed light on how various institutions help agents overcome dilemmas and enable cooperation. At the heart of the concept of an institution lies the notion of behavior that complies with rules (Langlois and Hodgson 1992: 165). Institutions coordinate the actions of agents by means of orienting these agents to

a common understanding and evaluation of a situation. Thus, they make behavior more predictable as they contain and transmit knowledge that helps people to interpret various situations and to find ways and means to cooperate (Greif 2000; North 1991). Institutions facilitate cooperation when they solve the problem of competitive exploitation; that is, when they change the incentives that a particular context offers to the participants in a way that cooperators do not bear the risk of being exploited by free-riders, either because cooperation is rewarded or because free-riding is punished (not necessarily in monetary terms).

A few examples may illustrate the value of the findings of behavioral economics for a rational-choice approach to behavioral ethics. A wide range of studies in experimental economics has compared different institutional arrangements with regard to their capability to enforce socially accepted norms such as cooperation. For example, Andreoni et al. (2003) conducted variants of a proposer–responder game to analyze how punishments and rewards can stabilize cooperation among the players. Based on the results of their experimental study, the authors found that the participants achieved the highest levels of cooperation and social welfare when they had the option of rewarding or punishing their peers. Similarly, Gächter et al. (2008) concluded from their experiments on public goods that having the option of punishing others in long-term interactions leads to greater cooperation and higher profits. Also, Casari and Luini (2009) showed that having the option of punishing others helps foster cooperation when this is supported by the majority of a group. These and similar works provide important insights into the conditions under which mutual punishment may function as a norm-enforcing institution.

We argued that, from the perspective of the participants in laboratory experiments, institutions such as punishment (e.g., the punishment of free-riders) offer them the opportunity to protect themselves in social-dilemma situations. With this in mind, we would expect that, if the participants in experiments are given the opportunity to establish such institutions, many participants would be willing to take it even when it is costly. Indeed, a recent strand of experimental research suggests that, under certain conditions, the members of a group are willing to invest and engage in building such an institution (Gürerk et al. 2006, 2014; Kosfeld et al. 2009; Putterman et al. 2011; Sutter et al. 2010; Traulsen et al. 2012; Zhang et al. 2014).

One example is the experiment of Andreoni and Gee (2012), who explored how two alternative and endogenously chosen institutions of control influence the levels of cooperation among the participants in a public-goods game. In the case of peer-to-peer punishment, participants could choose how much to contribute to the public good. In addition, they could penalize each other for uncooperative behavior. Alternatively, they could centralize the right to punish and delegate it to an independent policing agency (called "common pool institutions", cf. Guala 2012, p. 12). In the latter case, the player with the lowest contribution to the public good automatically receives a penalty, making it optimal for each player to provide the second-lowest contribution. Collectively, the players' intention to contribute more than the lowest contribution causes an upward spiral, ultimately leading to relatively high contributions.

As these examples of institutions show, the findings of experimental economics are of utmost importance to a rational-choice approach to behavioral ethics because

they shed light on how institutions that can overcome dilemmas and enable cooperation emerge and function. These insights illuminate the conditions under which it is plausible to expect behavior that systematically deviates from the HEM's predictions.

Proposition 8 *Experimental economics is relevant to a rational-choice approach to behavioral ethics because it helps understand the situational conditions under which different forms of morality emerge, persist and erode.*

In conclusion, approaches to moral behavior that are based on social preferences are not alternatives to but refinements of the HEM. In our view, the purpose of these approaches is to contribute to a contingency theory of moral behavior, rather than falsifying and replacing the HEM.

6 Normative implications

Positive and normative statements are different in many respects (Küpper 2011, 2018). Our proposed rational-choice approach to behavioral ethics constitutes positive research and, as such, has no direct normative consequences. The HEM is a positive analytical tool designed to model behavior in social dilemmas. We believe that it should not be used as a normative decision-making principle to guide human behavior (Harsanyi 1977: 16; Sugden 1991: 752). However, our positive analysis clearly has indirect normative implications for business ethics. A full discussion of these normative implications is beyond the scope of this paper, but we would like to use the concluding section to summarize the most important of these.

The first implication follows from the fact that every participant in a social interaction is able to affect its collective outcome (Homann and Suchanek 2005: 424). As we argued in detail above, every participant in social interactions that involve dilemmas can prevent the socially optimal outcome by choosing to defect. Moreover, a single free-rider can drive all other participants to what we called preemptive defection. In sum, any individual can inhibit cooperation, on which, however, everybody depends to achieve his or her individual goals. Thus, in social interactions, the success of any single agent depends at least in part on others cooperating (Buchanan 1995).

This destructive potential of every participant in an interaction can be used to justify basic human rights. If all individuals have the power to affect a society's outcome significantly, the members of a social group are well advised to take into account every other member's interests and to acknowledge the other members' legitimate claims in order to secure their cooperation. Once the destructive potential of others has become apparent, granting them basic rights becomes a matter of prudence. In the context of business ethics this argument can serve as an economic justification of stakeholder rights.

The second normative implication of our approach refers to how normative ethics should deal with uncooperative behavior that violates norms. Drawing on classic

Hobbesian contract theory, we argued that in general there are two reasons for defection in social dilemmas: greed (*glory*), and the desire to protect oneself from competitive exploitation (*diffidence* or *defensio*). Although both lead to the same behaviors, they differ in terms of the underlying ethics. While greed is a character flaw or a lack of virtue, preemptive defection results from a flaw in the structure of an interaction.

From our perspective, the results of behavioral economics suggest that in dilemma situations the participants in laboratory experiments defect preemptively. To begin with, many participants, being morally minded, follow a cooperative strategy. However, if they become exploited by others repeatedly, they eventually defect periods (Andreoni and Miller 1993; Fehr and Gächter 2000). There may always be some people who refuse heroically to defect in the face of a dilemma, but the majority respond with defection when they are repeatedly exploited. If it is true that one central reason for defection is self-protection, the normative thrust of ethics should shift away from character and focus more on institutional design.

On the basis of our positive analysis, we conclude that normative ethical propositions need to be backed up institutionally. Ethical propositions need to be accompanied by institutions that support and enable moral behavior by overcoming the problem of exploitation. Formal and informal institutions need to provide incentives that make defection a non-beneficial strategy. Put differently, any rule change that normative ethics proposes should pass what we might call an "HEM-test." Only proposed changes that lead to socially desirable results under the assumptions of the HEM should be implemented. For institutions that have passed the HEM-test, moral choices will lead to advantages (in a broad sense) and immoral choices will lead to sanctions (also in a broad sense). In essence, from the perspective of our approach, one central task of normative ethics is to suggest formal and informal institutional designs that prevent the exploitation of moral behavior.

Seen through the lens of a rational-choice approach to behavioral ethics, the risk of being exploited is the biggest obstacle to moral behavior. And in the long run, individual heroism will not be able to compensate for the structural deficits of social dilemmas. Any normative approach to ethics needs to account for this matter of fact, because failing to appreciate this fact may lead "to misguided efforts to attain positions that may be imagined but that are beyond the limits of behavioral feasibility" (Buchanan 1995: 141).

References

Anderson E (2000) Beyond *H. economicus*: new developments in theories of social norms. Philos Public Aff 29(2):170–200

Andreoni J, Gee LK (2012) Gun for hire: delegated enforcement and peer punishment in public goods provision. J Public Econ 96(11):1036–1046

Andreoni J, Miller JH (1993) Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. Econ J 103(418):570–585

Andreoni J, Samuelson L (2006) Building rational cooperation. J Econ Theory 127(1):117–154

Andreoni J, Harbaugh W, Vesterlund L (2003) The carrot or the stick: rewards, punishments, and cooperation. Am Econ Rev 93(3):893–902

Ariely D (2011) The (Honest) truth about dishonesty. How we lie to everyone—especially ourselves. HarperCollins, New York, p 2011
Aristotle (1925/1998) The nicomachean ethics. translated by ross. Oxford University Press, New York
Axelrod R (1980) Effective choice in the prisoner's dilemma. J Confl Resolut 24(1):3–25
Axelrod R (1984) The evolution of cooperation. Basic Books, New York
Bauman Y, Rose E (2011) Selection or indoctrination: Why do economics students donate less than the rest? J Econ Behav Organ 79(3):318–327
Becker GS (1976) The economic approach to human behavior. Chicago University Press, Chicago
Becker GS (1993) Nobel lecture: the economic way of looking at behavior. J Polit Econ 101(3):385–409
Blaug M (1992) The methodology of economics: or, how economists explain. Cambridge University Press, Cambridge
Boehm C (2012) Moral origins. The evolution of virtue, altruism, and shame. New York (Basic)
Boland LA (1981) On the futility of criticizing the neoclassical maximization hypothesis. Am Econ Rev 71(5):1031–1036
Brennan G, Buchanan JM (1985) The reasons of rules. Cambridge University Press, Cambridge
Buchanan JM (1995) Individual rights, emergent social states, and behavioral feasibility. Ration Soc 7(2):141–150
Camerer C, Weigelt K (1988) Experimental tests of a sequential equilibrium reputation model. Econometrica 56(1):1–36
Casari M, Luini L (2009) Cooperation under alternative punishment institutions: an experiment. J Econ Behav Organ 71(2):273–282
Chaudhuri A (2011) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Exp Econ 14(1):47–83
Churchich N (1994) Marxism and morality: a critical examination of marxist ethics. James Clarke & Co., Cambridge
Combs JG, Jr K, David J (1999) Can capital scarcity help agency theory explain franchising? Revisiting the capital scarcity hypothesis. Acad Manag J 42(2):196–207
Cooper DJ, Kagel JH (2016) Other-regarding preferences: a selective survey of experimental results. In: Kagel JH, Roth AE (eds) Handbook of experimental economics, vol 2. Princeton University Press, Princeton, pp 217–289
Cornelissen J (2017) Editor's comments: developing propositions, a process model, or a typology? Addressing the challenges of writing theory without a boilerplate. Acad Manag Rev 42(1):1–9
De Waal Frans (1996) Good natured. Harvard University Press, Boston
Di Stefano G, King AA, Verona G (2015) Sanctioning in the wild: rational calculus and retributive instincts in gourmet cuisine. Acad Manag J 58(3):906–931
Dierksmeier C (2011) The freedom-responsibility nexus in management philosophy and business ethics. J Bus Ethics 101(2):263–283
Donaldson L (1990) The ethereal hand: organizational economics and management theory. Acad Manag Rev 15(3):369–381
Donaldson T, Dunfee TW (1999) Ties that bind. A social contracts approach to business ethics. Harvard Business School Press, Boston
Dow SC (1997) Mainstream economic methodology. Camb J Econ 21(1):73–93
Etzioni A (2010) Behavioral economics: a methodological note. J Econ Psychol 31(1):51–54
Fehr E, Fischbacher U (2002) Why Social preferences matter—the impact of non-selfish motives on competition, cooperation and incentives. Econ J 112(478):C1–C33
Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. Am Econ Rev 90(4):980–994
Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114(3):817–868
Fehr E, Schmidt KM (2003) Theories of fairness and reciprocity: evidence and economic applications. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) Advances in economics and econometrics. Cambridge University Press, Cambridge, pp 208–257
Frank RH, Gilovich T, Regan DT (1993) Does studying economics inhibit cooperation? J Econ Perspect 7(2):159–171
Friedland J, Cole BM (2017) From homo-economicus to homo-virtus: a system-theoretic model for raising moral self-awareness. J Bus Ethics (online first). https://doi.org/10.1007/s10551-017-3494-6
Friedman M (1953) The methodology of positive economics. In: Friedman M (ed) Essays in positive economics. University of Chicago Press, Chicago, pp 3–43
Gächter S, Renner E, Sefton M (2008) The long-run benefits of punishment. Science 322(5907):1510

Gauthier D (1986) Morals by agreement. Clarendon Press, Oxford
Ghoshal S (2005) Bad management theories are destroying good management practices. Acad Manag Learn Edu 4(1):75–91
Gintis H (2016) Individuality and entanglement: the moral and material bases of social life. Princeton University Press, Princeton, p 2016
Gintis H, Bowles S, Boyd R et al (eds) (2005) Moral sentiments and material interests—the foundations of cooperation in economic life. MIT Press, Cambridge
Glass JC, Johnson W (1988) Metaphysics, MSRP and economics. Br J Philos Sci 39(3):313–329
Greene J (2013) Moral tribes. Emotion, reason, and the gap between us and them. Penguin Press, New York
Greene JD, Nystrom LE, Engell AD et al (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron 44(2):389–400
Greif A (2000) The fundamental problem of exchange: a research agenda in historical institutional analysis. Eur Rev Econ Hist 4(3):251–284
Guala F (2012) Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. Behav Brain Sci 35(1):1–15
Gürerk Ö, Irlenbusch B, Rockenbach B (2006) The competitive advantage of sanctioning institutions. Science 312(5770):108–111
Gürerk Ö, Irlenbusch B, Rockenbach B (2014) On cooperation in open communities. J Public Econ 120:220–230
Güth W, Kocher MG (2014) More than thirty years of ultimatum bargaining experiments: motives, variations, and a survey of the recent literature. J Econ Behav Organ 108:396–409
Hahn FH, Hollis M (1979) Philosophy and economic theory. Oxford University Press, Oxford
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108(4):814–834
Haidt J (2008) Morality. Perspect Psychol Sci 3(1):65–72
Haidt J (2012) The righteous mind: Why Good People are divided by politics and religion. New York (Pantheon)
Harsanyi JC (1977) Rational behavior and bargaining equilibrium in games and social situations. Cambridge University Press, Cambridge
Hauser MD (2006) Moral minds: How nature designed our universal sense of right and wrong. New York (HarperCollins)
Heath J (2014) Morality, competition, and the firm: the market failures approach to business ethics. Oxford University Press, Oxford
Hirschman AO (1985) Against parsimony: three easy ways of complicating some categories of economic discourse. Econ Philos 1(1):7–21
Hobbes T (1650/1994) The elements of law, natural and politic. In: Gaskin JCA (ed) Edited with an introduction. Oxford University Press, Oxford
Hobbes T (1651/2005) Leviathan. Continuum, New York
Homann K (1994) Homo oeconomicus und dilemmastrukturen. In: Sautter H (ed) Wirtschaftspolitik in offenen Volkswirtschaften - Festschrift zum 60. Geburtstag von Helmut Hesse. Göttingen, Vandenhoeck und Ruprecht, pp 387–412
Homann K (2014) Sollen und Können. Grenzen und bedingungen der individualmoral. Ibera/European University Press, Wien
Homann K, Suchanek A (2005) Ökonomik—Eine Einführung, 2nd edn. Tübingen, Mohr Siebeck
Hosmer LT, Chen F (2001) Ethics and economics. growing opportunities for joint research. Bus Ethics Q 11(4):599–622
Hühn MP (2014) You reap what you sow: How MBA programs undermine ethics. J Bus Ethics 121(4):527–541
Joyce R (2007) The evolution of morality. MIT Press, Cambridge
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–292
Kant I (1785/2013) Groundwork of the metaphysics of morals. In: Shafer-Landau R (ed) Ethical theory, an anthology. Wiley, Chichester **(Reprinted)**
Keynes JN (1917) The scope and method of political economy, 4th edn. Macmillan, London
Kirchgässner G (2008) Homo oeconomicus. The economic model of behavior and its applications in economics and other social sciences. Springer, New York

Kluver J, Frazier R, Haidt J (2014) Behavioral ethics for *Homo economicus*, *Homo heuristicus*, and *Homo duplex*. Organ Behav Hum Decis Process 123(2):150–158
Korth C (2009) Game theory and fairness preferences. In: Korth C (ed) Fairness in bargaining and markets. Springer, Berlin, pp 19–34
Kosfeld M, Okada A, Riedl A (2009) Institution formation in public goods games. Am Econ Rev 99(4):1335–1355
Langlois RN, Hodgson GM (1992) Orders and organizations: toward an austrian theory of social institutions. In: Caldwell B, Boehm S (eds) Austrian economics: tensions and new directions. Springer, Heidelberg, pp 165–192
Laville F (2000) Should we abandon optimization theory? The need for bounded rationality. J Econ Methodol 7(3):395–426
Lazear EP (2000) Economic imperialism. Quart J Econ 115(1):99–146
Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? J Econ Perspect 21(2):153–174
Levitt S, List JA (2008) Economics: *Homo economicus* evolves. Science 319(5865):909–910
Lindenberg S (1990) Homo socio-oeconomicus: the emergence of a general model of man in the social sciences. J Inst Theor Econ (JITE) 146:727–748
Loewenstein G, Thaler RH (1989) Anomalies: intertemporal choice. J Econ Perspect 3(4):181–193
Mackenzie KD, House R (1978) Paradigm development in the social sciences: a proposed research strategy. Acad Manag Rev 3(1):7–23
Marx K (1959) Economic & Philosophic Manuscripts of 1844, transl. and edited by Martin Milligan, Moscow (Foreign Languages Publishing House)
McKenzie RB (2009) Predictably rational? In search of defenses for rational behavior in economics. Springer, New York
McWilliams A, Siegel D (2001) Corporate social responsibility: a theory of the firm perspective. Acad Manag Rev 26(1):117–127
Meckling WH (1976) Values and the choice of the model of the individual in the social sciences. Schweizerische Zeitschrift für Volkswirtschaft und Statistik 112(4):545–560
Mill JS (1836/1967) On the definition of political economy. In: Robson JM (ed) Collected works, vol IV. University of Toronto Press, Toronto, pp 309–339
Mueller F (1995) Organizational governance and employee cooperation: can we learn from economists? Hum Relat 48(10):1217–1235
Néron P-Y (2015) Rethinking the very idea of egalitarian markets and corporations: Why relationships might matter more than distribution. Bus Ethics Q 25(1):93–124
North DC (1991) Institutions. The Journal of Economic Perspectives 5(1):97–112
Nowak MA, Page KM, Sigmund K (2000) Fairness versus reason in the ultimatum game. Science 289(5485):1773–1775
Ostrom E (1998) A behavioral approach to the rational choice theory of collective action: presidential address, 1997. Am Polit Sci Rev 92(1):1–22
Ostrom E (2000) Collective action and the evolution of social norms. J Econ Perspect 92(1):137–158
Pareto V (1907/1971) Manual of political economy (transl. by A. Schwier). London, MacMillan
Pies I, Hielscher S (2014) Verhaltensökonomik versus Ordnungsethik? Zum moralischen Stellenwert von Dispositionen und Institutionen. Zeitschrift für Wirtschafts- und Unternehmensethik 15(3):398–420
Pies I, Hielscher S, Beckmann M (2009) Moral commitments and the societal role of business: an ordonomic approach to corporate citizenship. Bus Ethics Q 19(3):375–401
Popper KR (1945/2011) The open society and its enemies. Routledge, London
Popper KR (1959/2005) The logic of scientific discovery. Routledge, London
Popper KR (1963/1985) The rationality principle. In: Miller DW (ed) Popper selections. Princeton University Press, Princeton, pp 357–365
Popper KR (1964/1994) Models, instruments, and truth. The status of the rationality principle in the social sciences. In: Popper KR, Mark AN (eds) The myth of the framework: in defense of science and rationality. Routledge, London, pp 154–184
Poundstone W (1992) Prisoner's dilemma: john von neuman, game theory, and the puzzle of the bomb. Doubleday, New York
Putterman L, Tyran J-R, Kamei K (2011) Public goods and voting on formal sanction schemes. J Public Econ 95(9):1213–1222

Schreck P, van Aaken D, Donaldson T (2013) Positive economics and the normativistic fallacy: bridging the two sides of CSR. Bus Ethics Q 23(2):297–329
Scott-Phillips TC, Dickins TE, West SA (2011) Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. Perspect Psychol Sci 6(1):38–47
Sen AK (1977) Rational fools: a critique of the behavioral foundations of economics. Philos Public Aff 6(4):317–344
Simon HA (1957) Models of man: social and rational. Mathematical essays on rational human behavior in a social setting. Wiley, New York
Slomp G (1990) The significance of glory in the political theory of thomas hobbes. UMI, London
Sober E, Wilson DS (1999) Unto others: the evolution and psychology of unselfish behavior. Harvard University Press, Cambridge
Sterelny K, Joyce R, Calcott B et al (eds) (2013) Cooperation and its evolution. MIT Press, Cambridge
Sugden R (1991) Rational choice: a survey of contributions from economics and philosophy. Econ J 101(407):751–785
Sutter M, Haigner S, Kocher MG (2010) Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. Rev Econ Stud 77(4):1540–1566
Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
Tinbergen N (1963) On aims and methods of ethology. Zeitschrift für Tierpsychologie 20(4):410–433
Tomasello M (2009) Why we cooperate. MIT Press, Boston, p 2009
Tomasello M (2014) A natural history of human thinking. Harvard University Press, Boston, p 2014
Tomasello M (2016) A natural history of human morality. Harvard University Press, Boston, p 2016
Tomasello M, Melis AP, Tennie C et al (2012) Two key steps in the evolution of human cooperation: the interdependence hypothesis. Curr Anthropol 53(6):673–692
Traulsen A, Röhl T, Milinski M (2012) An economic experiment reveals that humans prefer pool punishment to maintain the commons. Proc R Soc B 279(1743):3716–3721
Vanberg VJ (2002) Rational choice vs. program-based behavior: alternative theoretical approaches and their relevance for the study of institutions. Ration Soc 14(1):7–54
Vanberg VJ (2004) The rationality postulate in economics: its ambiguity, its deficiency and its evolutionary alternative. J Econ Methodol 11(1):1–29
Weizsäcker CF (1964) The relevance of science. Creation and cosmogony. Gifford lectures 1959–60. Collins, London
Zhang B, Li C, De Silva H et al (2014) The evolution of sanctioning institutions: an experimental approach to the social contract. Exp Econ 17(2):285–303
Zintl R (1989) Der *Homo Oeconomicus*: ausnahmeerscheinung in jeder Situation oder Jedermann in Ausnahmesituation? Analyse Kritik 11:52–69