Morality Beyond Social Preferences: Smithian Sympathy, Social Neuroscience and the Nature of Social Consciousness

Sylvie Thoron

Œconomia

History, Methodology, Philosophy

6-2 | 2016 :
Psychology and Economics in Historical Perspective
Psychology and Economics in Historical Perspective (Part 2)

Morality Beyond Social Preferences: Smithian Sympathy, Social Neuroscience and the Nature of Social Consciousness

La moralité au-delà des préférences sociales. La sympathie Smithienne, les neurosciences sociales et la nature d'une conscience sociale

SYLVIE THORON

p. 235-264

Résumés

English Français
The theory of social preferences expanded the definition of the utility function in order to reproduce the pro-social behavior observed in experiments. Does this then mean that this is the route towards a positive theory of morality in economics? We do not think so. Our claim is that there is an epistemic contradiction between methodological individualism which assumes that the economic agent's rationality is autonomous from society, and the nature of social consciousness. Therefore, we argue that a positive theory of morality should rely on social mechanisms that could not fit in this framework. We give two examples of lines of research that go in this direction. The first one is drawn from 18th century moral philosophy, this is the approach adopted by Adam Smith in *The Theory of Moral Sentiments*. The other one is drawn from a very recent domain of research, that of social neuroscience. We show how, in both cases, the objective is not only to understand how moral judgements shape behavior but also to get an understanding of how people form these moral judgements. Smith's thought and recent developments in social neuroscience seem to be mutually illuminating on this second aspect. Smith's model is an essential model of a moral agent embedded in society. The sympathy operator and the impartial spectator find an echo in the way in which social neuroscience tries to understand how emotional and cognitive empathy intermesh. Furthermore, social neuroscience attempts to go further in the understanding of the complex empathy mechanism, by considering it as a learning process.

La théorie des préférences sociales a élargi la définition de la fonction d'utilité de façon à pouvoir reproduire le comportement pro-social observé dans les expériences. Cela signifie-t-il pour autant que nous sommes sur la bonne voie vers l'élaboration d'une théorie positive de la morale en

économie? Nous considérons, au contraire, qu'il existe une contradiction épistémique entre l'individualisme méthodologique, qui suppose que la rationalité de l'agent économique est autonome par rapport à la société, et la nature d'une conscience sociale. Ainsi, nous montrons qu'une théorie positive de la morale doit reposer sur des mécanismes sociaux qui ne pourraient entrer dans ce cadre. Nous donnons deux exemples de lignes de recherche qui vont dans ce sens. Le premier est tiré de la philosophie morale du 18ème siècle, et il s'agit de l'approche adoptée par Adam Smith dans *La Théorie des sentiments moraux*. L'autre est tiré d'un domaine de recherche très récent, celui des neurosciences sociales. Nous montrons comment, dans les deux cas, l'objectif est de comprendre non seulement comment les jugements moraux modèlent les comportements, mais aussi la façon dont les gens forment ces jugements moraux. La pensée de Smith et les récents développements en neurosciences sociales semblent s'éclairer mutuellement au sujet de ce second aspect. Le modèle de Smith est un modèle essentiel d'un agent moral encastré dans la société. L'opérateur de sympathie et le spectateur impartial trouvent un écho dans la manière dont les neurosciences sociales cherchent à comprendre comment l'empathie émotionnelle et l'empathie cognitive s'articulent. En outre, les neurosciences sociales tentent d'aller plus loin dans la compréhension du mécanisme complexe de l'empathie, en le concevant comme un processus d'apprentissage.

Entrées d'index

Mots-clés : préférences sociales, économie expérimentale, sympathie, spectateur impartial, sentiments moraux, individualisme méthodologique, neuroscience, empathie émotionnelle, empathie cognitive
Keywords : social preferences, experimental economics, sympathy, impartial spectator, moral sentiments, methodological individualism, neuroscience, emotional empathy, cognitive empathy

Texte intégral

1 The theory of social preferences[1] was developed in order to provide an explanation for the experimental evidence of pro-social behavior that flourished during the 1980s and challenged the model of the selfish rational agent in mainstream economics. This theory gave the impression that it was not necessary to call into question the model of rationality to incorporate considerations of morality and justice into the standard theory. Indeed, the theory of social preferences expanded the definition of the utility function in order to reproduce the pro-social behavior observed in experiments and which is now regarded as a stylized fact.[2] Does this then mean that we are on the route towards a positive theory of morality in economics? By a positive theory of morality we mean a theory that can explain how, what is referred to by the generic term 'morality', shapes individuals' behavior in society. It is therefore not a normative theory but a theory that could incorporate some of the ideas from the behavioral sciences into economics. The claim of this paper is that, if we are to make real progress on this route, economics will have to incur a much higher epistemological cost. Indeed, we argue that methodological individualism based on the maximization of utility, inherent in standard economics, is a poor framework within which to understand the mechanisms that generate socially conscious behavior. In fact, there is an epistemic contradiction between methodological individualism, which assumes that the economic agent's rationality is autonomous from the society, and the nature of social consciousness.

2 In this paper, we propose first to see, through different examples, how the literature emerged from the dialog between experimental results and theory, without questioning the fundamental axioms of rationality developed by economists. We will highlight the difficulties of the different attempts to model a moral agent's rationality and the contradictions that resulted. The stability of social preferences that is assumed in this literature, in particular, is difficult to reconcile with the idea of a "dependence of justice evaluation on the context" (Konow, 2003, 1189) or with the idea that principles of justice are not necessarily immutable fairness ideals (Rodriguez and Moreno, 2012) and that moral principles in general are not innate but acquired by the agents. However, we will see that the assumption of stability of social preferences is fundamental in economics and has driven some well-known researchers to try to find an 'essential' model of the moral rational agent to represent the true "nature of human altruism" (Fehr and Fischbacher 2003). Of course one major difficulty in doing this is that there

are different principles, from social efficiency to inequity aversion and reciprocity among many others that could be candidates for this essential model of the theory of social preferences. We will also discuss an alternative and more modest approach, which consists in the proposal of models that are only valid in the 'short run'. Unfortunately, we will explain why, if economics wants to be a science of public decision, this will not be satisfactory either as an answer to the previous criticisms concerning the stability of social preferences. We will conclude this first section by explaining how the difficulties encountered in making progress result from the constraints of methodological individualism.

3 Then, we will argue that a positive theory of morality should rely on social mechanisms that could not fit in this framework. We will give two examples of lines of research, which, in our opinion, go in this direction. The first one is drawn from 18th century moral philosophy, this is the approach adopted by Adam Smith in *The Theory of Moral Sentiments*. The other one is drawn from a very recent domain of research, that of social neuroscience. We will see how, in both cases, the objective is not only to understand how moral judgements shape behavior but also to get an understanding of how people form these moral judgements. We will conclude with a discussion as to how these two approaches seem to be mutually illuminating and will suggest what contribution they can make to economic methodology. The epistemic cost that would be necessary to integrate morality in economics is certainly high but we think that rather than eliminate the identity of economics as a specific social science it may strengthen that identity.

1. Does a Positive Theory of Morality Already Exist?

1.1. The Theory of Social Preferences as an Answer to Experimental Evidence Against Homo Œconomicus

4 In the 1980s, experimental economics started to produce evidence that highlighted pro-social behavior[3] and were inconsistent with what would have been expected within the standard model. The experiments revealing these "anomalies" were initially based on simple games, in which two individuals interact with each other, the ultimatum game, the dictator game or the investment game. However, later, games involving groups of individuals, of varying sizes, such as the public goods game were also studied (see Camerer, 2003 for a review of this literature). According to the standard model, in the ultimatum game[4], the recipient should accept any proposal that leaves him a payment strictly higher than zero and the subject who makes the offer should therefore leave the smallest amount possible. However, experiments show that the recipient rejects, with a high frequency, any proposal less than 25% of the total amount and proposals close to 50/50 appear with a relatively high frequency. In the dictator game[5], if the frequency of selfish proposals increases, deviation from the behavior assumed in theory persists. Even when the protocol ensures perfect anonymity between subjects and between each subject and the persons organizing the experiment, the gap still persists (see the different protocols, for example tested by Güth, Schmittberger and Schwarze, 1982 and Hoffman et al., 1996).

5 These observations have been subject to different types of interpretation that led to a reconsideration of the standard model.[6] Here, we are interested in one of the most famous interpretations, which is that people have *social preferences* and the theoretical developments that followed. Fehr and Fischbacher (2005, 151) define social preferences as "other-regarding preferences in the sense that individuals who exhibit them behave as if they value the payoff of relevant reference agents positively or negatively." Indeed, Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) propose models that,

without abandoning the standard framework, can reproduce the observed behavior. The basic idea is simply to change the utility function by incorporating a component that represents the interest that the individual has in others' gains. In the model proposed by Fehr and Schmidt, the so-called *model of inequity aversion*, this component is simply the difference between the payment of the individual and that of the other or an average of those of others. The utility function is a weighted sum of the player's material payment and this component. Furthermore, an additional subtlety is introduced. The player is more sensitive to inequalities that are unfavorable to him than to inequalities which are in his favor. The player values equality of payments but he or she, is also more or less biased in his own interest. This model has been extremely well received in the economic literature.[7] Among the articles published in the *Quarterly Journal of Economics* since its inception, it was the fourth most cited item in 2015.[8]

6 This seminal paper and the ensuing literature on social preferences, constitute a turning point in the evolution of economics. Hitherto, the exclusion of interpersonal comparison of utilities was considered as a theoretical justification for the elimination of morality from economics. Indeed, as a consequence of this technical impossibility, economics could not provide any help to the policy maker in comparing individuals' welfare. However, if people do, in fact, compare their welfare with that of others, in other words if the comparison is internal to the agent's rationality and not external to it and determined by a policy maker, then, incorporating such comparisons into the analysis becomes feasible. Therefore, economics reformulated in this way enables the policy maker to take this comparison into account and the debate can be moved from the question of the possibility to the question of the means. However, from the outset, the theory of social preferences has been subjected to criticism. Some have said that the parameters allow so many degrees of freedom that the conclusion that the model can reproduce the experimental evidence is somewhat unconvincing (Binmore and Shaked, 2010). Indeed, in this approach, each individual is characterized by a parameter of his utility function that measures his degree of inequity aversion. Thus, an underlying assumption is that inequity aversion is an intrinsic and stable characteristic of each individual and that the different levels of inequity aversion in a population may generate a variety of different types of behavior.

7 Other models of social preferences have been proposed, in which the social component of the utility function can take different forms, as a measure of social welfare or the welfare of the poor. Experiments have been designed to try to distinguish and compare the relative importance of these different forms of social preferences (Charness and Rabin, 2002; Engelmann and Strobel, 2004). Although the results of these experiments have not always been favorable to inequity aversion, the latter form of social preferences has remained the most prominent. Other models, again, are a bit more sophisticated in their form. Modeling reciprocity, for example, has occupied a particularly important place in the literature. Reciprocity was proposed to interpret the results of an experimental literature showing that a subject's generosity seemed to depend on what he had observed in the past behavior of another subject[9] (see, for example, among the most recent references Ben-Ner et al., 2004; Servátka, 2010; Bowles and Gintis, 2011; Herne et al., 2013). Specifically, these studies showed that, in a dictator game, the dictator's generosity could be strongly correlated to the generosity from which he had previously benefited. However, the different models that have been proposed to allow homo oeconomicus to show some reciprocity could not be limited to introducing into the utility function of an individual a component that depends on his partner's behavior. Modeling requires, in this case, changing the equilibrium concept (see Falk and Fischbacher 2005). In the most sophisticated versions, first proposed by Matthew Rabin (1993), the individual is not only concerned with what he or she observes about the behavior of the other, but also with what he or she can infer from it about the intentions of the latter. The model, then, has to focus on beliefs. These models, however, are very complicated and difficult to use and solve. Yet, if modeling reciprocity is so complex, it is perhaps because this behavior is fundamentally different from simple inequity aversion. Reciprocity cannot easily be reduced to a component of the utility function which constitutes an intrinsic characteristic of the individual. Indeed, modeling reciprocity requires reflecting on individuals' interactions.

8 This multiplicity of possible models causes another problem. It does not allow economics to replace the standard model of the homo oeconomicus by a single new model integrating social preferences. For this reason Fehr and Fischbacher (2005) claim that reciprocity is a special case of social preference, but more sophisticated than inequity aversion: "Strong reciprocity means that individuals behave as if their positive or negative valuation of the reference agent's payoff depends on the actions of the reference agent." (Fehr and Fischbacher, 2005, 152) and their conclusion is that inequity aversion, although not a direct reflection of this more sophisticated behavior, is a good approximation while being much simpler and more tractable. Nevertheless, if we try to give a coherent account of this diverse literature, we may note that there is an increase in the depth of analysis, when one moves from the idea that subjects are concerned by the gains of others, to the idea that they are concerned with the behavior of others and finally to the idea that they are concerned with the intentions of others. Payments may result from behavior that may itself result from intentions. But what is the origin of these intentions?

9 In spite of these diverse levels of analysis, these models share a common approach. They interpret the pro-social behavior of the experimental stylized facts as resulting from an intrinsic moral rationality. Another literature which invokes different principles of justice, although related to these models, challenges some aspects of them.

1.2. Intrinsic Inequity Aversion versus Acquired Principles of Justice

10 Shortly after the first models were proposed to reproduce this experimental evidence on pro-social behavior, Cherry et al. (2002) published new experimental findings that challenged the experimental approach that has been used to generate this evidence. In their paper, they proposed adding to the traditional protocol, based on a dictator game, a first step during which the dictator has to perform a task to earn the amount he can share in the second stage. It appears that, in this case, the dictator becomes very selfish and his behavior now becomes consistent with what the theory predicts. However, Cherry et al.'s article is not an approval of the still dominant theory of the homo oeconomicus, but shows that we have perhaps been a little too hasty in adopting an explanation based on inequity aversion. Indeed, this work has shown that the experimental protocols that had been at the basis of a huge literature, stemmed from the false assumption that subjects did not attach any importance to the origin of the amount they have to share or invest. In those protocols, as in the utility function, money had no odour. Yet, this hypothesis should have been questioned at the same time as the hypothesis of perfectly selfish and egocentric behavior. Indeed, it appears that the way people take their decisions could depend on information which was not considered to be relevant before.

11 Cherry et al.'s paper opened the way for a new experimental literature. This literature is based on a standard protocol consisting of two steps. In a first step, unlike in Cherry et al.'s framework, all subjects are involved and can earn the money that can be shared in the second step. In this setting, it becomes clear that inequity aversion had been an appropriate interpretation, because giving equal rights to both subjects of an ultimatum game or dictator is the obvious "fair" solution in protocols without a production phase. Equal sharing is the basic principle of justice best suited to this context. A more general interpretation was then proposed, which argued that subjects refer to principles of justice (see in particular Cappelen et al., 2007; Rodriguez and Moreno, 2012). The way was opened for a literature on the role of principles of justice.

12 In the protocol proposed by Cappelen et al. (2007), the distribution phase, which is a dictator game, is preceded by a production phase during which subjects must answer a general knowledge quiz. At the end of this production phase, each subject is assigned an amount that depends on the number of his answers which are correct, but also on a price to be applied to each one of the latter. This price differs from one subject to the

other. Therefore, in each pair, the total amount that will be shared depends on the number of correct answers, but also on the two different prices. So that three focal points may appear, depending on three fundamental principles of justice: strict equal sharing of the total amount, libertarian sharing that gives each subject a share proportional to what he or she earns in the production phase—regardless of whether he or she is responsible or not for the price that applies to his correct answers—and finally the libertarian-egalitarian distribution, which consists of sharing in proportion only of the number of correct answers from each of the participants. These three fundamental principles emerged from a long debate on responsibility (see Konow, 2003). Cappelen et al. show that, indeed, the shares proposed by dictators can be categorized into groups which correspond closely to the three principles of justice. They propose a model that allows them to replicate adequately what they observed. The model incorporates in the utility function a component representing the valuation of a principle of justice, the agent's fairness ideal. They replace the difference between the agent's material payment and the average of others' payments that constitutes the inequity aversion of the model of Fehr and Schmidt (1999), by a difference between the player's material payment and a share corresponding to an ideal of justice. This ideal of justice characterizes the agent, in addition to the weights of the different components of the utility function. So, we passed from the theory of inequity aversion to the theory of ideals of justice.

13 Fundamentally, the models of Fehr and Schmidt (1999) and Cappelen et al. (2007), have much in common. First, they share *a common objective* which is to show that the homo oeconomicus is not purely selfish and values morality. Second, they share *the assumption that individual preferences are stable*, in agreement with mainstream economics. In order to reach their common objective under this common assumption, they propose to re-introduce morality in economics by incorporating, in the utility function, a social component that the economic agent values. In the first model, the economic agent is characterized by his degree of inequity aversion that is measured by the parameters constituting the weights in his utility function. In the second model, the individual is characterized both by the weights and his fairness ideal. However, Cappelen et al. (2007) are much more cautious than Fehr and Schmidt when they come to justify their model and the stability of social preferences. Of course, they insist on the fact that an individual has stable fairness ideals at least at one point in time; otherwise, we would not speak of fairness ideals. But they recognize that economic agents are not born with fairness ideals, and that the latter are the result of education and a long life experience.

14 More recently, in 2012, Rodriguez and Moreno have proposed an alternative conception of how principles of justice count in subjects' choices. Their experimental protocol is very similar to that of Cappelen et al. (2007) and like the latter, they show that the dictator is not purely selfish and seems to refer to one or the other of the three principles of justices. However, comparing a treatment in which the dictator is assigned a higher price than the recipient and another one in which the opposite prevails, they show that dictators' proposals in the first treatment are more frequently close to the libertarian principle, while dictators' proposals in the second treatment are more frequently close to the libertarian-egalitarian principle. The difference between the two treatments is significant. The statistical analysis appears to show that the dictator *chooses* the principle of justice that maximizes his payment. Following their interpretation, on the basis of what they observed by using the same protocol as Cappelen et al. (2007), subjects are not characterized by fairness ideals but adopt different principles of justice in different situations. More precisely, the dictator chooses the principle that maximizes his material payment. This would suggest that one can hardly consider that he is characterized by a fairness ideal as defined in the model of Cappelen et al., even though he may justify his behavior as corresponding to such a principle of justice. This result challenges the stability of social preferences and can be considered as a way back to the model of homo oeconomicus as a selfish utility maximizer. Notwithstanding the fact that Rodriguez and Moreno do not propose a model themselves, we might imagine that this new interpretation could open the way towards a modeling approach in which principles of justice would play the role of a social constraint.

15 Although these different approaches presented above seem to be mutually challenging, in fact, they can be thought of as belonging to a common framework. In what follows, we discuss the limits of this framework if the overall objective is to tackle the ambitious project of integrating morality in economics.

1.3. On the Possibility of Building an Essential Model of the Moral Agent

16 It seems a commonplace to say that standard economics falls into the framework of methodological individualism. However, that expression has often been criticized as not being well defined or as having different meanings, so that for example Geoffrey Hodgson (2007) proposed to abandon it. His objection is that "it is unclear whether it means that explanations should be in terms of individuals plus relations between individuals, or in terms of individuals alone." (Hodgson, 2007, 222) Nevertheless, because we think that it is of a particular interest for the purpose of this paper we will explain here what we mean by methodological individualism. Indeed, the first component of this paradigm is the idea that the individual agent is autonomous from social institutions, in the sense that he is provided with an independent intentionality. Thus, in economics, individual rationality is fundamental and must be at the basis of every explanation of social phenomena and collective behavior. An implication of this idea is that standard economics is concerned with what is so essential to the agent's rationality that it does not depend on his social and institutional environment. This will be the foundation on which the theory will be elaborated to understand individual *and* collective behavior. We will say that this model of the economic agent, described as a utility maximizer, and which does not depend on the environment, is *essential*. A second important component of methodological individualism is the idea that, starting from this essential model of the agent, it is necessary to understand how the latter is connected to his environment. This leads one to take into account the interactions between agents. The form of these interactions and the outcomes which result from them, may depend on the social environment in which institutions play an important role. Therefore, we have to be careful to make the distinction between the individual agent's behavior and the model of the individual agent's rationality that generates this behavior. The fact that economics has adhered to the paradigm of methodological individualism does not mean that the individual agent's behavior does not depend on his environment. Even in the standard model of general equilibrium, for example, the agent is a utility maximizer but under constraints given by the prices which are generated by his environment. In Fehr and Schmidt (1999), the stable utility function that incorporates inequity aversion has to be considered in the context of a game and it can generate very different and even opposite types of behavior, cooperative or competitive, depending on the type of game, ultimatum or market game, in which subjects are involved. Therefore, while Fehr and Schmidt suggested modeling stable preferences, the way these preferences are manifested in terms of behavior depends on the environment and the type of social interaction.

17 The role economics wants to play as a science of public decision depends on the existence of this essential model of the agent. Indeed, because economics wants to help society to choose its institutions, it needs to know what it is, in people's behavior, which depends on these institutions, and, above all, what does not. As we have already said, according to methodological individualism, a model of the individual agent's rationality is needed, on which the theory which describes the interactions between the agents and institutions is built. As a consequence, changes in institutions modify the interaction between agents but do not modify their rationality. The policy maker can act on the form of institutions, taking into account this rationality. It should be clear that this argument is an extension of the Lucas critique[10] from macro-economics to every domain in economics that contributes to make it a science of public decision. Therefore, the prevailing belief is that the role of economics is to make explicit what Lucas calls a structural model, and what we call here an essential model. In this

framework, the stability of preferences in general, and of social preferences in particular, which are constitutive elements of the essential model of the individual agent's rationality, is also fundamental.

18 The contribution of the theory of social preferences to economics is to argue that, although it is possible to keep the essential model of the economic agent, who maximizes his utility and has stable preferences, it is necessary to recognize that this model of standard economics is incomplete. Indeed, there is something over and above what can be directly or indirectly converted into the material gains or selfish welfare, that is essential in the utility function. The main idea is that people have social values, they value moral principles. The theory does not explain why this is so, and this question makes no sense if social preferences are innate, which is the underlying assumption of an essential model. However, as discussed before, there are different models of social preferences and they cannot all pretend to complete the essential model. For example, the approach of Cappelen et al. (2007) that consists in the characterization of a model of the moral agent based on the utility function enters clearly in the framework of methodological individualism. However, Cappelen et al. would not pretend that principles of justice are innate. They just consider that their model is valid for the 'short run'.

19 On the contrary, we think that Fehr and Schmidt (1999) had the ambition to propose a new essential model of the economic agent, as witnesses the very name of their model, the model of inequity *aversion,* and this could explain its remarkable success. Part of Fehr's subsequent research has been devoted to find some evidence of the essentiality of inequity aversion and social preferences. So, he participated in an ambitious research program whose objective was to test the explanatory power of the theory of social preferences in different social environments, different institutions and different cultures. He first joined a research network led by Herbert Gintis and Rob Boyd, focused on the "nature and origin of preferences" (Henrich et al., 2004, 2).[11] Later, one member reported results from a field experiment based on the ultimatum game among the Machiguenga, an ethnicity living in the south-eastern Peruvian Amazon basin. The behavior of Machiguenga subjects was very different from the stylized facts previously obtained in economic laboratories and much less pro-social. What was then called "the Machiguenga outlier" (Henrich et al., 2004, 11) became famous among behavioral economists. Stimulated and intrigued by these results, the group decided to launch a program of cross-cultural experimental work and invited a number of anthropologists to participate. The conclusion of the program is far from reinforcing the idea of an essential model of social preferences. It highlights the influence of culture on *behavior* but raises more questions than answers.

> It is tempting to react to the widespread experimental evidence of non-selfish behaviors by replacing the selfish axiom with some equally and universal assumption about human behavior. If Homo oeconomicus has failed the experimental test, maybe Homo altruistic, Homo reciprocans, or some other simplified version of a panhuman nature will do better. The diversity of behaviors we have observed leads us to doubt the wisdom of this approach. (Henrich et al., 2004, 50)

20 Our opinion is that there is a profound contradiction between methodological individualism and morality. The theory of social preferences tries to explain how morality shapes agents' behavior without explaining how the agents acquire their morality. The models proposed simply try to represent the idea that people value morality. Those that are best suited to this approach, try to capture something in human nature, which could be thought of as an intrinsic characteristic of the individual, his essence, and which could generate pro-social and moral behavior. But morality, or more precisely, moral rationality, is a social phenomenon. Our claim is that the basic unit that has to be studied in order to understand morality, the indivisible unit, may not be the rationality of an individual but rather the interaction between several individuals' rationality. If this is true, the research program which tries to explain moral behavior in the framework of methodological individualism is as hopeless as trying to explain sexual reproduction by analyzing a unique representative

human being. On the contrary, we think that it is worth seeking an alternative approach which takes interactions seriously and whose ultimate objective would be to explain morality as a *social mechanism*, which could explain *both* how the agents' moral rationality shapes their behavior and how the interaction between agents structures this moral rationality. Can we conceive of such mechanisms? Is this approach a feasible alternative? In what follows, we give two examples of lines of research, which, in our opinion, go in this direction. The first one is drawn from 18th century moral philosophy and the other from a very recent domain of research.

2. The Smithian Mechanism at the Origin of Moral Sentiments

2.1. On the Variability of Moral Sentiments

21 The *Theory of Moral Sentiments* by Adam Smith starts with this sentence:

> How selfish man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it. (Smith, 2005, 4)

22 Later, he explains how the individual's beneficence towards others decreases with some sort of affective distance, the individual devoting more attention to his family's members than to unknown people, to the members of his own society than to those living in foreign societies; but he also explains how the individual is capable of benevolence towards the whole Universe. While he seems, in general, to be focused on direct relations, in which the protagonists can listen to and observe each other, he also refers to very abstract relations and explains, for example, how somebody in Europe can sympathize "with the myriads of inhabitants of the great empire of China who would be suddenly swallowed up by an earthquake." (Smith, 2005, 119-120) Elsewhere again, Smith explains how moral sentiments depend on the position of people in the society and how, for example, these sentiments are corrupted by a "disposition to admire the rich and the great, and to despite or neglect persons of poor and mean condition" (Smith, 2005, 53).

23 Therefore, if Smith is interested in the *nature* of man, his conclusion is that moral sentiments depend on his social environment. In fact, this variation of the individual's moral sentiments and beneficence towards others is a consequence of the way those attitudes are formed. Indeed, Adam Smith's *Theory of Moral Sentiments* is devoted to the question of *how people form moral sentiments*. The Smithian individual agent forms his moral sentiments through an informal learning process that is realized through the vector of *social interactions*. By rephrasing the question in contemporary terms, we would say that in the five first parts of the *Theory of Moral Sentiments*, Smith's objective is to understand the psychological mechanisms that shape individual thinking and behavior, when he is involved in social interactions. The richness and complexity of Smith's though may well explain the existence of a prolific literature discussing these mechanisms (from Morrow, 1923 and Anspach, 1972, to Sugden, 2002; Dellemotte, 2005; Raphael, 2007; Forman-Barzilai, 2010; or Bréban, 2017). We will focus here on the main components of Smith's theory that are particularly relevant for the purpose of this paper.

2.2. The Sympathy Operator as Part of an Essential Model of the Individual?

24 According to Smith's conception of the individual, the latter has passions, among which hunger, pain, anger or sexual desire. These passions can be interpreted as emotional reactions to the environment, that may trigger actions. But the individual is also able to feel the passions of others, by means of what Smith calls sympathy, that of the mother for her child, that of the individual who feels the suffering of another, that of somebody who shares the happiness of a friend. Therefore, we can consider that sympathy falls, in one sense, into the same category as other passions to which the individual is subject, because it is an emotional mechanism and can be interpreted as an instinct. But sympathy, as an operator, occupies a special place among these emotional mechanisms because it is, in one sense, indirect. It does not generate emotions as a direct reaction to a situation, but permits the transmission of these direct emotions from the individual who experiences the situation to another individual.

25 However, the operator of sympathy also includes a cognitive dimension. Indeed, the way emotions are transmitted is through the individual's imagination. The individual can imagine the situation that triggers the other's feeling and feels, in a certain degree, the same emotion, by imagining himself in the same situation.[12] However, this imagined situation may not coincide with the perception of the situation by the person who is the object of sympathy. In fact, there is, in general, a gap between, on the one hand, the original feeling of the person directly concerned and that is triggered by a situation and on the other hand, the feeling of the spectator transmitted through the sympathy operator. There is in general a loss of intensity of emotions, whether the latter are "negative" such as pain or sadness or "positive" such as happiness, during transmission through the sympathy operator. However, the spectator may also overestimate the feelings of others. Taking the example of the mother, Smith says:

> What are the pangs of a mother, when she hears the moanings of her infant that during the agony of disease cannot express what it feels? In her idea of what it suffers, she joins, to its real helplessness, her own consciousness of that helplessness, and her own terrors for the unknown consequences of its disorder; and out of all these, forms, for her own sorrow, the most complete image of misery and distress. (Smith, 2005, 7)

26 A fundamental point in Smith's theory, explained in chapter two of the first part entitled "On the pleasure of mutual sympathy", is that, whether the transmitted emotions are positive such as happiness for example or negative such as pain, sympathy generates a certain pleasure for both individuals:

> As the person who is principally interested in any event is pleased with our sympathy, and hurt by the want of it, so we, seem to be pleased when we are able to sympathize with him, and to be hurt when we are unable to do so. (Smith, 2005, 11)

27 This idea that sympathy generates pleasure is even more interesting when we realize that it would be very difficult to interpret it in the framework of modern economics, by changing the components of the utility function or by taking into account indirect effects. In fact, according to Smith, this feeling cannot be derived from any sort of strategic reasoning. "... both the pleasure and the pain are always felt so instantaneously, and often upon such frivolous occasions, that it seems evident that neither of them can be derived from any such self-interested consideration." (Smith, 2005, 8)

28 It is precisely this convergence of emotions that generates pleasure. The pleasure increases with the narrowing of the unavoidable gap between the original feeling and the transmitted one, which results from this imperfect transfer. This pleasure generated by the coincidence of emotions is such that both parties, the object of sympathy who is affected directly and the spectator, subject of sympathy who imagines the situation, try to make their feelings closer to each other. Each member of the pair makes an effort to modulate his emotion in the direction of the other's. This mechanism is the foundation of virtues in Smith's moral philosophy. A virtuous person, spectator or object of sympathy, is a person who has the highest capacity for this modulation.

> Upon these two different efforts, upon that of the spectator to enter into the sentiments of the person principally concerned, and upon that of the person principally concerned, to bring down his emotions to what the spectator can go along with, are founded two different sets of virtues. (Smith, 2005, 18)

29 However, convergence of emotions does not always occur. This is the case when the spectator is guided by his own passions or when he disapproves of the other person's behavior. But Smith conflates moral approbation and the sharing of others' emotions when he says that

> [t]o approve or disapprove, therefore, of the opinions of others is acknowledged, by everybody, to mean no more than to observe their agreement or disagreement with our own. But this is equally the case with regard to our approbation or disapprobation of the sentiments or passions of others. (Smith, 2005, 12)

30 In fact, the sympathy operator, precisely described in the beginning of the *Theory of Moral Sentiments*, and which works on a bilateral relationship, is also the basis for a more general and complex mechanism, which describes how the individual forms judgements in society and whose outcomes are the moral sentiments of that individual. In this general mechanism, there are now three points of view: the point of view of the person who acts, the agent, the point of view of the person who is concerned by the agent's action, and the point of view of a third person who might observe the two others, *the impartial spectator*. Here again, the way the impartial spectator forms a judgement on the agent's behavior is based on the sympathy operator, which works towards the two others. Indeed, the impartial spectator's judgement of the agent's behavior depends on the distance between what the former imagines to be the sentiments of the latter when he acts and his own sentiment in the same situation. However, the role of the impartial spectator is particularly important in another type of situation, that of introspection, in which the individual assesses the morality of his own behavior. Indeed, as Raphael (2007, 42) says, "an agent can judge his own character and conduct only if he imagines himself in the position of a spectator". In that case, the agent, by changing his perspective, imagines himself in the role of an impartial spectator. Then, the operator of sympathy works between the individual agent and the object of his action, but it also works in the case where the individual takes the perspective of the impartial spectator and judges his own behavior. Therefore, the individual has the capacity to take different perspectives in a given situation. In this play of mirrors, he is, alternately, the agent and the subject of sympathy, the person who is concerned by the action and object of sympathy, and the impartial spectator. It should be said however, that the word "alternately" is a bit misleading since, in a given situation, the individual can adopt all the different perspectives. However, in all these situations, the aim of these changes of perspective and these adjustments of emotions is the coincidence with the others' sentiments and the seeking for approbation. Therefore, if the impartial spectator is at the origin of moral behavior, this is because the individual sees his own behavior from the point of view of others and can modulate his passions because he wants to gain their approval. However, without entering into the debate about the nature of the impartial spectator (see for example Raphael, 2007; and Keppler, 2010) it is worth saying that he "is a higher authority than actual spectators in moral judgment." (Young, 2013, 2)

2.3. Smith's Theory and Contemporary Economics in Perspective

31 If the idea that people attach some importance to their image in others' eyes has frequently been developed in economics, it would be wrong to interpret the impartial spectator in this sense. In contemporary economics, and in particular in the literature on repeated games and signaling (see, for example, Kreps and Wilson, 1982; Aumann and Hart, 2002), people give importance to what others think about them because they believe that they can draw material gains from it, thanks to a good reputation or

because they fear some punishment. In Smith's view, the main benefit that people draw from the account of their fellows is only a feeling of harmony. "These affections, that harmony, this commerce, are felt ... to be more important to happiness than all the little services which could be expected to flow from them." (Smith, 2005, 34) It is difficult to see how, by simply using the utility function, we could model this "fellow feeling [that] does not fit into the ontological framework or rational choice theory" (Sugden, 2002, 63).

32 Furthermore, Smith is at pains to disconnect his fundamental explanation of moral sentiments from utility. In the fourth Part of the *Theory of Moral Sentiments*, after a long advocacy for "the beauty which the appearance of utility bestows upon the characters and actions of men," he "affirm[s], that it is not the view of the utility or hurtfulness which is either the first or principal source of our approbation and disapprobation. These sentiments ... are originally and essentially different from this perception." (Smith, 2005, 167-169)

33 In fact, Smith's *Theory of Moral Sentiments* differs from current attempts to introduce morality into economics in several fundamental aspects. Before going into the details of these aspects, we note that, as a consequence, and contrary to the essential model of the individual in the theory of social preferences, in Smith's theory, the individual's moral sentiments are not an element of an essential model but the outcome of a complex social mechanism. But why should we choose to speak about a social mechanism instead of an essential model of the individual's moral rationality?

34 There is an essential model in the *Theory of Moral Sentiments* based on the sympathy operator, which is an innate capacity, as passions are. However, should we really consider the Smithian system as an essential model of the individual when each of its components changes with the social environment? On the one hand, physiological mechanisms that generate emotions are certainly a modern conception of what could be an intrinsic model of passions in Smith's theory. But even the most basic passions such as hunger or pain can be modulated by the eye of the spectator, just to attract his sympathy. On the other hand, the sympathy operator could be considered as an essential model of the individual because it relies on the idea that one has an innate ability to put oneself in another's shoes. But this ability itself depends a lot on the social environment. What differs fundamentality from the theory of social preferences is that this essential model is not a model of maximization of individual utility or of anything else, it is a model of *convergence*. Indeed, the key point of the model is that the individual wants his sentiments to be as close as possible to those of the others. The convergence of emotions and sentiments provides mutual pleasure, independent of the particular emotions and sentiments, whether positive or negative. In fact, this convergence seems to be the fundamental objective of the agent. Therefore, if this is so, can we really consider that the agent has an independent intentionality as assumed in the paradigm of methodological individualism? Our opinion is that the Theory of Moral Sentiments describes a complex social mechanism that has to be understood as a model of entanglement, a model of the individual involved in society.

35 It is interesting to remember what Morrow (1923) said a long time ago, to qualify the singularity of Hume, and after him Smith, in the context of the individualism of the eighteenth century:

> To abandon the assumption that human nature—with its motives to activity, its forms of approbation, its selfish or benevolent or neutral impulses—is 'given'; to point the way toward an explanation of human nature through the association of individuals with one another, is to give up the rigid individualism for a more comprehensive point of view which recognizes both the individual and the social factors in experience. (Morrow, 1923, 62)

3. Morality and Empathy in Contemporary Behavioral Sciences

3.1. Empathy as a Complex Phenomenon

36 Jean Decety, who is a researcher at the Social Cognitive Neuroscience Laboratory at the University of Chicago and one of the founders of a new field called social neuroscience, said in an interview that "the true origin of the concept [of empathy], underpinned by a naturalistic vision of psychological phenomena, derives from what the Scottish philosophy called 'sympathy'." (Decety, 2009) This concept of empathy that, indeed, shares common characteristics with Smith's sympathy, occupies an important place in the different behavioral sciences. According to the ethologist Frans de Waal (2010), empathy is the basis of the ability to cooperate in humans, as in all mammals. He gives a rather general and encompassing definition of empathy, which is the ability of the individual to identify himself with those who are in need and pain; this identification provoking an emotion that induces the individual to try to help. In neuroscience, research on empathy has accelerated since the second half of the 1990s, following the controversy provoked by the discovery by Giacomo Rizzolatti's team (Rizzolatti et al., 1996) of *mirror neurons*, also called *empathetic neurons*. These neurons of the brain have the peculiarity of being activated both when an individual performs an action and when he observes another individual performing the same action. This discovery seemed decisive because it reinforces a naturalistic approach of empathy.[13] What if empathy could be described by the operation of a category of neurons? But the function of mirror neurons, whose existence in humans has only recently been proved (Mukamel et al., 2010) and which remains controversial, is a fledgling area of research; and empathy remains a multifaceted and complex mechanism.

37 Making this complexity intelligible, provides a first challenge for behavioral sciences, social psychology and neuroscience, which is to disentangle different components of empathy. A clear consensual categorization has still to be established, except for the agreement on the distinction between an emotional and a cognitive dimension. The *emotional component* would be the ability of an individual to become affectively aroused by the emotions of another while the *cognitive component or theory of mind, or affective perspective-taking,* refers to the ability to understand the other's state of mind by putting oneself in the other's shoes and representing his intentions, beliefs, and emotions. Decetey and Cowell (2014) also distinguish a third *motivational* component, which refers to the willingness of an individual to take care of another and that is often associated with the emotional component. According to these two researchers, these different components of empathy correspond to different mechanisms that may interact or otherwise operate in parallel. Given that each of these different components can itself be decomposed, one can easily understand the complexity of the phenomenon that researchers are trying to analyze and the extensive literature that emerged as a consequence. Within this vast literature, however, we can distinguish two approaches that are of particular interest for the purpose of this paper. The first one focuses on the measurement of the different forms of empathy, while the other tries to understand empathy as a process.

3.2. Empathy as a Measurable Characteristic

38 The first approach proposes different indexes based on psychometric tests, such as the Interpersonal Reactivity Index (IRI) proposed by the psychologist Davis (1983) or the Toronto Empathy Questionnaire (TEQ) proposed by Spreng et al. (2009). The idea is to describe a series of interactive situations with which the subjects could be confronted and to ask them to imagine what they would do, feel or think in each of these situations and to choose the corresponding answer in a list. Scores are then calculated, which are considered to measure the subject's degree of empathy in its various forms. Some tests propose to measure the different forms of empathy by defining sub-scales, such as the IRI that distinguishes *Perspective Taking*—that is the tendency to spontaneously adopt the psychological point of view of others—from

Fantasy—that taps into respondents' tendencies to transpose themselves imaginatively into the feelings and actions of fictitious characters in books, movies, and plays—*Empathic Concern*—which assesses 'other-oriented' feelings of sympathy and concern for unfortunate others—and *Personal Distress*—which measures 'self-oriented' feelings of personal anxiety and unease in tense interpersonal settings. Others, such as TEQ, that conceptualizes empathy as a primarily emotional process, propose to define a unique score, which would measure a common denominator of the different components. In all of these cases, it appears that substantial individual differences can be observed. The neuroscientist Tania Singer and the economist Ernst Fehr (2005) have together emphasized the fact that there is the same heterogeneity in terms of empathy measured by these tests, as is assumed in the model of social preferences. In their article they promote the new field of neuroeconomics, the development of a program whose objective would be to find, in the analysis of empathy, a contribution "to the micro-foundation for theories of social preferences" (Singer and Fehr, 2005, 343). One might wonder what exactly is meant by this. In fact, it seems that they see in the two main components of empathy, theory of mind and emotional empathy, a justification of the essential model of the homo oeconomicus with social preferences. On the one hand, the rationality assumed in game theory, which considers that people are capable of predicting others' actions, could find its justification in the theory of mind. On the other hand, social preferences modeled by a utility function with a social component, could find their justification in emotional empathy as a motivation. In this framework, the heterogeneity of behavior could have two origins; in the heterogeneity of rationality and theory of mind and in the heterogeneity of inequity aversion and of the degree of emotional empathy. They propose to assess the relative importance of theory of mind and empathy for the prediction of motives and actions of others in different situations and for altruistic behavior. In order to do so, they suggest two testable predictions.

1. People with stronger empathic abilities are better predictors of others' motives and actions.

2. People who exhibit more affective concern are more likely to display altruistic behaviors.

39 Therefore, it appears that, in spite of the conclusions of the research program on small scale societies which we presented in the first section, the objective remains to find the essence of the individual agent that could explain the individual's behavior in society. Empathy with its different components could play a crucial role, and presents the advantage that it can be studied by neuroscientists. One might fear that the undeclared objective of this program is to show that the empathy mechanisms, that psychologists have learned to measure using tests and for which neuroscience researchers are seeking material foundations in the brain, may become a physiological justification of the theory of social preferences, at the risk of a somewhat extreme naturalism.

40 Artinger, Exadaktylos, Koppel and Saaksvuori (2014) have responded to Singer and Fehr's proposal to launch a research program in neuroeconomics and tested the relationship between the different components of empathy and a type of observed behavior assumed to be generated by social preferences. They designed a protocol based on a dictator game and an ultimatum game and attempted to correlate the generosity of proposers with the results of various questionnaires assessing the emotional component of empathy and what has been referred to as the *theory of mind*.[14] Given that the theory of mind is the capacity to understand the intentions, beliefs and reasoning of others, they also used as a simple measure, the capacity of subjects to guess somebody else's choice. Indeed, each subject had to make a choice sequentially as a dictator, as a proposer in an ultimatum game and as a recipient. She also had, at each stage, to guess the choice of another subject drawn at random and the accuracy of this guess was also used as a measure of the theory of mind. Their results did not confirm Singer and Fehr's second prediction that people who exhibit more affective concern are more likely to display altruistic behaviors.[15] Indeed, they found no significant

relationship between the empathy measures and offers in the dictator game and the ultimatum game. Measures of theory of mind did weakly correlate with offers in the dictator game. Finally, this is the accuracy of the guess that appeared to be the best predictor of proposers' generosity. These results, they are also contradict many results from previous experiments conducted by psychologists that seem to confirm the hypothesis of a correlation between altruism and empathy designated as emotional, something which is referred to as the empathy-altruism hypothesis and which was formulated by Batson (2011).

41 Artinger et al. (2014) admitted, however, that these results may be largely explained by the neutral framework they used, in accordance with the rules of experimental economics. Indeed, the literature in psychology shows that the ability to experience emotional empathy is triggered by images or representations, or, in any case, by the possibility of identifying the object of empathy (see, e.g Blanchette and Campbell, 2012). Economists Small, Loewenstein and Slovic (2007) also showed how effective this mode of empathy is. They used an experimental protocol in which photos of individuals in need were shown to subjects before they were asked for a donation and observed that the latter are much more generous to the victims whom they have identified. Although Artinger et al's. laboratory experiments respecting the strict rules of experimental economics seem rather artificial, this was perfectly justified by their objective which was to understand the more realistic relationships people often experience today. Indeed in our large societies bathed in high technologies, people have often to take decisions that have consequences for others, in a very abstract environment. In a recent paper, Jean Decety and Keith Yoder (2016) found a type of result comparable to Artinger et al's. but in a more realistic environment, by using an online survey. In their experiment, they are interested in the sense of injustice and show that it is not the emotional but the cognitive component of empathy that predicts the importance of this feeling in individuals.

42 This literature that uses the measurement of the different components of empathy as a key element of its methodology presents the same drawback that we criticized in the theory of social preferences. By focusing on empathy as a characteristic of the individual, it does not pay enough attention to the fact that empathy is a mechanism. Questionnaires are used in order to assess the variability of behavior in different social situations. Therefore, the assumption is that a given degree of empathy generates different behavior depending on the social environment, in the same way that a given degree of inequity aversion can generate cooperative or competitive behavior depending on the game to be played. However, we do not know to what extent the degree of empathy itself is innate and to what extent it depends on the social and institutional environment. The answer to this question is a matter of considerable importance for the design of institutions. Should we simply assume that institutions just take into account this heterogeneity in individuals or should we also take into account the fact that institutions themselves can have an impact on this heterogeneity?

3.3. Empathy as a Process

43 Jean Decety proposes a different approach. He argues that the study of the brain must be accompanied by an overall vision of society and accuses conventional neuroscience of tending to focus its analysis on the brain, treating it as an isolated object, whereas the brain's main function is to interact with its environment, in particular its social environment. Decety tries, rather, to change the viewpoint and to develop hypotheses about the social mechanisms that could shape the individual's empathetic ability. This is one of the purposes of social neuroscience. Scholars like him who adopt this approach, advocate investment in the field of developmental social neuroscience. Indeed, they propose to conceive of empathy as a learning process starting during childhood and continuing throughout life (Cowell and Decety, 2015). It has already been established that manifestations of emotional empathy appear in the very early stages of a baby's life. According to Davinov et al. (2013), children's capacities to respond emotionally to the joys and sorrows of others and to express

empathic concern are present during the first year of life. However, children show strong bias toward individuals and members of groups with which they identify. Although this bias persists when the child grows up, the reference group changes and widens. Also, the child learns to respects concerns for morality, and the way this happens does not seem to depend directly on this emotional empathy. Between the ages of 3 and 9, the child feels intrinsic interpersonal obligations towards in-group members, while towards out-group members her obligations are only contingent on the existence of rules of morality he has to obey. The ability to experience emotional empathy remains in the adult, throughout his life. However, the adult also has the capacity to put himself in another's shoes by using a mechanism based on cognition.

44 The conditions under which this mechanism based on cognition works are still a subject for debate (Jackson, 1999; Gordon, 2009). In philosophy and psychology, in a more general framework, proponents of the *simulation theory* oppose the *theory theory*. According to the latter theory, individuals use a basic or 'naïve' theory of psychology, 'folk psychology', to infer the mental states of others, such as their beliefs, desires or emotions. This is the result of a learning process starting in childhood, the child observing his environment, and in doing so, gathers data about the world's true structure. As more data accumulates, he can revise his naive theories accordingly. Therefore, the *theory theory* posits that people are capable of mindreading because they have acquired an internal store of causal laws or principles corresponding to rules of logic and rational argument. The theory of mental simulation claims that this is not specific to mindreading since we use these same rules for our own reasoning (Gordon, 1986) and gives an alternative interpretation. According to this theory, as presented for example by the psychologists Davies and Stone (1995), the individual first, makes a projection and then, modifies it, establishing a difference between himself and the other. This correction is done through a series of adjustment criteria drawn from the consideration of the social, historical and cultural environment of the target person, and depends on the age of the simulator. One way to assess the respective validity of the two theories might be found in neuroscience. Indeed, the simulation theory would be consolidated if one could demonstrate the existence of a *double duty neural system* that is activated when a person reacts to a given situation and when she imagines somebody else reacting in the same situation. This brings us back to mirror neurons. However, we know that this is not true, at least for emotional empathy. Indeed, using an experiment in which a person in an fMRI observes her and her lover's hand while they are alternately pricked by a needle, Singer et al. (2004) deduce that some parts, but not the entire "pain matrix" is activated when empathizing with the pain of the other. In fact, in spite of this apparent contradiction, many scholars argue for a hybrid simulation-and-theory account and the debate is, then, about how the two theories intermesh.

45 As in the simulation theory, Decety considers that people operate a type of correction, but his conception of cognitive empathy is rather different from Davies and Stone's mindreading. Indeed, the correction imagined by Davies and Stone, as briefly presented above, does not involve any moral consideration. Contrary to this vision, Decety strongly emphasizes the role of morality. According to him, morality intervenes through cognition by modulating the emotional process of empathy. In recent articles co-written with Cowell (Decety and Cowell, 2014; Cowell and Decety, 2015), the authors hypothesize that morality is an attempt, by society, to extend the emotional component of empathy, which is usually thought of as being triggered through relationships with close relatives, to the relationships between all its members. Ultimately, Humanism is an attempt to extend this feeling to all humanity. But in that case, through which vectors can morality act on the individual? Decety considers that literature and art could be some of these vectors. However, what are the mechanisms by which social institutions might shape people's empathy? A first answer provided by neuroscience, using different methodologies, from functional neuroimaging, and lesion studies, to studies involving psychopaths, is that there is some evidence that both empathetic concern and moral reasoning require the involvement of the same region in the brain, the ventromedial prefrontal cortex (vmPFC). In other words, the vmPFC may *bridge* conceptual and emotional processes. This more detailed account of the

functioning of the brain provides some support for a vision in which the individual, connected to the society through her cognition, may modulate her emotional processes.

4. Conclusion

46 The theory of social preferences is an attempt to complete the basic model of the individual based on utility maximization, in order to introduce justice and morality into economics, without challenging that model. In this paper, our aim has been to demonstrate the limits of this approach. Certainly, the models of a moral agent proposed in this theoretical framework generate behavior that is consistent with some of the stylized facts observed in laboratory experiments. Moreover, one particular modification, the model of inequity aversion, can even explain how economic agents change their behavior depending on their institutional environment. However, in this model as in the other models in the theory of social preferences that has been proposed to date, moral principles are integrated in the utility function to modulate selfish motives in the determination of behavior, without explaining how and why. As a consequence, the choice of the moral principle that has to be integrated would seem to be arbitrary and there is no guarantee that this principle is independent of the social and institutional environment. The only way to ensure that a model of a moral agent is essential is to ensure that the model can explain both how moral principles influence behavior and how they are incorporated by the agent. In other words, we have to answer the question as to how moral judgements are formed.

47 This is exactly the object of Adam Smith's theory, which is entirely devoted to an explanation of how agents form moral sentiments. The operator of sympathy that allows the individual to share her fellows' feelings is at the basis of Smith's analysis. If there is an essential model of the individual in Smith's analysis, it is the model of an individual made of selfish passions such as hunger, thirst, sexual desire and other instincts that allow him to survive; but also an individual who would not be human without this operator of sympathy that allows him to feel also, indirectly, his fellows' passions. Finally, this individual would not be entirely human, without the complex mechanism of the impartial spectator which allows the individual to form moral judgements. In sum then, Smith's model is not an essential model of the isolated individual but an essential model of a moral agent embedded in society.

48 In contemporary behavioral sciences emotions are transferred between individuals through a complex phenomenon called empathy. In order to understand this phenomenon, scholars have tried to decompose it and rebuild it. On the one hand, they have tried to disentangle the different components of empathy, and came to distinguish, in the coarsest categorization, emotional from cognitive empathy. On the other hand, they also needed to mesh these different components, and to understand how they might explain pro-social behavior. This question is still the subject of debate. While for some scholars the origin of pro-social behavior is to be found in emotional empathy, for others this role is attributed to cognitive empathy. The difficulty with this debate may lie, precisely in the fact that emotional and cognitive empathy are just two components of the same mechanism. Among contemporary behavioral sciences the new field of social neuroscience holds a special place, since it is particularly interested in understanding empathy as a whole as a complex social mechanism. Smith's thought and recent developments in social neuroscience seem to be mutually illuminating.

49 According to social neuroscience, emotional empathy appears very early in a baby's life. As such it seems almost innate and should be likened to Smith's sympathy operator. But there is an important difference. In contemporary terms, we say that empathy is a sharing of emotions, even if imperfect, while in our interpretation the sympathy operator aims at the convergence of emotions. This means in particular that, in the first conception of sharing emotions, what the individual feels, at least partly, is only the emotion of the other. In Smith, in addition to this shared emotion, the moral agent feels something else, which is the pleasure of mutual sympathy. This 'fellow-feeling' already emphasized by Sugden (2002) is at the origin of the existence of moral

sentiments and seems to be, in Smith's view, the ultimate objective pursued by the moral agent. This leads us to two reflections. First, it would be interesting for neuroscience to try to single out this particular feeling. Second, the specificity of economics may be found somewhere else other than in methodological individualism. It might be better to consider economics as the discipline which views the individual as pursuing objectives, even if these objectives are not necessarily autonomous from the society. In the *Theory of Moral Sentiments* and contrary to contemporary economics, this objective is not the maximization of utility, although reflecting social preferences, but an objective of harmony and convergence towards others. The sympathy operator and the impartial spectator are only means towards this objective.

50 When we compare social neuroscience and Smith's theory, the other interesting aspect is that social neuroscience attempts to go further in the understanding of the empathy mechanism, from how people form moral sentiments to how they learn to form them. Indeed, scholars of social neuroscience have as a goal to better understand how emotional and cognitive empathy intermesh by studying empathy as an integral part of the development of the individual, a learning process starting in childhood and continuing all along life. To our knowledge, there is no conception of this kind in the literature on Smith's theory on the nature of the impartial spectator; even in Part VI of the last edition of the *Theory of Moral Sentiments*. The latter is an addition to the previous editions and may reflect the most mature part of Smith's thought. According to Morrow (1923), this part shows that "[Smith's] purpose is to set forth the stages by which the moral consciousness develops and the individual passes beyond himself and his individual concerns. ... The social consciousness thus begun in the family group grows as his sympathies spread out in widening circles, first to his clan or neighborhood, then to his nation, and finally to the whole system of the universe." (Morrow, 1923, 74) "The book [by Forman-Barzilai's] is the first full scale analysis of the role of geographic distance in Smith's moral philosophy" (Young, 2013, 6) and gives a very interesting analysis of these circles of sympathy. Our interpretation is that the objective of social neuroscience is to better understand this geographical dimension of the empathy mechanism by also taking into account its temporal dimension; Our hope is that a conception of the empathy mechanism as a learning process may help to reconcile the two dimensions.

The author would like to thank two anonymous referees and the editors for their helpful comments that allowed her to considerably improve a previous version of the paper, making clear the essential message.

Bibliographie

Anspach, Ralph. 1972. The Implications of the Theory of Moral Sentiments for Adam Smith's Economic Thought. *History of Political Economy*, *4*(1): 176-206.
DOI : 10.1215/00182702-4-1-176

Artinger, Florian, Filippos Exadaktylos, Hannes Koppel and Lauri Saaksvuori. 2014. In Others' Shoes: Do Individual Differences in Empathy and Theory of Mind Shape Social Preferences? *PLoS ONE, 9*(4): e92844. doi:10.1371/journal.pone.0092844.

Aumann, R.J. and Sergiu Hart (eds). 2002. *Handbook of Game Theory with Economic Applications*. Oxford: Elsevier, edition 1, volume 3, number 3.

Batson, C. Daniel. 2011. *Altruism in Humans*. New York: Oxford University Press.

Ben-Ner, Avner, Louis Putterman, Fanmin Kong and Dan Magan. 2004. Reciprocity in a Two-Part Dictator Game. *Journal of Economic Behavior & Organization, 53*(3): 333-352.

Binmore, Ken and Avner Shaked. 2010. Experimental Economics: Where Next? *Journal of Economic Behavior & Organization, 73*(1): 87-100.

Blanchette, Isabelle and Michelle Campbell. 2012. Reasoning about Highly Emotional Topics: Syllogistic Reasoning in a Group of War Veterans. *Journal of Cognitive Psychology*, *2*(24): 157-164.

Bolton, Gary E., and Axel Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review, 90*(1): 166-193.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Bréban, Laurie. 2017. An Investigation into the Smithian System of Sympathy: from Cognition to Emotion. *The Adam Smith Review, 10,* forthcoming.

Camerer, Colin. 2003. *Behavioral Game Theory*. Princeton: Princeton University Press.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen and Bertil Tungodden. 2007. The Pluralism of Fairness Ideals: An Experimental Approach. *American Economic Review, 97*(3): 818-827.

Charness, Gary and Matthew Rabin. 2002. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics, 117*(3): 817-869.

Cherry, Todd L., Peter Frykblom and Jason F. Shogren. 2002. Hardnose the Dictator. *American Economic Review, 92*(4): 1218-1221.

Cowell, Jason and Jean Decety. 2015. Precursors to Morality in Development as a Complex Interplay Between Neural, Socio-Environmental, and Behavioral Facets. *Proceedings of the National Academy of Sciences USA, 112*(41): 12657-12662.

Davis, Mark H. 1983. Measuring Individual Differences in Empathy: Evidence for a Multi-Dimensional Approach. *Journal of Personality and Social Psychology, 44*(1): 113-126.

Davies, Martin and Tony Stone.1995. *Mental Simulation: Evaluations and Applications - Reading in Mind and Language*. Oxford: Blackwell.

Davidov, M., C. Zahn-Waxler, R. Roth-Hanania, and A. Knafo. 2013. Concern for Others in the First Year of Life: Theory, Evidence, and Avenues for Research. *Child Development Perspectives, 7*(2): 126-131.

Decety, Jean. 2009. L'acquisition de l'empathie. *Pour la Science,* 63.

Decety, Jean and Jason M. Cowell. 2014. The Complex Relation Between Morality and Empathy. *Trends in Cognitive Sciences, 18*(7): 337-339.

Decety, Jean and Keith J. Yoder. 2016. Empathy and Motivation for Justice: Cognitive Empathy and Concern, but not Emotional Empathy, Predict Sensitivity to Injustice for Others. *Social Neuroscience*. *11*(1): 1-14.

Dellemotte, Jean. 2005. Sympathie, désir d'améliorer sa condition et penchant à l'échange. *Cahiers d'Économie Politique / Papers in Political Economy, 1*(48): 51-78.

Engelmann, Dirk and Martin Strobel. 2004. Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. *American Economic Review, 94*(4): 857-869.

Falk, Armin and Urs Fischbacher. 2005. Modeling Strong Reciprocity. In Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr (eds), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press, 194-214.

Fehr, Ernst, and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics, 114*(3): 817-868.

Fehr, Ernst and Urs Fischbacher. 2003. The Nature of Human Altruism. *Nature, 425*: 785-791.

Fehr, Ernst and Urs Fischbacher. 2005. The Economics of Strong Reciprocity. In Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr (eds). *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press, 151-191.

Forman-Barzilai, Fonna. 2010. *Adam Smith and the Circles of Sympathy; Cosmopolitanism and Moral Theory*. Cambridge: Cambridge University Press.

Gordon, R. 1986. Folk Psychology as Simulation. *Mind and Language, 1*(2): 158–171.

Gordon, Robert M. 2009. Folk Psychology as Mental Simulation. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/fall2009/entries/folkpsych-simulation/>

Güth, Werner, Rolf Schmittberger and Bernd Schwarze. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization, 3*(4): 367-388.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. New York: Oxford University Press.

Herne, Kaisa, Olli Lappalainen and Elina Kestilä-Kekkonen. 2013. Experimental Comparison of Direct, General, and Indirect Reciprocity. *The Journal of Socio-Economics, 45*: 38-46.

Heyes, Cecilia. 2010. Where Do Mirror Neurons Come From? *Neuroscience Biobehavioral Review, 34*(4): 575-583.

Hoffman, Elizabeth, Kevin McCabe and Vernon Smith. 1996. Social Distance and Other-Regarding Behavior in Dictator Games. *American Economic Review, 86*(3): 653-660.

Hodgson, Geoffrey. 2007. Meanings of Methodological Individualism. *Journal of Economic Methodology, 14*(2): 211-226.
DOI : 10.1080/13501780701394094

Jackson, Frank. 1999. All That Can Be at Issue in the Theory-theory/Simulation Debate. *Philosophical Papers, 28*(2): 77-96.
DOI : 10.1080/05568649909506593

Kagel, John H. and Alvin E. Roth. 1995. *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

Keppler, Jan H. 2010. *Adam Smith and the Economy of the Passions*. London and New York: Routledge.

Konow, James. 2003. Which Is the Fairest One of All? A Positive Analysis of Justice Theories. *Journal of Economic Literature, 41*(4): 1188-1239.
DOI : 10.1257/002205103771800013

Kreps, David M. and Robert Wilson. 1982. Reputation and Imperfect Information. *Journal of Economic Theory, 27*(2): 253-279.
DOI : 10.1016/0022-0531(82)90030-8

Lucas, Robert Jr. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, *1*(1): 19-46.

Morrow, Glen R. 1923. The Significance of the Doctrine of Sympathy in Hume and Adam Smith. *The Philosophical Review, 32*(1): 60-78.
DOI : 10.2307/2179032

Mukamel, Roy, Arne D. Ekstrom, Jonas Kaplan, Marco Iacoboni, Itzhak Fried. 2010. Single-Neuron Responses in Humans during Execution and Observation of Actions. *Current Biology, 20*(8): 750-756.

Rabin, Matthew. 1993. Incorporating Fairness into Game Theory and Economics. *The American Economic Review, 83*(5): 1281-1302.

Raphael, David D. 2007. *The Impartial Spectator; Adam's Smith's Moral Philosophy*. Oxford: Clarendon Press.

Rizzolatti Giacomo, Fadiga L., Fogassi L., Gallese V. 1996. Premotor Cortex and the Recognition of Motor Actions. *Cognitive Brain Research, 3*(2): 131-141.
DOI : 10.1016/0926-6410(95)00038-0

Rodriguez-Lara, Ismael and Luis Moreno-Garrido. 2012. Self-interest and Fairness: Self-serving Choices of Justice Principles. *Experimental Economics, 15*(1): 158-175.

Servátka, Maros. 2010. Does Generosity Generate Generosity? An Experimental Study of Reputation Effects in a Dictator Game. *The Journal of Socio-Economics, 39*(1): 11-17.

Singer, Tania, Ben Seymour, John O'Doherty, Holger Kaube, Raymond J. Dolan, Chris D. Frith. 2004. Empathy for Pain Involves the Affective but not Sensory Components of Pain. *Science, 303*(5661): 1157-1162.

Singer, Tania, and Ernst Fehr. 2005. The Neuroeconomics of Mind Reading and Empathy. *American Economic Review, 95*(2): 340-345.

Singer, Tania. 2008. Understanding Others: Brain Mechanisms of Theory of Mind and Empathy. In: Glimcher Paul W., Colin F. Camerer, Ernst Fehr and Russel A. Poldrack (eds), *Neuroeconomics: Decision Making and the Brain*. London: Academic Press, 251-268.

Small, Deborah A., George Loewenstein and Paul Slovic. 2007. Sympathy and Callousness: The Impact of Deliberative Thought on Donations to Identifiable and Statistical Victims. *Organizational Behavior and Human Decision Processes, 102*(2): 143-153.

Smith, Adam. [1790] 2005. *The Theory of Moral Sentiments*. David D. Raphael and A. L. Macfie (eds). Indianapolis: Liberty Fund.

Spreng, R. Nathan, Margaret C. McKinnon, Raymond A. Mar and Brian Levine. 2009. The Toronto Empathy Questionnaire: Scale Development and Initial Validation of a Factor-Analytic Solution to Multiple Empathy Measures. *Journal of Personality Assessment, 91*(1): 62-71.

Sugden, Robert. 2002. Beyond Sympathy and Empathy: Adam Smith's Concept of Fellow-feeling. *Economics and Philosophy, 18*(1): 63-87.

Takagishi, Haruto, Shinya Kameshima, Joanna Schug, Michiko Koizumi and Toshio Yamagishi. 2010. Theory of Mind Enhances Preference for Fairness. *Journal of Experimental Child Psychology, 105*(1-2): 130-137.

de Waal, Frans. 2010. *The Age of Empathy: Nature's Lessons for a Kinder Society*. London: Potter Style.

Young, Jeffrey T. 2013. A Review of Some Recent Smith Scholarship. *Œconomia – History/Methodology/Philosophy, 3*(1): 147-164.

Notes

1 See in particular Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Fehr and Fischbacher (2005), Falk and Fischbacher (2005).

2 See Camerer (2003) for a review of this literature.

3 In fact, the first results that challenged the standard model are much older. See Kagel and Roth (1995) for a brief history of experimental economics.

4 In the ultimatum game, a first person (player A) must decide how to share a certain amount of money with a second person, the recipient, (player B). In a second step, player B must decide whether to accept or reject the offer. In case of rejection, neither individual receives any money.

5 The dictator game differs from the ultimatum game in that the recipient does not have the possibility to reject the proposal, strategic aspects being thus eliminated.

6 Two categories of interpretations can be distinguished: In a first category interpretations are in terms of bounded rationality and challenge the agent's logic capacity to process the relevant information; in the second one the interpretations challenge the homo oeconomicus' selfishness.

7 A second influential model of social preferences proposed by Bolton and Ockenfels (2000) was different in several dimensions and somewhat less successful, maybe because it was less tractable.

8 See the ranking on the Web site of the *Quarterly Journal of Economics*

9 The literature (see for example Ben-Ner et al., 2004; Servátka, 2010; Bowles and Gintis, 2011; Herne et al., 2013) distinguishes at least three different notions of reciprocity that can be easily analyzed in the context of a dictator game. Strict reciprocity is observed in a situation where an individual B who was the recipient of a dictator A and therefore suffered the consequences of A's choice of sharing, now finds himself dictator for that same individual A. In this type of protocol, the behavior of type B subjects is interpreted as an expression of strict reciprocity when there is a strong correlation between the choice of what A gives to B and the choice of what B gives to A. Subject B is generous if A was generous to him and gives little or nothing if subject A was selfish. Generalized reciprocity can be observed in a protocol where the recipient B who observes what dictator A gave him, then, becomes the dictator for a third subject C. The term generalized reciprocity is used when the choices of B subjects are highly correlated with the choices of A subjects, regardless of the fact that C subjects are in no way responsible for the choice of A subjects. Finally, indirect reciprocity can be observed in a protocol where a subject C observes the behavior of an A dictator toward a subject B and then becomes himself a dictator for subject A. The literature shows generally the existence of reciprocity, especially strict and indirect. Results regarding generalized reciprocity are much less conclusive.

10 Robert Lucas (1976) formulates a systematic critique of the assumption that economic agents do not change their behavior ac as a function of any economic policy that is implemented. This is an assumption that is used in economic forecasting methods and especially to analyze the impact of economic reforms; The Lucas critique is significant in the history of economic thought as a representative of the paradigm shift that occurred in macroeconomic theory in the 1970s as macroeconomists tried to build sound micro-foundations for their macroeconomic models.

11 The history of this collaboration is described in Chapters 1 and 2 of Henrich et al. (2004).

12 See Bréban (2017) for an interpretation of Smithian sympathy in terms of emotions and cognition and an explanation of the passage from the cognitive to the emotional realm.

13 It should be noted however that some authors claim that the reaction of mirror neurons may be the result of adaptive social learning rather than inherited (see Heyes 2010).

14 See for example Takagishi et al. (2010) and Tania Singer (2008).

15 They did not confirm the first prediction neither. In particular, the measure of capacity of mindreading did not correlate with the accuracy of stated beliefs.

Pour citer cet article

Référence papier
Sylvie Thoron, « Morality Beyond Social Preferences: Smithian Sympathy, Social Neuroscience and the Nature of Social Consciousness », *Œconomia*, 6-2 | 2016, 235-264.

Référence électronique
Sylvie Thoron, « Morality Beyond Social Preferences: Smithian Sympathy, Social Neuroscience and the Nature of Social Consciousness », *Œconomia* [En ligne], 6-2 | 2016, mis en ligne le 01 juin 2016, consulté le 11 octobre 2016. URL : http://oeconomia.revues.org/2373 ; DOI : 10.4000/oeconomia.2373

Auteur

Sylvie Thoron

Université Paris-Est Créteil, LIPHA Paris-Est, sylvie.thoron@u-pec.fr

Droits d'auteur