INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH
ICER

WORKING PAPER SERIES

M. Stimolo

MULTIPLE-SELF MODELS IN NEUROECONOMICS
A METHODOLOGICAL CRITIQUE

Working Paper No. 7/2012

Marco Stimolo

Multiple-self models in neuroeconomics
A methodological critique

University of Naples “Federico II” Department of Economics & ICER

September 2012

Abstract. The idea of multiple-self models in economics is that individual identity is the equilibrium result of the strategic interaction between sub-personal selves. These models fill the gap of standard rational choice theory in explaining inter-temporal inconsistency of choices. This modelling procedure requires an extension of revealed preference theory to the sub-personal level. This extension is grounded in the assumption that sub-personal selves are economic agents to whom analytical tools of microeconomics apply. I claim that this assumption is false and entails the empirical methodology of functional localization that fails to provide robust results.

Keywords. Multiple-self, rationality, as if, functional localization, robustness.

1. Introduction

The idea of multiple-self models (MSMs) in economics is that individual selfhood is the equilibrium result of the strategic interaction amongst sub-personal selves. This view is at odds with the approach of mature neoclassicism in standard rational choice theory (RCT) that implicitly assumes selfhood to be unitary (Becker 1976).

The alternative way of modelling selfhood provides three major results. First, it offers a foundation of resolute choices as either a sequential or simultaneous game between sub-personal selves (Elster 1986, Schelling 1980). Furthermore, it explains weakness of will (*akrasia*) through hyperbolic discount curves (Ainslie 2001). Thirdly, MSMs explain the dynamic inconsistency of preferences in terms of unstable equilibria of the interaction between sub-personal selves (Montague et al. 2006, Ross et al. 2007, 2008).

The paper focuses only on the third subset since it refers to a particular use of MSMs in neuroeconomics where sub-personal selves are identified with neural sites[1]. In this field, the recent contributions characterize the brain as an equilibrium state of the strategic interaction between neural sites. This modelling procedure bears on the extension of revealed preference theory (RPT) to the neural level. This extension allows neural sites to be described as economic agents that interact amongst each other to bring about a given choice behaviour.

I criticize this subset of MSMs at the theoretical and empirical level. On the theoretical level, I show that the extension of RPT to the neuro-level is devoid of causal assumptions because the principle of utility maximization is a systematic description *and not* a causal explanation of behaviour[2]. For this reason, modelling the interaction amongst neural sites in RPT terms misses the target of a causal explanation of behaviour.

Yet RPT extension to the neuro-level has empirical consequences since it characterizes in linear terms the causal relation between neural activity and behaviour. In neuroscience a causal relation is linear when a behaviour directly follows from the

[1] These models differ from those proposed by Schelling, Elster and Ainslie that identify sub-personal selves at the behavioural level and do not require a neural foundation: the argument of the paper applies to MSMs in neuroeconomics and not *necessarily* to those in behavioural economics.

[2] The distinction between systematic description and causal explanation of behaviour *does not* entail a consideration of RPT as a system of non-falsifiable tautologies. On the contrary, a systematic description of behaviour as consistent can be empirically tested even if it black-boxes the *causes* of consistency.

activity of a neural site. Hence MSMs entail the empirical methodology of functional localization that aims to identify the activity of neural sites as the linear cause of behaviour. In this regard I analyze examples of how functional localization fails to provide robust results. This methodology produces false positives since it finds «regions of the brain that seem to be identified with a certain task in one study, but not in other studies» (van Rooij and Van Orden 2011: 40).

Critics of MSMs in neuroeconomics limit themselves to questioning their theoretical legitimacy, whilst retaining functional localization as reliable empirical methodology (Camerer 2004). They refute the accusation of the lack of robustness by referring to the contingent limitation of the instruments of observation[3]. Conversely, I show that lack of robustness arises from the theoretical problems of MSMs. Identification of the intrinsic link between the methodological fallacies of MSMs at the theoretical and empirical level is my intended contribution.

The paper is organized as follows. In section 2 I present MSMs as an integration of the explanatory lack of RCT. In section 3 I illustrate how MSMs bear on the extension of RPT to the neural level. In section 4 I claim that this extension is not grounded in causal assumptions. In section 5 I show how this extension limits the explanatory power of MSMs by entailing functional localization methodology that fails to provide robust results.

2. The basic argument of multiple-self models

Here I describe how MSMs fill the gap of RCT by explaining the intrinsic link between self-defeating behaviour and dynamic inconsistency of preferences in terms of the strategic interaction amongst neural sites. This modelling procedure is an integration of this explanatory shortcomings since the interaction amongst reward centres is described by means of the analytical tools of RCT. In sum, MSMs require an extension of RCT to the neuro-level.

Dynamic inconsistency of preferences is so described: when a smaller\earlier reward and a delayed\larger one are distant in time individuals rationally prefer the

[3] The main instruments of observation are functional magnetic resonance (FMRI) and positron emission tomography (PET). Both of them aim at identifying neural activity as linear cause of behaviour. My critique focuses on the general assumptions of functional localization and not on its technical applications.

delayed\larger one, but when the smaller\earlier reward gets closer they may reverse their preferences. Under this description fall the empirical instances of addiction and pathological gambling. Indeed, these kinds of self-defeating behaviour are characterized by dynamic inconsistent preferences. For example, an alcoholic might prefer not to drink (delayed/larger reward) when the drinking occasion (earlier\smaller reward) is far off. However, when the possibility of drinking is at hand (s)he might reverse his/her preferences. Why does this happen?

In standard RCT, self-defeating behaviour is either an error in computing the long-run effects of the two rewards or a matter of pathological preferences. If self-defeating behaviour is a computational mistake, then it can be corrected by learning. Conversely, if it is pathological, then economics can only explain how subjects, *given* their pathological preferences, maximize their utility. To analyze these phenomena RCT endorses the exponential discount rule, according to which individuals discount future rewards at a *constant* rate (Koopmans 1960). Once the discount rate is given, individuals' preferences for either smaller or larger rewards are *fixed* and they consistently maximize their inter-temporal utility function.

The case of computational error refers to the fact that if an agent at time t chooses A over B and at time t_{+1} B over A, then this is due to a change in rational beliefs in light of new information. The case of pathological preferences refers to the theory of rational addiction (Becker and Murphy 1988). In this framework alcoholics sharply devalue the future. As a result, they consider the utility of drunkenness to exceed high opportunity costs and maximise their expected utility *given* their pathological preferences. Their loss in welfare is explained by the fact that alcohol consumption raises their tolerance threshold. Hence, opportunity costs rise while pathological preferences persist: addiction is explained through rational consistency. In both cases RCT, by assuming a constant discount rate, fails to explain the dynamic inconsistency of preferences at the roots of self-defeating behaviour.

MSMs fill this gap by explaining self-defeating behaviour in terms of an internal conflict amongst rewards centres partitioned into two categories: far-sighted selves that prefer the delayed\larger reward and follow the exponential discount rule; short-sighted selves that prefer earlier\smaller rewards. The latter follow the hyperbolic discount rule according to which future rewards are discounted in proportion to their delay. When

short-sighted selves “win” the sub-personal conflict then the subjects engage in self-defeating behaviour; this means that they discount the future hyperbolically so that when the smaller reward gets closer in time they value it more than the larger\delayed one.

MSMs are consistent with recent developments in neurosciences, according to which preferences are the result of the coordination of different reward centres. This coordination is variable and determines preference instability (Dolan & Sharot 2012). MSMs in neuroeconomics seek to integrate RCT with this branch of neurosciences.

In this modelling procedure, the dynamic inconsistency of choices is explained in terms of unstable equilibria of the sub-personal interaction. The intra-individual dynamics is described as follows: short-sighted selves “win” the conflict and cause self-defeating behaviour. This equilibrium can be forestalled by far-sighted selves that induce time-consistent behaviour. Again, consistent behaviour can be forestalled by short-sighted selves and so on. Consider for example the trigger mechanism of addictive behaviour. Addictive agents (short-sighted self) crowd out salient stimuli that are not drug predictors, thereby focussing attention on stimuli that *are*. By the same token, addictive agents can be crowded out by the orbifrontal cortex (far-sighted self) that prevents them taking exclusive control of behaviour (Shiv et al. 2005).

However, self-defeating behaviour is a case study of MSMs and not their exclusive domain of application. Indeed, the extension of RCT to the sub-personal level can explain other kinds of heterodox behaviour in terms of the interaction amongst reward centres. An example of this would be cooperative behaviour (Sanfey et al. 2003) and the interaction effect between the structures of beliefs and payoffs in choice problems under risk and ambiguity (Hsu et al. 2005).

This bottom-up approach can be applied to several instances of behavioural instability. However, the assumption of stable preferences is crucial for the analytical tools of RCT to work. Thus, MSMs in neuroeconomics require that reward centres be “economic agents” with “stable preferences” revealed in stable behavioral responses. Besides the metaphorical phrasing, MSMs have to assume behavioral stability at the neuro-level to explain unstable behaviour at the individual level. The following section analyses this implication.

3. Neural site as the economic agent

In this section I show that RPT, by virtue of its behaviourist interpretation, provides the theoretical framework to model neural sites as economic agents. More precisely, I claim that the behaviourist interpretation of RPT provides the two core features of MSMs in neuroeconomics: the definition of rationality as behavioural stability and the duality relationship between equilibrium and its individual components. On these grounds MSMs can implement the analytical tools of standard RCT at the neuro-level.

RPT defines preferences in terms of choices that in turn are defined in terms of consequent outcomes. With this definitional *continuum* the problem of the indirect observation of preferences can be avoided since they are determined by *counterfactual* choices among alternatives. The necessary and sufficient condition for a duality relationship between preferences and choices is the strong axiom (SA) of revealed preferences. By SA a rational preference relation generates a preference-revealing choice structure; analogously, a preference-revealing choice structure is rationalized by a consistent preference relation[4].

The interpretation of RPT based on SA has two major implications: rationality is a *positive* property of behavioral stability; preferences and equilibrium stand in a duality relationship. These two implications allow RPT to be extended to the neuro-level.

As regards the former, if preferences stand in a duality relationship with choices, then they are tantamount to an empirical matter and do not need any normative justification. This means that if SA is satisfied, then rationality is a *positive* property of behavioral stability. The latter is defined as a stable relation between goals and actions. This definition is behaviourist since it black-boxes any internal structure that could be separated from observable behaviour. Moreover, identification of rationality with behavioral stability means that intentions and deliberation are not *necessary* conditions for rational behaviour.

This behaviourist interpretation is consistent with the absence of ontological commitment of RPT. By ontological commitment I mean identification of a kind of entity as the exclusive bearer of the properties that the theory postulates. In this

[4] Notice the difference with the weak axiom of RPT. This is a necessary but not sufficient condition imposed on the relation between preferences and choices. This means that rational preference ordering generates a preference-revealing choice structure. By the same token, we cannot *always* claim that this choice structure can be rationalized in terms of a consistent preference relation (Mass-Colell 1995).

behaviourist framework RPT does not require identification of human beings as exclusive bearers of the positive properties of rationality. The only condition imposed on identification is the definition of behavioral stability. Hence, *any entity* exhibiting a high degree of behavioral stability is amenable for RPT-based explanations.

With regard to the second implication, if rational choices satisfy SA, then the defining assumptions on equilibrium and those on preferences stand in a duality relationship. By virtue of that, rational preferences can be deduced from equilibrium and *vice versa*. An example is provided by Ross (2005: 237-258), who claims that rational preferences can be attributed to economic agents by deduction from equilibrium. Albeit not explicit, Ross's assertion can be supported only if it is assumed that choices at equilibrium satisfy SA.

These two implications allow the interaction of reward centres to be modelled in RPT terms. In neuroeconomics reward centres are assumed to be behaviourally stable. At the neuro-level the definition of behavioral stability is further qualified as the *invariance* of the function of a reward centre *vis-à-vis* its interaction with other reward systems. For example, time-inconsistent behaviour is explained as the equilibrium result of the interaction between prefrontal cortex – far-sighted self - and the limbic system – short-sighted self (McClure et al. 2004). The dynamics of the isolated functional responses of the two systems is not modified by their interaction (van Rooij and Van Orden 2011). Hence RPT, by virtue of the definition of rationality as behavioral stability, can be extended to the neuro-level. That is, reward centres can be modelled as economic agents interacting to maximise their utility.

The extension of RPT to the neuro-level postulates a duality relationship between neural equilibria and the functions of reward centres. This means that we can derive the functions of reward centres from the equilibrium of their interaction and *vice versa*. In other words a behaviour can be *correlated* to the activity of neural sites. For example, from time-inconsistent behaviour we can derive the fact that the limbic system (short-sighted self) is more active than the orbifrontal cortex (far-sighted self). By the duality relationship, from the higher activation of the limbic system we can infer time-inconsistent behaviour (McClure et al. 2004).

In sum, MSMs explain time-inconsistent behaviour at the personal level by assuming that reward centres are rational in the positive sense of behavioral stability and that their

stable functions stand in a duality relationship with the equilibrium of their interaction. In the next section I analyse the methodological legitimacy of this move at the theoretical level.

4. Methodological fallacies of multiple-self models

Here I criticise the extension of RPT to the neuro-level by referring to the argument that causal statements are not derivable from axioms of rationality and the principle of utility maximisation[5]. On this basis I identify *three* methodological fallacies of MSMs in neuroeconomics. The first concerns the fact that modelling reward centres as economic agents blurs the distinction between systematic description and causal explanation of behaviour. Moreover, the evolutionary process that brought about sub-personal economic agents is modelled by *analogy* with the principle of utility maximisation. Thirdly, these two methodological issues imply an *unjustified* equivalence between *utility* and *fitness* maximisation.

The claim that the interaction amongst utility-maximising reward centres *causes* self-defeating behaviour overlooks the distinction between systematic description and causal explanation of behaviour. Indeed, even in a behaviourist interpretation of RPT, rationality axioms systematically describe the properties of consistency between preferences and choices on a *pure* logical level. Thus, the theoretical axiomatic core establishes the conditions to describe behaviour as utility-maximising, but it is silent on the *causes* of utility maximisation.

The systematic description of behaviour is synthesised by the representation theorem of expected utility theory: individuals whose preferences satisfy axioms of rationality behave *as if* they were maximising an expected utility function (von Neumann, Morgenstern 1944). The clause "as if" means that the description of behaviour as rational depends on whether preferences and choices satisfy such axioms (Lehtinen and Kuorikoski 2007). The fulfilment of axioms is compatible with a wide range of causes, but axioms themselves are not causal assumptions. This means that the theoretical axiomatic core black-boxes the mental operations compatible with utility maximisation, which is not a mental process.

[5] I refer to the axioms of rationality and the related principle of utility maximization provided by expected utility theory.

In MSMs reward centres are systematically described *as if* they were maximising their utility functions (Montague and Berns 2002, Montague et al. 2006, Ross 2005, Ross et al. 2008)[6]. In this regard, the methodological fallacy is the following: reward centres are intended to be the empirical basis to derive from an *as if* assertion – rational agents behave *as if* they were maximising their utility function – the conclusion that the interaction amongst utility-maximising reward centres *causes* self-defeating behaviour. This conclusion is devoid of explanatory power[7].

The same kind of methodological fallacy concerns the MSMs' evolutionary hypothesis according to which natural selection follows the principle of utility maximisation in selecting reward centres. This hypothesis is the result of an inference: if reward centres *are* utility maximisers then natural selection follows the principle of utility maximisation. However, the description of sub-personal selves in terms of utility maximisation is devoid of causal assumptions. By contrast evolution is causal in character (Sugden 2001). In sum, from the premise of the inference we cannot derive the *causal* conclusion. Therefore sub-personal selves are not the empirical basis to derive from an *as if* proposition the *causal conclusion* that evolution follows the principle of utility maximisation.

Nonetheless, the assumption of utility-maximising natural selection is necessary for MSMs to characterize the brain as an equilibrium of the strategic interaction amongst sub-personal rational agents. However, the recent developments of evolutionary biology and game theory provide counter-examples of the fact that nature does not maximise utility in selecting simple organisms. The following example by Sugden clarifies the point:

«Suppose that for some locus on a chromosome there are two possible genes, A and a. Thus there are three possible genotypes, AA, Aa and aa. Suppose that of these, Aa confers the greatest reproductive success. Because of the facts of genetics, it is not possible to have a population which contains only Aa genotypes: the equilibrium state of the gene pool is a mix of A and a, and so all three genotypes survive. The population thus contains stable proportions of […] unfit phenotypes which correspond with AA and

[6] This *as if* methodology is implicit in most of the MSMs analyses in neuroeconomics. In this regard Berg and Gigerenzer (2010) provide a lucid critique of *as if* methodology in neuro and behavioral economics.

[7] Neuroeconomists are aware of the methodological fallacy of blurring the difference between systematic description and causal explanation. To avoid such fallacy, they derive the utility functions of a given neural network from the determination of its computational algorithm that tracks a causal regularity. However, this strategy bears on a linear concept of causation and on the methodology of functional localization. The latter produces many false positives in the identification of a causal nexus between neural activity and behaviour. I analyze the point in section 5.

aa. […] The moral of this example is that the phenotypes that are selected by evolutionary processes do not necessarily maximize anything at all» (*Ivi*: 224).

Natural selection can be "forced" to optimise if we model it through choice functions that are *assumed* such that they give the actual results[8]. MSMs in neuroeconomics are an instance of this *as if* methodology. In this framework the assumption of optimising natural selection is an analytical truth and not an empirical result. However, given that the target of *explanandum* of MSMs is empirical in character, the models run up against the fallacy of grounding their explanations in an analytical truth rather than in realistic causal properties.

Due to these two fallacies MSMs present an unjustified equivalence between *utility* and *fitness* maximisation. Even the behaviourist interpretation of RPT defines utility functions as a numerical representation of well-ordered preferences (Ross 2005). In this fashion, utility functions are unrelated to any psychological quantity to be maximised. Therefore, the principle of utility maximisation is an analytical assertion from which we cannot derive causal statements on choice behaviour. By contrast, the evolutionary process is grounded in the principle of fitness maximisation. The latter is a *tautology* but it tracks the empirical regularity of *reproductive success*. Hence evolutionary theory lends itself to causal explanations of behaviour while RPT does not.

MSMs in neuroeconomics assume that rewards centres maximise fitness in the specific sense of maximising their level of activity. Utility functions are constructed on the range of these activity levels. However, utility maximisation is not a causal process, unlike the maximization of activity levels. Hence, modelling the activity of reward centres *as if* they were maximizing a utility function does not capture a causal process.

This methodological critique shows how the description of reward centres as sub-personal economic agents is devoid of causal assumptions. However, MSMs in neuroeconomics provide a characterization in linear terms of the causal nexus between neural activity and behaviour. Hence, the explanatory power of MSMs can be assessed at the empirical level. I tackle the problem in the next section.

[8] I am grateful to John Collier for this suggestion.

5. MSMs and functional localization in neuroeconomics

Here I claim that MSMs in neuroeconomics entail the empirical methodology of functional localization of neural sites as *linear* causes of choice behaviour (5.1). In this regard I give examples of how this methodology fails to provide robust results (5.2).

5.1. Grounding assumptions of MSMs and functional localization

Here I claim that MSMs entail functional localization methodology because they both bear on the assumptions of behavioral stability and equilibrium. These assumptions are necessary and jointly sufficient conditions to identify a *linear* causal relationship between neural activity and choice behaviour. More precisely, the two assumptions provide the operational hypothesis according to which a given behaviour can be inferred from the most active neural site.

In neuroscience a reward centre is assumed to be behaviourally stable because it implements the same function *independently* of its interaction with other neural sites. Consider the neural sites involved in time-inconsistent behaviour: prefrontal cortex and limbic system (McClure et al. 2004). The respective functions of these two systems are identified in isolation and they are assumed to be stable across interactions.

If neural functions are stable across interactions, then to localize the most active neural site it is enough to subtract the activity of the others that are involved. For example, in binary decision problems between earlier/smaller and delayed/larger rewards, the most active neural site is localized by subtracting the activity of the prefrontal cortex from the activity of the limbic system (*Ibidem*). Thus, behavioural stability, qualified in terms of functional invariance, is *necessary* to identify the difference in the activity levels between neural sites.

Moreover, the interaction between functionally stable neural sites is assumed to reach equilibrium. The fluctuation of brain dynamics around an equilibrium point is assumed to hold even when we compare the average values of the activity of a neural site in two different points in time. This assumption is crucial to explain time inconsistency as a suboptimal equilibrium result of interacting reward centres.

As claimed in section 3, extension of RPT to the neuro-level postulates a duality relationship between the activity levels of reward centres and equilibrium results of their interaction. As a result, we can infer behaviour from the different levels of activation of reward centres. For example, we can infer time-inconsistent behaviour from the higher activation of the limbic system with respect to the prefrontal cortex.

These two assumptions at the roots of both MSMs and functional localization are *necessary* and *jointly sufficient* conditions for a *linear* causal relationship between the interaction of reward centres and choice behaviour. If this is the case, then the explanatory power of MSMs in neuroeconomics can be assessed on the grounds of the robustness of the empirical results that functional localization provides. I encounter the problem in the next section.

5.2. Examples of non-robust results

An empirical result is robust if it does not vary with the controlled variation of the experimental design (Guala 2005). Grounding on this definition, I focus on three examples of how functional localization fails to provide robust results: inter-temporal choices, cooperative behaviour, choices under risk and ambiguity. Each example represents an implication of the problem of lack of robustness. Experimental evidence on inter-temporal choices does not distinguish dual and unitary models of the brain. The neural correlates of cooperative behaviour in ultimatum and trust games are distributed through the brain and functional localization fails to identify causation within networked systems. Experiments on choices under risk and ambiguity show that the same brain area can be associated to different behaviours.

Experimental evidence on inter-temporal choices is not robust since it does not distinguish between the dual and unitary models of the brain. Empirical evidence is distinguishing if it exclusively supports one of the explanatory hypotheses of a phenomenon. On this basis I compare experiments performed by McClure et al. (2004) and Glimcher et al. (2007).

McClure et al. tested the hypothesis of the dual model of the brain according to which in binary choices among smaller/earlier and larger/delayed rewards the lateral prefrontal cortex is correlated with long-run preferences and the limbic system with

those in the short run. In this framework, the prefrontal cortex exhibits an exponential discount rate while the limbic system a hyperbolic rate. The authors found the following evidence to support their hypothesis: the limbic system worked harder for smaller/earlier rewards and exhibited a hyperbolic discount rate; conversely, the prefrontal cortex worked harder for larger/delayed rewards and exhibited an exponential discount rate. Hence, the higher activation of the limbic system with respect to prefrontal cortex in choices of smaller/earlier rewards was the basis to infer time-inconsistent behaviour.

Glimcher et al. (2007) tested the hypothesis of a unitary neural correlate of time-inconsistent choices. They started gathering behavioural data in binary choices and then they localized the neural correlate. Behavioural data were explained by hyperbolic discount functions. This explanation is consistent with the dual model of the brain. However, neural data did not support the dual system hypothesis. The authors identified the ventrial striatum, medial prefrontal cortex and the posterior cingulate cortex as a *single* reward centre of subject's discounted utility. As they pointed out «we saw no evidence of separable neural agents that could account for the multiple selves that are used to explain hyperbolic-like discounting behavior» (*ivi*: 143).

It is evident that functional localization methodology provides *conflicting* neural data that fit the *same* behaviour. Hence, experimental evidence on inter-temporal choices does not distinguish between dual and unitary models of the brain.

Experimental evidence on ultimatum and trust games is not robust because functional localization fails to identify causation within the networked system at the roots of cooperative behaviour. To illustrate this point, I compare the studies of Sanfey et al. (2003), King-Casas et al. (2005) and McCabe et al. (2001).

In their study on ultimatum games Sanfey et al. tested the hypothesis according to which the decision of rejecting unfair offers is the result of the competition between cognitive and emotional motives. The former refers to the choice of accepting any offer to maximize the amount of money. The latter refers to the decision to decline unfair offers. The authors localized two brain areas involved in the reactions to unfair offers: the bilateral anterior insula, which was more active in rejecting unfair offers, and the dorsolateral prefrontal cortex, which was more active in deciding to accept unfair offers.

Thus, the higher activation of the bilateral anterior insula was the basis to infer the choice of rejecting unfair offers in ultimatum games.

In their experiment on trust games King-Casas et al. aimed to localize the neural correlate of the reaction to fair offers framed as an intention to trust. In trust games the proposer has to divide a surplus with the responder. The latter receives a tripled amount of the chosen share and then decides whether or not to return some money to the proposer. The game solution is that the proposer does not trust the responder to return something, so he does not offer anything. The game was repeated 10 times with the same subjects. The behavioral data falsified the game solution: agents were willing to trust and reciprocate with fair offers. As regards the neural correlate, the caudate nucleus was the most active area in responses to fair offers.

The reactions to fair and unfair offers are located in different brain areas: the anterior insula formulates responses to unfair offers and the caudate nucleus to fair offers. These systems should be connected by a dispatcher that sends the correct emotional signal to the most active brain area. However, evidence on this connection is lacking (van Rooij & Van Orden 2011: 36).

Moreover, the study of McCabe et al. on repeated trust and punish games provided conflicting results. The authors tested the hypothesis that cooperation requires overcoming the desire for an immediate reward to pursue a delayed/larger one. The neural hypothesis was that the prefrontal cortex binds the attention to cooperation by inhibiting the desire for immediate rewards. Their results showed that the neural basis of cooperative behaviour is a networked system distributed across the brain: the prefrontal cortex, the occipital lobe, the parietal lobe and the thalamus. Within this networked system functional localization cannot identify different levels of activity between brain areas as the empirical basis to infer cooperative behaviour.

Experimental results on choice under risk and ambiguity are not robust because the same brain area (*orbifrontal cortex*) is associated to different behaviours. I refer to studies of Smith et al. (2002) and Hsu et al. (2005) which aimed to identify the neural correlate of the interaction effect between the belief structure (risk and ambiguity) and payoff structure (gains and losses). This effect drives subjects to be risk averse in gains and risk seeking in losses, while they are ambiguity seeking neither in gains nor losses.

From this evidence the authors derived the hypothesis that different neural sites are correlated with choices under risk and ambiguity.

In the experiment run by Smith et al. subjects had to choose between two urns with a different number of red, blue and yellow marbles, associated to different payoffs. The level of risk was varied using the range of payoffs. For example, if in an urn there are 30 red, 30 blue, 30 yellow marbles with respective payoffs of $50, $6, and $4, then the expected payoff of the urn is $20. The range of payoffs can be varied so that 30 red marbles pay $30, 30 blue marbles pay $30, and 30 yellow marbles pay $0; the urn expected payoffs is the same ($20), but the gamble is riskier. Ambiguity was created by giving the exact number of marbles of one colour and only the total sum of other marbles.

The behavioural data confirmed the interaction effect between the structures of beliefs and payoffs. Moreover, the authors identified two complementary neural systems in evaluating these two dimensions: dorsomedial neocortical system including *orbifrontal cortex* for risky judgments of losses and the ventromedial system for other stimuli (Smith et al. 2002: 717).

In their similar study Hsu et al. (2005) provided contradictory results. Indeed, they identified a correlation of *orbifrontal cortex,* amygdala and dorsomedial prefrontal cortex with choices in conditions of ambiguity. Furthermore, they identified the dorsal striatum as the neural correlate of risky choices.

These two studies show how orbifrontal cortex is correlated with both the behavioral response to risk (Smith et al. 2002) and ambiguity (Hsu et al. 2005). Given this non-exclusive association, it is not legitimate to infer the behavioral response to ambiguity or risk from the higher activation of orbifrontal cortex with respect to other brain areas.

Problems of lack of robustness can be found in experiments in phonology and psycho-linguistics. Indeed, lack of robustness is a *systematic* and *domain-general* problem of functional localization. This quandary is due to a theoretical commitment to a linear concept of causation. In MSMs this theoretical commitment is determined by the extension of RPT to the neuro-level. These models, bearing on the assumptions of behavioral stability and equilibrium state, characterize in linear terms the causal nexus between neural activity and behaviour. Thus, MSMs entail functional localization that fails to provide robust results.

6. Conclusions

In this paper I criticized MSMs in neuroeconomics both at the level of the methodological legitimacy of extending RPT to the neuro-level and at the level of the empirical methodology of functional localization that this extension entails. I showed that the extension of RPT to the neuro level does not provide any realistic causal features grounding the explanation of behaviour. On this basis I claimed that MSMs run up against the general problem of deriving causal statements from the systematic description of behaviour in terms of utility maximization.

Furthermore, I showed how these models entail the empirical methodology of functional localization of brain areas as linear causes of behaviour. In this regard, I analysed examples of how this methodology provides empirical results that are not robust. From these examples I drew the conclusion that the empirical fallacies of MSMs in neuroeconomics are an implication of their theoretical inconsistency.

7. References

Ainslie G. (2001) *Breakdown of Will*, Cambridge: Cambridge University Press.

Becker G. (1976) *The Economic Approach to Human Behavior*, Chicago III: University of Chicago Press.

Becker G. and Murphy K. (1988), A theory of rational addiction, *Journal of Political Economy*, 96: 675-700.

Berg N. and Gigerenzer G. (2010), As If Behavioral Economics. Neoclassical Economics in Disguise? *History of Economic Ideas*, 1: 133-165.

Camerer C., Loewenstein G., Rabin M. edited by (2004), *Advances in behavioral economics*, Princeton, Princeton University Press.

Dolan R.J. and Sharot D. (2012), *Neuroscience of preferences and choice. Cognitive and Neural Mechanism*, Elsevier.

Elster J., edited by (1986), *The Multiple Self*, Cambridge University Press.

Glimcher, P.W., Kable, J.W., and Louie, K. (2007), Neuroeconomic Studies of Impulsivity: Now or Just as Soon as Possible? *American Economic Review*, 97(2): 142-147.

Guala F. (2005), *Methodology of Experimental Economics*, New York: Cambridge University Press.

Hsu M., Meghana B., and Ralph A. (2005), Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making, *Science*, 310(5754): 1680–83.

King-Casas B., Tomlin D., Anen C., Camerer C. F., Steven R. Quartz, and Montague P. R. (2005) Getting to Know You: Reputation and Trust in a Two Person Economic Exchange, *Science*, 308(5718): 78–83.

Koopmans TC (1960), Stationary ordinal utility and impatience, *Econometrica*, 2: 287 309.

Kuorikoski J., Lehtinen A. and Marchionni C. (2010), Economic Modelling as Robustness Analysis, *British Journal for the Philosophy of Science*, 61: 541-567.

Lehtinen A. and Kuorikoski J. (2007), Unrealistic Assumptions in Rational Choice Theory, *Philosophy of Social Sciences*, 37: 115-138.

Mass-Colell A., Whinston M. D. and Green J. R. (1995), *Microeconomic Theory*, Oxford: Oxford University Press.

McCabe K. Houser D., Ryan L., Smith V. and Trouard T. (2001), A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange, *PNAS*, 98(20): 11832–35.

McClure, Laibson D. I., Loewenstein G., Cohen J. D. (2004), Separate Neural Systems Value Immediate and Delayed Monetary Rewards, *Science*, 15: 503-507.

Montague P.R., King-Casas B. and Cohen J. D. (2006), Imaging valuation models in human choice, *Annual Review of Neuroscience*, 29: 417-448.

Montague, P.R. and Berns G. (2002), Neural economics and the biological substrates of valuation, *Neuron*, 36: 265-284.

Montague, P.R., King-Casas, B. and Cohen, J. (2006), Imaging valuation models in human choice, *Annual Review of Neuroscience*, 29: 417-448.

von Neumann J. and Morgenstern O. (1944) *Theory of games and economic behaviour*, Princeton University Press, Princeton.

van Rooij M. and Guy Van Orden G. (2011), It's about Space, It's about Time, Neuroeconomics and the Brain Sublime, *Journal of Economic Perspectives*, vol. 25, 4: 31-56.

Ross D. (2005) *Economic Theory and Cognitive Science: Microexplanation*, MIT press.

Ross D., Spurrett D., Kincaid H., and Stephens L. edited by (2007), *Distributed Cognition and the Will: Individual Volition and Social Context*, MIT Press.

Ross D., Sharp C., Vuchinich R. and Spurret D. et al. (2008), *Midbrain Mutiny: The Picoeconomics and Neuroeconomics of Disordered Gambling*, MIT Press.

Rustichini A. (2009), Is There a Method of Neuroeconomics? *American Economic Journal: Microeconomics*, 1(2): 48–59.

Sanfey A. G., Rilling K. J., Aronson J. A., Nystrom L. E., and Cohen J. D. (2003), The Neural Basis of Economic Decision-Making in the Ultimatum Game, *Science*, 300(5626): 1755–58.

Schelling T., (1980), The Intimate Contest for Self-Command, *The Public Interest*, Vol. 60: 153-178.

Smith K., Dickhaut J., McCabe K. and Pardo J. V. (2002), Neuronal Substrates for Choice under Ambiguity, Risk, Gains, and Losses, *Management Science*, 48(6): 711–18.

Sugden R. (2001), Ken Binmore's Evolutionary Social Theory, *The Economic Journal* 111: 213-243.