RATIONAL CHOICE THEORY BETWEEN CAUSATION AND EXPLANATION

Jens Harbecke

Witten/Herdecke University, Germany
jens.harbecke@uni-wh.de
www.jensharbecke.com

Abstract

This paper focuses on two arguments recently developed in the literature against the interpretation of rational choice theory as an empirical theory. It starts with a reconstruction of a historical analysis, according to which rational choice theory has mostly been used in the past as a methodological principle and rarely as a deep empirical theory. In a next step, it challenges an argument found in the literature that social and economic phenomena are ontically emergent and that they by themselves can enter genuine explanations. Subsequently, it criticizes the methodological assumption about the irrelevance of psychological mechanisms of the individual for economic models. The main reason offered is the observation that such models, even if predictively adequate, will be very limited in their explanatory power. The overall conclusion of the paper is that rational choice theory ought to be treated as a theory after all – and potentially extended by future empirical research.

1 Introduction

Criticizing rational choice theory (RCT) for its presumed irrealistic assumptions and its alleged detachment from reality has been popular throughout its history inside and outside academic science. The most widespread line of argument against RCT has emphasized its empirical shortcomings as a theory of the cognitive life of human economic agents. Mill's notion of the *homo oeconomicus* has served as the prototypical target of this critical appraisal. For Mill, the concept of homo oeconomicus represented "(...) a being who desires to possess wealth, and who is capable of judging the comparative efficacy of means for obtaining that end." (Mill 1843, 137) The modern mathematical formulations of RCT have translated the notions of a "judgement of the comparative efficacy of means" and the "desire to possess wealth" into the formal notions of a "preference relation" and "utility maximization rule or algorithm". They usually define a complete and transitive (and potentially continuous,

monotone, locally non-satiated, convex) preference relation over a set of choices and a choice rule singling out the optimal choice under constraints. The preference relation is typically associated with a utility function specifying for each action a utility represented by ordinals, which can be fed into a maximization function.

The above-mentioned charge of a detachment from reality has emphasized that real humans of flesh and bones satisfy neither Mill's description of homo oeconomicus nor modern formulations of RCT. On the one hand, human action is caused by various motives. The desire to possess wealth is only one among many, and it is by far not shared by all humans. Secondly, choices in real-life seem to be highly context-dependent. For instance, the phenomenon of anchoring-and-adjustment in pricing decisions has been documented even for experts (cf. Northcraft and Neale 1987). Thirdly, it is argued that choices in real life can impossibly be based on a consideration and evaluation of all options. Rather, humans mostly make choices on the basis of "rules of thumb" or heuristic learning strategies. Moreover, they make regular mistakes in caculating optimal choices (cf. Thaler 1980).

It has recently been has argued by Catherine Herfeld (2014) that RCT in its historical versions was not as one-dimensional as the criticism voiced against it usually wants us to believe. Rather, RC theorists were usually very much aware of the fact that human nature involves many motives apart from pure self-interest. Moreover, many theorists did not even intend RCT to be a theory in the strict sense. In many cases, RCT rather served as a "methodological rule of modeling social phenomena with reference to the behavior of economic agents (individual or not individual), accounting for their behavior in terms of a highly idealized account of economic agents drawing upon the rationality-concept." (*op. cit.*, 269) Herfeld calls this status of RCT a *methodological rationalism*. "Economists' commitment to this doctrine has been unceasing in their attempts to solve manifold problems with different tools." (*op. cit.*, 269)

Herfeld herself defends a non-individualism of economics that makes only minimal rationality assumptions about economic agents. These minimal assumptions fall short of constituting a psychological theory. As she says, "(...) I doubt that focusing exclusively on the individual and his choices in economic investigation is an undertaking as worthwhile as the recent boom in behavioral economics makes us believe." (*op. cit.*, 272)

The two main arguments are offered for this non-individualist and minimal approach: one is based on methodological, the other on metaphysical considerations. The first starts from the observation that "on the aggregate level and within market contexts, individual deviations cancel each other out." (*op. cit.*, 272) The second departs from the metaphysical contention that "[s]ocial phenomena are not the objects of individual choice but rather emerge from the interactions of individuals. Under such conditions the explanation of individual behavior by a theory of individual choice appears to be useful only to a very limited extent." (*op. cit.*, 275) Herfeld admits that economics needs a minimal theory of the individual for its models. However, "it does not necessarily require a *psychological theory*." (*op.*

cit., 276) Both arguments have precursors in the literature (cf. Arrow 1986; Arthur 1999; Chen and Yeh 2002; Charness and Sutter 2012), and the non-individualist and minimal approach defended on their basis is popular and influential in many branches in schools of economics.

In this paper, I side with Herfeld's historical analysis of RCT. Only in some of its many versions RCT was actually intended to form something like a psychological theory of humans in economic contexts. In most of its versions, rational choice analysis (RCA) either took on a minimal stance on human psychology or declared only a portion of human cognition as relevant for economic contexts in the first place.

Notwithstanding, I want to challenge the two arguments accepted by Herfeld and others for the contention that minimal and non-psychological non-cognitive RCT is *adequate* for economic theorizing. In a first step, I will focus on the metaphysical consequences of the non-individualist and emergentist view. More specifically, I am interested in the contention about an emergence of social phenomena from individual actions and the metaphysics of causation. In a second step, I am inquiring into the explanatory power of RCT and economics that widely denies the relevance of the cognitive life of economic individuals.

With respect to the first issue, I will defend the claim that the status of social phenomena as emergent is problematic from a metaphysical point of view. Hence, I challenge the assumption that the emergence provides a good reason against individualist RCT. Secondly, although I agree that behaviourist approaches will not necessarily provide more accurate predictions on a macrolevel, I will defend the claim against Herfeld and others that economic theory lacks explanatory power unless an extended version of RCT is placed at its foundations.

The structure of this paper is the following. Section 2 reviews Herfeld's analysis of the history and nature of RCT and her conclusions with respect to its status as a theory of human agency. Section 3 discusses Herfeld's metaphysical argument against the requiredness of RCT as a theory of action for economics. It shows that the ermergent social phenomena Herfeld envisages face the "causal exclusion argument". Section 4 focuses on the methodological argument for the minimality of RCT. It emphasizes that, for economics and RCT to be truly explanatory, RCT either needs to be interpreted as a psychological theory or it needs to be supplemented with one, potentially with a neurocognitive theory. Section 5 summarizes the main points of the paper and lists some questions for further research on the topic of RCT, its metaphysical stance, and its explanatory power.

2 The Historical Faces of RCT

In her recent analysis of the history, interpretation, and reception of RCT, Herfeld convincly argues that the various accounts that have made use of a rationality concept are far more complex than critics tend to acknowledge. "Psychologists and be-

havioral economists that are mainly interested in the explanation of human behavior have, as philosophers, questioned RCT on the grounds that it does not adequately capture the causal processes behind individual behavior or frequently becomes rejected on the grounds that it is not testable." (Herfeld 2014, 17) She cites Philip Pettit as an example who rejects RCT for failing to provide a genuine causal account of human action and as such not offering adequate explanations of behavior in most social interactions. (*op. cit.*, 18) However, in Herfeld's opinion, "the psychologist and philosopher's view [of RCT] does not necessarily correspond to the economist's actual practice and goal." (*op. cit.*, 23)

Herfeld's own view is that RCT is not "(...) a specific theory for analyzing choice, although this might be one purpose for which it is used. Rather, RCT can best be understood as a highly flexible programmatic framework consisting of a family of models and theories that share certain features, i.e. family resemblances, but are also characterized by some fundamental differences." (*op. cit.*, 25)

In a first step, Herfeld shows that the concept of a self interested individual played a theoretical role at least from Adam Smith's *Wealth of Nations* onwards. Smith saw the efficiency of markets and a society's welfare to result from the self-interested actions of individuals. However, neither did he believe that self-interest was the only driving force of humans, nor did he aim at a detailed theory of individuals. The principle "(...) was not used as a theory that explained individual behavior by referring to the actual causes responsible for decision-making in particular cases. Smith did also not provide causal explanations of social phenomena by making precise the exact mechanism by which self-interested individuals, motivated as broadly understood as Smith did, would directly produce effects on market outcomes." (*op. cit.*, 72) His main focus was on the unintended consequences on the intermediate level within economic situations. In this sense, "Smith's principle of self-interest does not have to have the status of a psychological theory in order to make this explanation work." (*op. cit.*, 73)

Mill's well-known concept of *homo oeconomicus* can be characterized in a similar fashion. Mill was influenced by the utilitarian tradition of Bentham and others who developed the idea of a utility maximizing calculus of pleasures and pains. The homo oeconomicus judges means according to their calculated expedience for a specific goal. In economic contexts, this goal is usally "wealth". But, as Herfeld points out, also Mill's concept is not to be confused with a psychological theory of humans. Rather, it is a methodological prerequisite for an abstract or "pure" political economy. "Mill believed that narrowing down the set of psychological laws to only the pursuit of wealth while abstracting from all other human motives would enable the political economist to deduce economic laws (...)." (*op. cit.*, 89) Hence, it would be misleading to consider Mill's economizing agency as an exhaustive theory of human behaviour.

The Marginalist revolution commonly associated with William Stanley Jevons (1835-1882), Carl Menger (1840-1921), and Leon Walras (1834-1910) added to the toolbox of RCT the ideas that economic value is subjective and that the utility

or benefit derived from the last portion of a consumed good decreases in degree. These assumptions made the original concepts of utility maximization and choice calculation more fine-grained and detailed. To the extent that they believed in the applicability of their theories to real-world cases, the Marginalists did aim at a more detailed psychological theory of humans as consumers. Moreover, Jevons even believed in the measurability of utility through observable behavior and its quantifiability through prices. The error that could be expected in any actual measurements of behaviour Jevons believed to cancel out in the aggregate. In Herfeld's view, this shift in focus to the aggregate shows that also the Marginalists were far from believing their principles to apply to real human beings.

The Marginalists' shift away from individual psychology became more obvious through Pareto's ordinal revolution. Pareto offered a formal model for a preference ordering that does not have to satisfy the property of cardinality but instead can be represented by an ordinal scale. His proposal partly stemmed from his skepticism about the possibility to measure utility in practice (cf. Herfeld 2014, 119). The only thing that could be observed and tested is that humans order form and order options in a relatively consistent way. However, this minimal assumption hardly forms a psychological theory. It merely forms the basis for the definition of indifference curves and, if aggregated, as an explanation for the shape of the demand curve.

The axiomatized versions of RCT developed in the 20th century stayed true to this minimal approach to human psychology. As Herfeld points out, "(...) introducing the axiomatic method into economics required that human action could be formally represented by a set of axioms that were either validated by empirical observation or of self-evident character." (*op. cit.*, 156)

Axiomatized RCT typically contains two general elements. First, it specifies a set of choices over which it defines a binary preference relation roughly corresponding to the two-place predicate "...equal to, or better than,...". The relation is defined at least as complete and transitive. For some models, it is also defined as continuous, monotone, locally non-satiated, and/or convex through further axioms. A choice rule, or more specifically an algorithm, is then explicated that singles out the set of choices that are equal to, or better than, all choices. Secondly, the axiomatized RCT typically contains a number of theorems deducible from the axioms.

In Herfeld's view, axiomatized RCT is not immediately to be considered a theory in the strict sense. "[It] is a theory only insofar as we accept axiomatic theories as theories; they do not have any immediate resemblance to the common understanding of theories as providing empirical hypotheses. As axiomatic theories are generally not interpreted in the first instance, [axiomatized RCT] has no obvious application; it is a purely 'formal' theory." (*op. cit.*, 158)

At the same time, a purely formal RCT can receive empirical relevance through bridge laws or statements involving empirical entities. Hence, "[RCT] becomes an applicable framework of individual choice only under a specific interpretation." (*op. cit.*, 160) Because of its purely formal nature, axiomatized RCT by itself does not make any empirical predictions. It is a highly abstract, yet very flexible framework

that can in principle be set into relation with a wide range of phenomena. The main criterion for any application is that its internal consistency remains unharmed. "This flexibility makes [axiomatized RCT] generally attractive for scientists. Yet it also becomes immune to critique based upon empirical arguments and no refutable consequences can be derived." (*op. cit.*, 162)

Herfeld concludes from these examples that RCT has taken very different shapes in its historical development. Moreover, it is not clear that it was ever intended by its proponents as a substantial psychological theory. Rather, it has served various goals and theoretical purposes, only some of which were immediately empirical. It follows that various kinds of criticism that have been raised against RCT are in need of re-evaluation.

As already mentioned in section 1, RCT has been often characterized as too far remote from reality to be true. In contrast to the core assumptions of RCT explicated in its axioms, human action is caused by various motives and not exclusively by a desire to maximize utility. Moreover, choices in real-life are highly context-dependent. Thirdly, it is often argued that choices in real life can impossibly be based on a consideration and evaluation of all options. Rather, humans make choices often on the basis of "rules of thumb" or certain learning strategies.

Herfeld counters these criticisms by pointing out that they apply only to those versions of RCT that take on a substantial theory of human psychology. None of the examples from Smith to the axiomitized versions of RCT fits this description. In particular, rational choice analysis as presented by these authors does not form a monolithic paradigm explaining individual choice; it "(...) is not committed to egoism or utility-maximization, two characteristics of human action that [RCT] has often been directly linked with." (*op. cit.*, 265)

In Herfeld's view, RCT is probably best thought of as a methodological tool applicable in the explanation of various phenomena.

> What can be said is that rationality can and often has been specified in different ways; the axiomatic approach simply associates rationality with behavior that complies with certain conditions, whereby the specific conditions and their detailed characteristics are often not agreed upon (...). Furthermore, in its axiomatic version, RCA is not restricted to any particular (economic) interpretation but rather turns out to be a highly flexible set of tools that can be modified for the purpose for which it is applied. (Herfeld 2014, 265)

But Herfeld goes beyond this diagnosis of the actual state of RCT in science by speculating what could and could not be achieved by the transformation of the existing approaches into more detailed theories of economic agents. In her view, a deeper understanding of human choice making will not substantially alter the predictive and explanatory success of economics and social science: "As the characteristics inherent in complex phenomena, such as markets, do not allow for any more detailed explanations and predictions, any attempt to investigate further into

the goals, intentions, and desires of individuals interacting in the market would be a very costly, yet unnecessary, undertaking." (*op. cit.*, 267/68) Consequently, she doubts "(...) that focusing exclusively on the individual and his choices in economic investigation is an undertaking as worthwhile as the recent boom in behavioral economics makes us believe." (*op. cit.*, 272)

An important methodological reason for this irrelevance of a behaviouristic and psychological theory of economic behaviour is that, "(...) on the aggregate level and within market contexts, individual deviations cancel each other out." (*op. cit.*, 272) A second reason is metaphysical. Such a theory with a focus on the psychological states and mechanism of the individual might imply that "(...) social phenomena are reduced to individual agency, which in turn might imply a commitment to psychologism, the doctrine that all social phenomena can be reduced to the mental states of the individual that bring those phenomena about." (*op. cit.*, 275) However, in Herfeld's view the metaphysical status of social phenomena is a very different one: "Social phenomena are not the objects of individual choice but rather emerge from the interactions of individuals." (*op. cit.*, 275)

If social phenomena emerge in an metaphysical sense, and if they have causes and effects by themselves, it would indeed be futile to provide detailed accounts of the elements that social phenomena emerge from. In the next section, I will question this metaphysical status of emergent social phenomena on the basis of a well-known argument from the philosophy of mind and the philosophy of biology. If social phenomena are non-reducible and emergent, it becomes very difficult to see how they could themselves have effects in the real world – a prerequisite for their theoretical relevance.

3 Emergent Social Phenomena as Causes

As mentioned in the previous section, Herfeld (2014) advocates the view that economic modeling and explanation requires only certain minimal assumptions about the psychological life of economic agents. It is sufficient to focus on the aggregate level of social and economic phenomena, mainly because individual differences equal out in large groups. Social phenomena are not the objects of individual choice but rather emerge from the interactions of individuals. Before I argue that this view of the domain of economics, if read literally, is problematic from a metaphysical perspective, I will try to clarify what exactly are the phenomena that economic models aim to explain.

3.1 The Explananda of Economic Models

Introductory textbooks to economics sometimes leave the impression that economic models, such as a market equilibrium model consisting of demand and supply functions, are the primary *explananda* of economic theorizing. Demand and supply

functions are deduced from the preference functions of individual buyers and the production functions of individual firms along with certain further assumptions about perfect information, perfect competition, negligible transaction costs etc. In this sense, they are "explained" on the basis of more primitive assumptions. However, if a claim B is deduced from claim(s) A, then the conditional $A \rightarrow B$ cannot form an empiricially testable generalization. Hence, even if the emerging price in a market is "explained" deductively from the assumptions, this explanation is not yet empirical.

But economics obviously aspires to be an empirical science. So what are the empirical *explananda* of economics then? The obvious answer is that economic models such as market equilibrium models along with their real-world referents are themselves the *explanantia* of certain observable effects: exchanges of goods in the widest sense, or more specifically modifications of the rates of good exchanges under different initial conditions. For example, a price increase in a given market may be explained by a shift in demand resulting from an increase in income. Under this interpretation of the *explananda* of economic models, economic explanation ultimately is a kind of causal explanation. In the mentioned case, it identifies certain causes, such as the existence of a supply/demand constellation (*SDC*) in a given market and a change in a single or more factors underlying the *SDC*, and it singles out as an observable effect a difference in the exchange of goods (*EG*). Hence, in abstract form the explanation has the following conditional form representing a causal (and not simultaneous) relationship: $SDC \rightarrow EG$.

One interesting aspect of the causal dimension of this explanation is that it has both a social and a physical ingredient. An *EG* consists in a change of various social relations of ownership. At the same time, it has a detectable physical dimension: Goods and money physically change position, services are being carried out, representations of virtual money change their values etc. In other words, the *SDC* that, according to Herfeld, emerges from certain properties of consumers and suppliers has at least partially physical effects.

The term "emergence" has been used in various ways. I was made popular by the British Emergentists[1], who believed that certain higher level phenomena were not the mere "resultants" of basic physical forces but spontaneously occurred when matter arranges itself in a certain way. In this sense, occurrences of emergents were thought of as being unanticipated by the laws governing the fundamental physical kinds. The higher-level forces endow the relevant emergent kinds with powers to influence motion.

If social phenomena are thought of as emergent in the same sense, then they are taken to be endowed with causal powers over and above their constituent elements. And it is these powers that bring about the effects in question. Recall that this

[1]Brian McLaughlin (1992) counts Samuel Alexander, Alexander Bain, Charlie Dunbar Broad, Conwy Lloyd Morgan, John Stuart Mill, and George Henry Lewes among the British Emergentists.

picture was one of the reasons for Herfeld for abandoning the aim of developing RCT into a more realistic and appropriate psychological theory. Hence, the idea seems to be that a *SDC* is somehow autonomous and independent from the physical basis, but has its own causal powers as it brings about changes such as an *EG*. Since this causal relation is all that is required for an explanation in economics to get off the ground, it is not necessary to provide a more detailed analysis of the individuals underlying the emergent *SDC*.

From a metaphysical perspective, emergent phenomena and events are problematic in the sense that their causal efficacy is immediately called into question when certain well-established principles of physics are taken into account. A famous line of argument claims that emergent higher-level phenomena are excluded from causing physical phenomena. I will now turn to an explication of this argument in the context of Herfeld's overall picture of economic phenomena.

3.2 Causal Effects of Emergent Events

Suppose that the abstract causal conditional $SDC \rightarrow EG$ is in fact a typical explanation in economics. Suppose further, with Herfeld, that social phenomena such as the *SDC* referred to by the conditional emerge from the interaction of the individuals underlying it without being identical or reducible to the latter. Then we run into a logical inconsistency, given the observation that the physical world is causally complete and physical events are not generally causally overdetermined.

The claim about an emergence and non-reducibility of *SDC* presupposes that social and economic phenomena are not identical to the totality of interactions of economic individuals (premise 1). Suppose further that some social phenomenon such as *SDC* have effects such as *EG*, which are physical to a substantial extent (premise 2). If it is also the case that, as physics tells us, every physical effect has a complete physical cause (premise 3), then a phenomenon such as *SDC* can only be a redundant cause of the *EG*. But we have every reason to believe that systematically redundant causes are no causes at all (premise 4). Hence, we run into an inconsistency. Even though all of premises 1-4 seem plausible and true in isolation, it is clear that they cannot all be true.

An analogous problem is known from the philosophy of mind. The mental analogue of the argument is believed to show that mental phenomena cannot have physical effects unless they are identical to physical phenomena (cf. Kim 1989, 1998, 2003, 2005; for a recent formal analysis of the problem, cf. Harbecke 2013). In these contexts, the problem is usually formulated not for mental "phenomena" but for mental "events", since in most canonical theories of causation only events should be considered as the adequate relata of the causal relation.

For social and economic phenomena, a formulation for events can be developed in a similar way. To do so, we need to define class of economic types (= classes of phenomena) that can be instantiated by certain events, a class of physical types that can be instantiated by events, and a first-order relation of causation:

$\mathfrak{I}(C)$ = $\{(x, y) : x \text{ causes } y\}$
$\mathfrak{I}(\mathcal{E})$ = $\{\phi : \phi \text{ is an economic or social type}\}$
$\mathfrak{I}(\mathcal{P})$ = $\{\phi : \phi \text{ is a physical type}\}$

The word 'causes' should roughly be understood as '...is a sufficient direct singular cause of...'. With these presuppositions, a possible formalization of the assumptions constituting the simplest informal version of the problem of the exclusion of social or economic emergent causes is the following ((P1)–(P4) are considered to be equivalents of the original informal formulations of premises 1-4 mentioned above; (P1)′–(P4)′ are intended as adequate formalizations of the original informal assumptions; (P1)′*–(P4)′* are intended as direct explications of (P1)′–(P4)′ in natural language; ϕ, ψ are type variables, x, y, z are individual variables):

(P1) 'Economic or social events are not identical to physical events'
(P1)′ $\forall x \forall y \forall \phi \forall \psi(\phi x \wedge \mathcal{E}\phi \wedge \psi y \wedge \mathcal{P}\psi \rightarrow x \neq y)$
(P1)′* If any event instantiates any economic or social property and any second event instantiates any physical property, then the two events in question are not identical.

(P2) 'Some economic or social events cause physical events'
(P2)′ $\exists x \exists y \exists \phi \exists \psi(\mathcal{E}\phi \wedge \phi x \wedge \mathcal{P}\psi \wedge \psi y \wedge Cxy)$
(P2)′* There are some events that instantiate a property belonging to the class of economic or social properties that cause certain events that instantiate a property belonging to the class of physical properties.

(P3) 'Every physical event has a physical cause'
(P3)′ $\forall x \forall \psi(\psi x \wedge \mathcal{P}\psi \rightarrow \exists y \exists \gamma(\gamma y \wedge \mathcal{P}\gamma \wedge Cyx))$
(P3)′* If any event instantiates any physical property, then the event in question is caused by some event that instantiates a physical property as well.

(P4) 'Physical events are not causally overdetermined'
(P4)′ $\forall x \forall y(\exists \phi(\phi x \wedge \mathcal{P}\phi) \wedge Cyx \rightarrow \neg \exists z(Czx \wedge z \neq y))$
(P4)′* If any physical event is caused by some event, then there is no further event causing that physical event non-identical to the event causing the physical event.

Premises (P1)′–(P4)′ are provably inconsistent. And since the inconsistency allows a *reductio* proof for all premises, it can serve to reject the least plausible premise out of (P1)′–(P4)′. Since the rejection of the completeness of physics seems improper in the light of a large body of empirical evidence, and since the redundancy of causes seems conceptually implausible, the majority of philosophers has tended towards a rejection either of (P1)′ or (P2)′.

For our current context, this means that either social phenomena cannot serve as explanatory causes after all, or they are not emergent. Both consequences shatter Herferld's argument that had taken the emergence of social phenomena as a metaphysical reason for a methodological focus on aggregate phenomena and away from behaviourist or psychological theories of human economic agents. If the argument from the inconsistency to the rejection of (P1)′ or (P2)′ is sound, Herfeld's argumentative basis loses an important pillar: It is not obvious that economics can only focus on social phenomena understood as aggregates, because it is not clear that these kinds of phenomena by themselves can have the causal effects ascribed to them by economic theory.

4 RCT and Mechanistic Explanation

As mentioned in section 2, Herfeld's second argument against the necessity or fruitfulness to transform RCT into an empirical theory of economic agents is a methodological one. She predicts that, "on the aggregate level and within market contexts, individual deviations cancel each other out." (Herfeld 2014, 272) Among other authors, Herfeld cites Keith Arrow in support of this contention: "In the aggregate, the hypothesis of rational behavior has in general no implications, (...) if agents are different in unspecifiable ways, then (...) very little, if any, inference can be made" (Arrow 1986, 388). In other words, the predictive power of economic models is not necessarily improved if the minimal rational choice principles are replaced by a detailed theory of the individual, its bounded rationality, and the mechanisms underlying its cognitive life.

In my view, it is actually not clear that RCT, if transformed into a more realistic theory of human agents, will not improve the predictive success of our more general economic models. Nevertheless, I will grant Herfeld and Arrow this point for now. My main concern in this context is what I consider to be an important difference between the predictive success and the explanatory success of a theory. Economic models with minimal assumptions about the individual may yield relatively accurate predictions with surprisingly sparse means. However, they do not thereby yield a satisfactory explanation of the phenomenon. Only when the various mechanisms underlying the phenomenon have been successfully isolated, a genuine explanation is forthcoming.

Th position behind my argument is what has become known as the "mechanistic approach" to explanation (cf. Bechtel and Richardson 1993; Machamer et al. 2000; Craver 2002, 2007). The explanatory norm promoted by this account states that explanation in the special sciences essentially requires the identification, location, and analysis of the mechanisms underlying a to-be-explained phenomenon on several levels. The prototypical examples for this kind of explanatory model are research projects in neuroscience, in which a cognitive phenomenon such as spatial

memory has been successfully analysed by the isolation of the neural mechanisms underlying it on several levels.

Machamer et al. define a mechanism as consisting of "...entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer et al. 2000, 3). Bechtel and Abrahamsen extend this definition by describing a mechanism as "...a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena." (Bechtel and Abrahamsen 2005, 423)

As mentioned in section 2, I agree with Herfeld's analysis that the minimal RC assumptions used to ground more general economic models are not yet mechanistic in this sense. Axiomatized RCT remains almost mute about the "entities and activities" that underly a market dynamics and the agents acting within it. RCT so applied is neither a psychological nor a neurocognitive theory of human economic agents. Under sparseness and simplicity considerations this makes a lot of sense, of course. However, there are two reasons why, pace Herfeld's and Arrow's conclusion, ultimately economic theory requires a theory of the individual that goes beyond the minimal assumptions of axiomatized RCT.

The first reason is that a specification of the underlying mechanisms can be of great help to delineate the domain of economics, i.e. the natural class of economic agents. Without such a delineation, it would be quite unclear what economics is in fact about as a science (or as a non-science that is). In particular, it would be a science about "rational stones" as well that "decide" to remain at rest in order to maximize their utility. The specification of the mechanisms allows us to define what it means to be a human rational and economic agent in the first place. In this sense, an extended and adequately transformed RCT helps to define the foundations of economics.

The second reason is that, even when a model generally yields correct predictions about real-world exchange effects, the question still is to what extent the model has actually explained the effect. It seems that, as long as nothing is understood about the psychological and neurocognitive mechanisms, the explained phenomenon remains obscure to a large extent. This is because different explanations are in principle conceivable that yield the same predictions about the exchanges of goods but that invoke entirely ridiculous ontologies. For instance, it may be proposed that the observed economic exchanges are not determined by the self-interested choices of the agents, but by their altruistic desire to satisfy a potential future economist who analyses the exchange phenomena on the basis of RCT. Moreover, in an inverted feelings scenario, it may be the case that agents engage in economic exchange out of a self-aggressive tendency, because, as it happens, whenever they buy or sell a good they feel a terrible grief.

The absurdity of these scenarios only shows that there must be more to RCT than its formal axiomatization if it should form the basis of our explanatory models. Or more generally, to single out the how-actual explanation from a large range of how-

possible explanations, something must be said about the mechanisms underlying the cause phenomenon. The only way to achieve this goal is to increase effort in behavioral economics and neuroeconomics. Both of these fields are still in a fairly young stage as a science, and reliable results may not be immediate. But since behavioral economics and neuroeconomics mainly add new terms and parameters to the preference function and constraint set of RCT[2], this branch of recent science can be viewed as remaining inside the paradigm of RCT as an empirical theory.

Camerer et al. (2005) provide an excellent overview of how neuroeconomics may be able to achieve these goals, whilst the large number of research projects in behavioural economics starting with the well-known works of Tversky and Kahneman (1979; 1981; 1991) and continuing up until the most recent works of Loewenstein et al. (2013) offer good reasons to believe that a comprehensive theory of the psychological mechanisms underlying human economic behaviour will be avaible in the future. Hence, not only ought RCT be transformed into a theory if economic modeling is to maintain explanatory power, it also looks as though RCT is already turned into a comprehensive theory of human economic agency in this very phase of history.

5 Conclusion

In this paper, I have focused on two arguments by Herfeld (2014) against the interpretation of RCT as an empirical theory. I have agreed with Herfeld that in the history of economic science, RCT and RCA in general has mostly been used as a methodological principle containing sparse claims about the psychological mechanisms underlying economic decisions of human agents. However, I have challenged her argument that social and economic phenomena are ontically emergent and that they can thereby enter genuine explanations in disregard of the mechanisms underlying them.

Secondly, I have critically assessed the methodological assumption of the non-necessity to include the psychological mechanisms of the individual into economic models based on the observation that such models, even if predictively adequate, will be very limited in their explanatory power. It follows that RCT will have to be transformed into a psychological and neurocognitive theory of humans in economic circumstances if economics is supposed to maintain and expand its metaphysical plausibility and explanatory success. Hence, in my view, the behaviourist and neuroeconomic extension of RCT is a very valuable enterprise.

Due to limits of space, this paper was unable to offer a deeper analysis of how RCT will have to be extended. Moreover, it was not possible to discuss the possibility that behavioural economics and neuroeconomic research will eventually falsify

[2]I ignore here positions claiming that neuroeconomics will eventually falsify RCT entirely. In my point of view, neuroeconomics and behavioural economics will more likely vindicate and expand RCT.

RCT entirely – a view that I do not share. These questions will have to be left for future research on the status of RCT as a theory.

References

Arrow, K. J. (1986). Rationality of self and others in an economic system. Journal of Business, S385–S399.

Arthur, W. B. (1999). Complexity and the economy. science 284(5411), 107–109.

Bechtel, W. and A. Abrahamsen (2005). Explanation: A mechanist alternative. Studies in History and Philosophy of Biological and Biomedical Sciences 36(2), 421–441.

Bechtel, W. and R. Richardson (1993). Discovering complexity: Decomposition and localization as scientific research strategies. New York: Princeton University Press.

Camerer, C., G. Loewenstein, and D. Prelec (2005). Neuroeconomics: How neuroscience can inform economics. Journal of economic Literature, 9–64.

Charness, G. and M. Sutter (2012). Groups make better self-interested decisions. The Journal of Economic Perspectives, 157–176.

Chen, S.-H. and C.-H. Yeh (2002). On the emergent properties of artificial stock markets: the efficient market hypothesis and the rational expectations hypothesis. Journal of Economic Behavior & Organization 49(2), 217–239.

Craver, C. (2002). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. Philosophy of Science 69(3), 83–97.

Craver, C. (2007). Explaining the brain. New York: Oxford University Press.

Harbecke, J. (2013). On the distinction between cause-cause exclusion and cause-supervenience exclusion. Philosophical Papers 42(2), 209–238.

Herfeld, C. (2014). The Many Faces of Rational Choice Theory. Witten, Germany: Witten/Herdecke University Dissertation Library.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 263–291.

Kim, J. (1989). The myth of nonreductive materialism. Proceedings and Addresses of the American Philosophical Association 63, 31–47.

Kim, J. (1998). Mind in a physical world: An essay on the mind-body problem and mental causation. Cambridge: MIT Press.

Kim, J. (2003). Blocking causal drainage and other maintenance chores with mental causation. Philosophy and Phenomenological Research 67(1), 151–176.

Kim, J. (2005). Physicalism, or something near enough. Princeton: Princeton University Press.

Loewenstein, G., D. A. Asch, and K. G. Volpp (2013). Behavioral economics holds potential to deliver better results for patients, insurers, and employers. Health Affairs 32(7), 1244–1250.

Machamer, P., L. Darden, and C. Craver (2000). Thinking about mechanisms. Philosophy of Science 67(1), 1–25.

McLaughlin, B. (1992). The rise and fall of British emergentism. In A. Beckermann, H. Flohr, and J. Kim (Eds.), Emergence or reduction?: essays on the prospects of nonreductive physicalism, pp. 49–93. Berlin/New York: Walter de Gruyter.

Mill, J. S. (1882/1843). A System of Logic: Ratiocinative and Inductive, Book III (8th ed.). London: Harper and Brothers.

Northcraft, G. B. and M. A. Neale (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. Organizational behavior and human decision processes 39(1), 84–97.

Thaler, R. (1980). Toward a positive theory of consumer choice. Journal of Economic Behavior & Organization 1(1), 39–60.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. Science 211(4481), 453–458.

Tversky, A. and D. Kahneman (1991). Loss aversion in riskless choice: A reference-dependent model. The Quarterly Journal of Economics 106(4), 1039.