Preprint of an article in Economics and Philosophy 32(2): 249-281, 2016

Official Publication: https://www.cambridge.org/core/journals/economics-and-philosophy/article/mentalism-versus-behaviourism-in-economics-a-philosophyofscience-perspective/F4ED161738B25EFB44CC782F702984F1

Mentalism versus behaviourism in economics: a philosophy-of-science perspective

Franz Dietrich & Christian List*

First version 1 April 2012, this version 17 May 2015

Abstract

Behaviourism is the view that preferences, beliefs, and other mental states in social-scientific theories are nothing but constructs re-describing people's behaviour. Mentalism is the view that they capture real phenomena, on a par with the unobservables in science, such as electrons and electromagnetic fields. While behaviourism has gone out of fashion in psychology, it remains influential in economics, especially in 'revealed preference' theory. We defend mentalism in economics, construed as a positive science, and show that it fits best scientific practice. We distinguish mentalism from, and reject, the radical neuroeconomic view that behaviour should be explained in terms of brain processes, as distinct from mental states.

Keywords: Mentalism, behaviourism, revealed preference, decision theory, scientific realism.

1 Introduction

Economic theory seeks to explain the social and economic behaviour of individual human agents (and sometimes that of other agents, such as firms). It usually does so by (i) ascribing, at least in an 'as if' mode, certain mental states, such as beliefs and preferences, to the agents in question and (ii) showing that, under the assumption that those agents are rational, the ascribed mental states lead us to predict, and thereby to

*Contact details: F. Dietrich, Paris School of Economics & CNRS, CES-Centre d'Economie de la Sorbonne, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75647 Paris cedex 13, France; URL: <http://www.franzdietrich.net>. C. List, London School of Economics, Departments of Government and Philosophy, London WC2A 2AE, U.K.; URL: <http://personal.lse.ac.uk/LIST>.

make sense of, the behaviour to be explained.[1] For example, we explain why Franz went to buy a cappuccino at Starbucks, instead of staying in his office, by saying that he had a preference to drink coffee, and a belief that there was coffee at Starbucks but not in the office, so that it was rational for him to take the action. Standard economic theory formalizes this explanation by representing Franz's preferences in terms of an ordering or utility function over outcomes and his beliefs in terms of a subjective probability function over states of the world; an action is then deemed 'rational' (at least in the simplest case) if it maximizes expected utility. Setting aside technicalities, the logic underlying this explanation is very similar to the logic underlying folk-psychological reasoning with its ascription of beliefs and desires to explain behaviour. Economic decision theory can thus be seen as a more sophisticated and scientific reconstruction of folk psychology.[2]

But what is the status of the ascribed mental states, as formalized by objects such as preference orderings, utility functions, and subjective probability functions? Are they

(1) mere *re-descriptions of behavioural patterns* and perhaps *instrumentally useful constructs* for organizing and making sense of empirical regularities,

or are they

(2) *representations of real mental or psychological phenomena*, no less existent in the world than the (also not directly observable) electrons, neutrinos, and electromagnetic fields postulated in the natural sciences?

Roughly, *behaviourism* is the first of these two views, whereas *mentalism* is the second; we will make this more precise later.

Behaviourism used to be the dominant view across the behavioural sciences, including not only economics, where it was pioneered by scholars such as Vilfredo Pareto (1848-1923), Paul Samuelson (1915-2009), and Milton Friedman (1912-2006), but also psychology and linguistics, where it was prominently expressed, for example, by Ivan Pavlov (1849-1936), Leonard Bloomfield (1887-1949), and B. F. Skinner (1904-1990). Bloomfield (quoted in Langendoen 1998) wrote:

[1] We focus on micro-economic theory, which is based on decision- and game-theoretic models. For an overview of theories of choice and rationalization, see Bossert and Suzumura (2010). Non-human agents to which micro-economic theory is sometimes applied include corporate agents (see, e.g., List and Pettit 2011) and non-human animals (see, e.g., a special issue on group decision making in humans and animals, edited, with introduction, by Conradt and List 2009).

[2] Decision theory thereby exemplies a familiar feature of science more generally, which Quine described as commonsense gone self-conscious (Quine 1960). As Lewis (1983: 114) puts it: '[Decision theory] is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference, and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematized.' On decision theory's relation to folk psychology, see also Pettit (1991) and Mongin (2011).

> 'The terminology in which at present we try to speak of human affairs – [...] "consciousness", "mind", "perception", "ideas", and so on – [...] will be discarded [...] Non-linguists [...] forget that a speaker is making noise, and credit him, instead, with the possession of impalpable "ideas". It remains for the linguist to show [...] that the speaker has no "ideas" and that the noise is sufficient.'

In psychology and linguistics, especially since Noam Chomsky's influential critique (1959) of Skinner, behaviourism has long been replaced by some versions of mentalism (e.g., Katz 1964; Fodor 1975), though often under different names, such as 'cognitivism' or 'rationalism'. Many forms of behaviour, it is now widely accepted, are hard to explain unless we pay attention to the cognitive mechanisms underlying them. Chomsky argued that the way in which children learn languages would be difficult to explain if we thought of children as mere stimulus-response systems, without any innate language processing capacities (Pinker 1994; cf. Tomasello 1995). For instance, children's grammatical mistakes do not fit a random trial-and-error pattern; there are certain kinds of combinatorially possible mistakes that children hardly ever make.

In economics, unlike in those other sciences, behaviourism continues to be influential and, in some subfields, even the dominant orthodoxy.[3] The 'revealed preference' paradigm, in many of its forms, is behaviouristic, though there are more and less radical versions of it, as we will explain in detail later. Recently, Faruk Gul and Wolfgang Pesendorfer (2008) have offered a passionate defence of what they call a 'mindless economics', a particularly radical form of behaviourism.

In this paper, we aim to clear up some common confusions about behaviourism and mentalism in economics, situate the debate in the broader context of the philosophy of science, and defend a mentalist approach to economics, which we argue is in line with best scientific practice. We thereby reject Gul and Pesendorfer's case for behaviourism, though we do so from a different, more philosophy-of-science-oriented perspective than earlier responses to them, which often invoked either normative or neuroeconomic arguments (see, e.g., Kőszegi and Rabin 2007; Harrison 2008; and Caplin and Schotter's 2008 collection; some of our criticisms are shared by Hausman 2008). We show that a case for mentalism can be made even if economics is treated as a purely positive science of socio-economic behaviour and not as any sort of normative enterprise. We briefly review

[3] Behaviourism should not be conflated with behaviourial economics, which emphasizes the need for economic models to incorporate insights from psychology (e.g., Camerer *et al.* 2004). Arguably, the name 'behaviourial economics' is slightly misleading; 'psychological economics' would be more appropriate.

some other responses to behaviourism at the end of this paper.[4]

We agree with one methodological concern voiced by Gul and Pesendorfer: the concern about the appropriate *level of explanation* in economics. Here, we think, Gul and Pesendorfer are right in criticizing the attempts of the most radical neuroeconomists to reduce decision theory to neuroscience. But Gul and Pesendorfer draw the wrong conclusions from this. Far from supporting a 'mindless economics', rejecting the attempt to reduce economics to neuroscience is entirely consistent with accepting a mentalist appproach to economic theory. The failure to recognize this point may stem from a failure to distinguish clearly between the notions of 'mind' and 'brain'. The former is a 'higher-level', psychological notion, the latter a 'lower-level', physiological one. The most compelling forms of mentalism entail precisely the view that the study of rationality and action cannot be reduced to the neuro-physiological study of the brain and body.

The paper is structured as follows. In Section 2, we review and contextualize Gul and Pesendorfer's claims. In Section 3, we identify four misconceptions underlying them. In Section 4, we introduce some key concepts from the philosophy of science, which help us clarify the difference between behaviourism and mentalism. In Section 5, we distinguish between two kinds of 'revealed preference' approaches to economic theory – an 'epistemological' and an 'ontological' one – and show that only the more radical and less plausible approach commits us to behaviourism. In Section 6, we state our argument for mentalism more positively. In Section 7, we argue that the difference between mentalism and behaviourism is not just a metaphysical matter but relevant to the practice of economics. In Section 8, we distinguish mentalism from, and argue against, the radical neuroeconomic view that socio-economic behaviour should be explained in terms of agents' brain processes, as distinct from their mental states. In Section 9, we conclude.

[4]The parallels between the mentalism-behaviourism debate in psychology and the one in economics have received relatively little attention in the literature. For a brief historical discussion, unrelated to Gul and Pesendorfer, see Edwards (2008). In a manuscript that came to our attention after we completed the first version of this paper, Okasha (forthcoming) discusses how the contrast between mentalistic and behaviouristic interpretations of decision theory interacts with that between positive and normative uses of decision theory. He argues – in line with our conclusions here – that when decision theory is used as a positive, explanatory theory, there are good reasons to adopt a mentalistic interpretation. More provocatively from our perspective, he argues that when decision theory is used as a normative theory (to prescribe how choices should be made), a behaviouristic interpretation is appropriate. For an earlier discussion of the merits and demerits of different interpretations of decision theory, see Bermudez (2009).

2 The case for mindless economics

Gul and Pesendorfer's paper, 'The case for mindless economics' (2008), provides a useful starting point for our discussion. The paper makes at least three claims about economic science (i.e., the positive rather than normative part of economics):

- The only *evidence* that should be used to test economic theories is evidence about people's choice behaviour.
- The *content* of any economic theory consists solely in its choice-behavioural implications; two theories that are choice-behaviourally equivalent should be seen as equivalent simpliciter.
- Any economic theory should take the *form* of a representation of choice behaviour, and that representation should ideally take the form of attributing to the agents the maximization of some objective function.

The first of these claims concerns the *evidential base* of a theory in economics, the second its *semantic content* or *meaning*, and the third the *methodology of theory construction*. In addition to making these positive claims, Gul and Pesendorfer also express scepticism towards any form of normative economics that goes beyond a very thin kind of 'revealed-preference Paretianism', i.e., the assessment of socio-economic institutions or outcomes in terms of whether they are Pareto efficient relative to people's revealed preferences. For present purposes, however, we set the case of normative economics aside.

In essence, Gul and Pesendorfer hold that (positive) economics should be the science of choice behaviour, and that its evidence base, ontology of the world, and formal structure should focus solely on people's observed or observable choices. Although they do not situate their views in the context of earlier behaviouristic schools of thought in psychology and related disciplines, Gul and Pesendorfer's approach to economics mirrors Pavlov's and Skinner's approaches to psychology and the Vienna Circle's approach to the philosophy of science and language. In fact, each of their central claims corresponds to a different historical variant of behaviourism (using the taxonomy in Graham 2010).

The first claim – about the evidence base of economics – broadly corresponds to 'psychological behaviourism', the view that human (and animal) behaviour should be explained solely on the basis of behavioural evidence, such as evidence about 'external physical stimuli, responses, learning histories, and (for certain types of behavior) reinforcements' (Graham 2010). If anything, the evidence accepted by those earlier psychological behaviourists is *less* restricted than that accepted by Gul and Pesendorfer.

The second claim – about the semantic content or meaning of any theory in economics – corresponds to 'analytical or logical behaviourism', the view associated with the Vienna Circle, Gilbert Ryle (1900-1976), and some of Ludwig Wittgenstein's (1889-1951) work that 'the very idea of a mental state or condition is the idea of a behavioral disposition or family of behavioral tendencies' (Graham 2010). Accordingly, '[w]hen we attribute a belief ... to someone, we are not saying that he or she is in a particular internal state or condition. Instead, we are characterizing the person in terms of what he or she might do in particular situations or environmental interactions' (Graham 2010).

The third claim – about the methodology of theory construction in economics – is analogous to 'methodological behaviourism' in psychology in that it prescribes a focus on the representation of behaviour rather than the modelling of mental states and mental processes in theory construction. Historically, methodological behaviourism, as defended for instance by John Watson (1878-1958), is the view that 'psychology should concern itself with the behavior of organisms' and not 'with mental states or events or with constructing internal information processing accounts of behavior' (Graham 2010). Accordingly, 'reference to mental states, such as an animal's beliefs or desires, adds nothing to what psychology can and should understand about the sources of behavior' (Graham 2010), and so a theory's goal should simply be to represent behavioural patterns. Gul and Pesendorfer strengthen that demand by requiring that this representation take the form of attributing to the agent the maximization of some objective function.

Figure 1 summarizes the parallels between Gul and Pesendorfer's claims and their historical precursors in psychology and related disciplines. Given the extent to which Gul and Pesendorfer's claims mirror (and perhaps reinvent) earlier behaviouristic claims, one might ask whether their views suffer from the same problems that those earlier behaviourisms suffered from and which ultimately led to their demise. In what follows, we draw on insights gained from some of those other cases to see what lessons can be learnt for the case of economics.

3 Four misconceptions

We begin our defence of mentalism by arguing that Gul and Pesendorfer's three positive claims, like their historical precursors, rest on at least four misconceptions, which we will call the 'misconception of a fixed evidence base', the 'evidence/content conflation', the '"unobservable, therefore non-existent" fallacy', and the 'maximization dogma'.

Figure 1: Gul and Pesendorfer's claims and their precursors

of a theory in...	Gul & Pesendorfer's claims economics	Historical precursors psychology
Evidence base	agents' choice behaviour	external physical stimuli, responses, learning histories, reinforcements 'psychological behaviourism' (Pavlov, Skinner)
Semantic content	choice-behavioural implications	behavioural dispositions or behavioural tendencies described by the theory 'analytical / logical behaviourism' (Vienna Circle, Ryle, Wittgenstein)
Methodological form	representation of choice behaviour, in terms of the maximization of an objective function	representation of behaviour, no modelling of internal information processing mechanisms 'methodological behaviourism' (Watson)

3.1 The misconception of a fixed evidence base

In line with psychological behaviourism, Gul and Pesendorfer argue that the only evidence that should be used to test economic theories is evidence about people's choice behaviour. But there is no systematic reason why the evidence base of economics should be restricted in this way. Across the sciences, it is a common phenomenon that our available evidence base occasionally grows. Various things or phenomena that people could not observe in the past, and which earlier generations might have regarded as speculative, have eventually turned out to be observable, through the use of more advanced instruments, more creative experimental designs, and so on.

In physics, entities and phenomena such as the Higgs boson and various elementary particles, forces, and fields seemed at some point to be purely theoretical constructs, but are being increasingly turned into observable entities and phenomena – albeit indirectly observable ones – through the advances in sophistication in our instruments and experimental techniques. The advances in microscopy over the centuries are a perfect illustration of this point. (On the lack of a static distinction between what is observable and what is not, see, e.g., Maxwell 1962 and Shapere 1982.)

In short, the idea that the evidence base of a particular scientific discipline should be fixed once and for all lacks any justification, given the history of science and the experi-

ence of other scientific disciplines. Rather, the evidence base of any science is changeable and dynamic, and there is no reason why economics should be an exception. Accordingly, even if there was a period in the history of economics when people's choice behaviour was the only evidence used to test theories, there is no principled reason why other kinds of evidence – from people's verbal reports and communicative behaviour to physiological and neuroscientific evidence – could not also be relevant. Indeed, psychology has long moved on from its fixation on a behaviouristic evidence base.[5]

3.2 The evidence/content conflation

In line with analytical or logical behaviourism, Gul and Pesendorfer argue that the content of any economic theory consists solely in its choice-behavioural implications; two theories that are choice-behaviourally equivalent should be seen as equivalent simpliciter. But even if the *evidence base* of economic theories were restricted to observable choice behaviour alone – and, as we have seen, there is no principled reason why it should be – it would not follow that the *content* of any economic theory should consist solely in its choice-behavioural implications. Rather, the content of a theory can, and often does, go well beyond its evidence base. To see that this is not just an esoteric possibility, but the norm across many scientific disciplines, consider a few simple examples:

Archaeology and ancient history: The evidence base for theories in these subjects consists of various archaeological objects and artefacts found, for instance, in excavations. But the content of those theories goes well beyond these objects and artefacts. The content, ultimately, is the life, social organization, and culture of the ancient societies in question. The reason why we are interested in old pots, pans, and other broken items is not just that these objects are interesting in their own right, but that they tell us something we cannot directly observe: namely how people lived in the societies under investigation.

Paleobiology: A natural- rather than social-scientific discipline that illustrates our point is paleobiology. Here the evidence base consists of geological findings and fossils, but the aim of the discipline is to answer biological questions about the evolution of life and its underlying molecular-biological mechanisms. Again, the content of the relevant theories goes well beyond the evidence base.

The point of much of science is precisely to make creative use of what is observable

[5]Harsanyi (1955: 317) already observed the following: 'In general, we have two indicators of the utility that other people attach to different situations: their preferences as revealed by their actual choices, and their (verbal or nonverbal) expressions of satisfaction or dissatisfaction in each situation.'

in order to get a better understanding of certain phenomena that are not by themselves observable. Making sense of, and organizing, empirical regularities is just one aim – but not the only aim – of science. By organizing empirical regularities, we often find evidential support for the existence of certain hitherto unobserved aspects of reality.

The disciplines of psychology and linguistics, unlike economics, have fully embraced this point. Linguists in the tradition of Chomsky, for instance, argue that the observable regularities in people's language usage, grammaticality judgments, and initial language acquisition give us evidence for certain underlying cognitive structures that are innate features of human beings, but which are, of course, not themselves observable.

3.3 The 'unobservable, therefore non-existent' fallacy

The next misconception is also relevant to Gul and Pesendorfer's logical or analytic claim that the content of any theory in economics consists solely of its choice-behavioural implications, and that two choice-behaviourally equivalent theories should be seen as equivalent simpliciter. One route by which one might arrive at this claim is the following. Suppose one accepts, as Gul and Pesendorfer do, that observations about people's choice behaviour are the only evidence that we are entitled to use to test our economic theories. And suppose, further, one somehow accepts the principle that *anything that is not observable does not exist*. It then follows that we are not entitled to treat as 'real' or 'existent' any properties or entities in economics that go beyond what we can directly observe. And this, by stipulation, is people's choice behaviour alone.

But even if we were to suspend our criticism of the assumption that only choice behaviour is observable in economics, it should be obvious, as a matter of logic, that, from the fact that a particular entity or phenomenon is not observable, it does not follow that this entity or phenomenon does not exist. And the conclusion that the entity or phenomenon does not exist follows even less from the fact that something is not *currently* observable. Sometimes we can have strong indirect evidence for something, even though it is not directly observable.

Electrons and other elementary particles are not, strictly speaking, directly observable; we can only see their traces, for example, when they travel through a cloud chamber (as water vapour condenses upon the impact of ionizing particles). But few people would seriously doubt their existence.

'Occam's razor' principle tells us not to postulate the existence of any unnecessary entities. So, before we can hypothesize that something exists despite being unobservable, we need to come up with at least some indirect evidence for its existence. But if the best confirmed and most parsimonious theory of a particular phenomenon commits

us to postulating an entity, then it is fully consistent with Occam's razor principle to accept its existence. The key idea behind the principle is that we should not postulate too many entities, but neither should we postulate too few. Consistently with this principle, psychologists and linguists are perfectly prepared to postulate certain mental constructs, and to treat them as real, in order to explain the observable phenomena they are interested in.[6]

3.4 The maximization dogma

Implicitly relying on a particularly strong version of methodological behaviourism, Gul and Pesendorfer suggest that any economic theory should take the form of a representation of choice behaviour, and that this representation should ideally take the form of attributing to economic agents the maximization of some objective function. However, while it may be a useful *starting point* for the explanation of behaviour to search for some objective function that a given agent maximizes, there is no principled reason why our best theories of economic behaviour should *necessarily* be based on the notion of maximization.

Which *form* of a theory best explains human behaviour is a contingent, empirical question, which can be settled only by actual scientific research, not by methodological stipulation. Just as it has turned out to be wrong – given Einstein's general theory of relativity – that space and time must necessarily be Euclidean (as Immanuel Kant, for example, assumed), so there is no *a priori* reason to think that the explanation of social and economic behaviour must necessarily be based on the maximization of a single objective function. For example, an empirically adequate theory might model agents as being governed by a more complex system of constraints.

[6]We here accept that mental states are not directly observable and are similar in status to the unobservable entities and properties in other sciences. Hausman (1998) denies that the mental states posited in economics (e.g., the utility and subjective probability functions) are unobservables on a par with electrons or neutrinos, and argues instead that they should be seen as part of 'commonsense reality', like tables and chairs. This is because the functional role played by utilities and probabilities in economics is 'virtually identical' to that played by desires and beliefs in folk psychology, and the latter are already among our everyday ontological commitments. We accept the analogy between the mental states in economics and those in folk psychology and agree with Hausman that those mental states should be considered real. Yet, we think a further argument is needed to convince the skeptic that mental states in *both* folk psychology *and* economics can be seen as real, *despite their prima-facie unobservability (or at most indirect observability)*. Our argument in this paper is intended to fill this gap. Several contributions to the 'realism-antirealism' debate in economics (as reviewed, e.g., in Hausman 1998) either deny or do not develop the analogy between the mental states posited in economics and the unobservables posited in the natural sciences, and hence that debate is somewhat orthogonal to our concerns here.

Of course, current attempts to explain economic behaviour in a non-maximization-based way remain controversial. Examples are theories of satisficing as introduced by Herbert Simon (1956) and theories of fast and frugal heuristics as proposed by Gerd Gigerenzer and others (e.g., 2000). But the mere fact that these are well-defined and eligible contenders for economic theories illustrates that the maximization of a single objective function is not the only form an economic explanation can take. The reason economists are divided over Simon's and Gigerenzer's theories is *not* that these theories have the wrong form *per se*, but rather that it is unclear whether they offer the best explanations of the empirical phenomena they are intended to explain.

4 A primer in the philosophy of science

We have identified four misconceptions underlying Gul and Pesendorfer's (and no doubt others') arguments for behaviourism in economics. To clarify the distinction between behaviourism and mentalism further, we need to introduce some key concepts from the philosophy of science: (i) theory, (ii) model, (iii) empirical adequacy, (iv) ontological commitment, and (v) underdetermination of theory by evidence.

4.1 What is a theory?

We begin with the orthodox, 'syntactic' definition of a theory and subsequently amend it for our purposes. This definition underlies the so-called 'deductive-nomological account' of scientific explanation, which was famously defended by Karl Popper and Carl Gustav Hempel, among others (for a survey, see Woodward 2009). According to this account, to explain a body of observations is to formulate a hypothesis or set of hypotheses from which the observations can be logically derived. In Quine's formulation (1975), a *theory* is a set of sentences, which is ideally:

(i) closed under implication, and

(ii) expressible as the set of implications of a finite (ideally small) set of basic principles or axioms (called the *theory formulation*), perhaps together with some auxiliary assumptions.

Clause (i) is a technical stipulation which signals that a theory is taken to be logically committed to all its implications. We can thus formally identify a theory with its body of implications. There is no presumption that all those implications are *known*. In practice, the users of a theory will know only some of its implications and will represent the theory in terms of its theory formulation, as specified in clause (ii).

Newtonian physics is a paradigm example of a theory that fits this definition. Here, the theory formulation consists of Newton's three laws of motion and his law of universal gravitation, and the theory itself consists of all the implications of those basic principles. To arrive at a Newtonian theory of a specific physical system, such as the solar system, we must add to the original theory formulation some system-specific auxiliary assumptions and initial conditions. These must specify the relevant physical bodies and their initial configuration: their masses and positions, and the forces acting on them. The theory's predictions about the system's behaviour will then be among the resulting implications. (Formally, this means that we are taking the union of the original theory formulation and the set of auxiliary assumptions and initial conditions and are then defining the amended theory as the set of all sentences that are implied by that union.)

Of course, scientists normally do not think of a theory simply as an uninterpreted set of sentences, but have some interpretation in mind. An *interpreted theory* is a set of sentences that is endowed with an intended interpretation – or with a *set* of admissible interpretations. We now make this more precise. This will allow us to accommodate some insights from so-called 'semantic' conceptions of a theory, which define a theory, not as a set of *sentences* with a certain structure, but as a set of *models* with a certain structure (e.g., van Fraassen 1980).[7]

4.2 What is a model of a theory?

To define the notion of a model, we first introduce a basic notion from formal logic: that of a *semantic interpretation* of the language in which the theory is expressed. This is

- an *assignment of truth-values* to all sentences in that language,

which, in turn, is based on

- a definition of a *domain of objects* (or possibly a family of domains of objects of different types, depending on how many types of objects the theory refers to),
- an *interpretation of all predicates, relations, and functions* that the theory uses, as predicates, relations, and functions over the relevant objects, and
- an *assignment of objects to all constant symbols* used by the theory.

We call a semantic interpretation that renders a given theory formally true – i.e., which assigns the truth-value 'true' to all sentences of the theory – a *model* of that theory. Any

[7] It goes without saying that many subtly different variants of both syntactic and semantic definitions can be given; the details are not the focus of this paper.

consistent theory has at least one model, and typically many. Each model corresponds to one possible (hypothetical) way the world could be – one possible world – *consistent with the theory*. The model's domain of objects (or family of domains) then represents the objects that exist in that hypothetical world, and the predicates, relations, and functions correspond to the properties of, and relations between, those objects.

Now, some models of a theory, while being formally well-defined, do not match the intended interpretation of the theory, perhaps because the objects they specify, and the various relations in which they stand, have nothing to do with the real-world objects and real-world relations that the theory is intended to capture. Among the multiplicity of possible models, we may therefore pick out some that we deem admissible, and set aside others that we deem inadmissible. The latter might be too non-standard or too far-fetched, or simply too distinct from the intended interpretation.

We thus define an *interpreted theory* as a theory that is endowed with a set of admissible models. Formally, an *interpreted theory* can be viewed as a pair consisting of a set T of sentences, as defined earlier, and a set M of interpretations, where each interpretation in M qualifies as a model of T. The first component of this pair corresponds to a theory in the syntactic sense, the second component to a theory in the semantic sense. The notion of an interpreted theory thus combines syntactic and semantic elements.

This notion leaves room for 'thinner' and 'thicker' ways of interpreting a theory. Someone who opts for a very 'thin' interpretation of a given syntactic theory T will deem all or most formally well-defined models of T admissible, thereby specifying a very large set M. By contrast, someone who opts for a very 'thick' interpretation of T may have in mind a single, concrete model which he or she deems uniquely admissible, thereby rendering M maximally small. Typically, the intended interpretation of a theory falls somewhere between these extremes: i.e., several but not all models of T are deemed admissible.

4.3 What does it mean for a theory to be empirically adequate?

Informally, empirical adequacy can be defined as follows:

> 'a theory is empirically adequate exactly if what it says about the observable things and events in this world, is true – exactly if it saves the phenomena.' (van Fraassen 1980: 12; also quoted in Monton and Mohler 2012)

This can be spelt out either in syntactic terms or in semantic ones. From a syntactic perspective, a theory T is *empirically adequate* with respect to some body of observation sentences S – i.e., ideally the set of *all* sentences expressing our empirical observations –

if and only if these sentences are among the theory's implications, formally if and only if T logically entails S. This is the notion of empirical adequacy underlying the deductive-nomological account of scientific explanation. To apply this definition in practice, we must usually include some relevant auxiliary assumptions and/or initial conditions in T. In the example of a Newtonian theory of the solar system, as already noted, we must take T to be the body of implications of Newton's laws, together with our specification of the constituent celestial bodies and their initial conditions. The theory is then (at least approximately) adequate with respect to a set S of observation sentences about the motion of the planets around the sun. It is not adequate, on the other hand, with respect to a body of sentences about the behaviour of objects whose velocity is close to the speed of light, as Einstein famously pointed out.

From a semantic perspective, a theory, now endowed with a set M of admissible models, is *empirically adequate* if it 'has at least one model [in M] that all the actual phenomena fit inside' (van Fraassen 1980: 12). To make this precise, we must designate certain *parts* of any given model as 'empirical substructures', i.e., 'candidates for the direct representation of observable phenomena' (van Fraassen 1980: 64). Empirical adequacy then requires that all observed phenomena – definable as the 'structures which can be described in experimental and measurement reports' – are 'isomorphic to empirical substructures of that model' (van Fraassen 1980: 64).[8]

Empirical adequacy (or at least approximate empirical adequacy, a notion that could be analyzed further) is typically a *minimal* desideratum on a good scientific theory. Importantly, it is not the same as *truth* of the theory. Truth is a more demanding, and more elusive, notion. According to the *correspondence theory*, informally speaking, a theory is true if everything it says matches reality, not only with respect to observable features, but also with respect to the underlying unobservable features. Slightly more formally, this requires that one of the theory's models *in full* (not just certain *empirical substructures* of it) stands in a structure-preserving correspondence (a *homomorphism*) with the relevant structures in the world. In practice, of course, *approximate truth* is the most we can hope to achieve in science (for a review, see Chakravartty 2014).

[8]This is less demanding than requiring that the actual phenomena fit inside *every* model of the theory. In effect, van Fraassen's definition of empirical adequacy requires that the observable phenomena be *compatible* with the theory. The traditional, syntactic definition of empirical adequacy requires that the observable phenomena be *implied* by the theory, in line with the deductive-nomological approach. In principle, each notion of empirical adequacy (the one based on implication and the one based on compatibility) could be defined either syntactically or semantically. Here we set this issue aside.

4.4 What are the ontological commitments of a theory?

We have defined a theory as a set of sentences, endowed with a set of admissible models. By considering all these models, we can ask which kinds of objects, properties, and relations are present in *all* of them. These can be seen as the objects, properties, and relations that the theory is – in a logical sense – *minimally* committed to. We call them the theory's *ontological commitments*.[9] The theory, according to its own admissible models, could not be true without them. It is important to note that the admissible models need not all be isomorphic to one another. Some models may be 'sparser' than others, e.g., have a smaller domain of objects. Hence something will count as an ontological commitment of a given theory only if it is a common presence in *all* admissible models. (We set aside a number of subtleties here, which have been discussed in detail by logicians and model theorists. See, e.g., Bricker 2014.)

This notion of an ontological commitment is very natural. Consider, for example, the theory of arithmetic as defined by the Peano axioms, the fundamental axioms of arithmetic. Let any standard model of these axioms count as admissible; i.e., we adopt the ordinary interpretation of arithmetic. Any standard model has a domain of objects with the formal properties of the natural numbers. Therefore – and as we would intuitively expect – a set of objects playing the role of the natural numbers is among the ontological commitments of Peano arithmetic. In fact, we reach a similar conclusion even if we do not restrict our attention to standard models of arithmetic, but deem non-standard models admissible as well: such models include 'non-standard' numbers, in addition to the 'standard' ones. What *all* models of Peano arithmetic have in common is that they *include* a set of objects with the structure of the natural numers (even though some models – namely non-standard ones – include additional objects).

For another example, consider the standard theory of particle physics, which offers a unified theory of electromagnetic, weak, and strong nuclear interactions, while still leaving out gravity. Just as the natural numbers are a common presence in any model of Peano arithmetic (standard or non-standard), so certain kinds of elementary particles can be found in any non-trivial model of the standard theory of particle physics, such as quarks, leptons (of which electrons are special cases), and different kinds of bosons. Most of these have also been experimentally identified, using instruments such as the Large Hadron Collider at CERN, Switzerland, but at least until recently the Higgs boson remained empirically undiscovered. The theory has always been logically committed to

[9]This definition draws on Quine's account of ontological commitment, although it formally amends it by explicitly quantifying over admissible models. Quine's account is summarized by his famous slogan: '[t]o be is to be the value of a variable' (1948).

its existence, however, since the theory could not be true without it.

The present notion of ontological commitment is central to the so-called *naturalistic attitude* towards ontological questions (Quine 1948; Fine 1984; Musgrave 1989). To figure out what entities, properties, and relations there are in any given area, according to this attitude, we should not engage in armchair metaphysical reasoning, but consult our best scientific theories of that area. Unless we have independent reasons to doubt those theories, we should take their ontological commitments at face value. This means that we should (at least provisionally) accept that the entities, properties, and relations to which the theories are committed correspond to real entities, properties, and relations. This attitude is common in normal scientific practice, outside the philosophy classroom. For example, if our best physical theories tell us that there are quarks, leptons, and bosons, we have every reason to believe in these particles' existence, regardless of their unobservable status. This naturalistic attitude goes against a form of *anti-realism* according to which we may pick and choose among a theory's ontological commitments and take only some at face value (say, those which correspond to observables), while treating others as merely convenient constructs, fictions, or metaphors (say, those which correspond to unobservables). (For a comprehensive discussion and defence of scientific realism, see Psillos 1999.)

4.5 Underdetermination of theory by evidence

In principle but often also in practice, there can be two or more distinct theories that coincide in their observable implications, but which are logically incompatible with respect to some unobservable implications. In such a case, we speak of the *underdetermination of theory by evidence*. This problem was famously discussed by Quine (e.g., 1975; see also List 1999). Why does this problem arise?

In syntactic terms, a theory T logically entails its observable implications, but usually not the other way round. If the theory is empirically adequate, our body of observation sentences – call it S – will be a subset of T, but T will typically go beyond S. In particular, T may also have some unobservable implications. In semantic terms, the observed phenomena usually correspond not to an entire model, but only to certain 'empirical substructures' of such a model. Hence the observed phenomena are logically insufficient to fix an admissible set M of models. Distinct theories, with distinct sets of admissible models, may accommodate the same phenomena.

A simple illustration of this problem in economics is given by the assignment of a von-Neumann-Morgenstern utility function to an agent. For the purposes of the example, let us treat this utility function as our 'theory' of this agent's choice behaviour. As is well

known, there is not just one utility function that fits an agent's choice behaviour, but an infinite number. The function is unique only up to positive affine transformations.[10] Of course, in the present example, nothing much hinges on the properties of the function that are left underdetermined, such as whether the agent's utility in one situation is twice as large as that in another. Indeed, most economists would not consider such statements meaningful; they would regard the question of which *specific* von-Neumann-Morgenstern utility function (as opposed to which equivalence class) is the right one as indeterminate. The underdetermination problem would come to trouble us only if we wanted to use von-Neumann-Morgenstern utilities as a measure of satisfaction or well-being whose levels or units are meaningful (thereby enabling interpersonal comparisons), something many economists would reject. (For a discussion, see List 2003.)

Generally, however, the problem of underdetermination of theory by evidence raises important questions for the status of the unobservable implications of any theory and its ontological commitments. When a theory is underdetermined by the evidence (so that there could be an empirically equivalent theory with different unobservable implications and different ontological commitments), we face the question of whether there is a fact of the matter about the theory's unobservables and what would settle this fact. In particular, we will have to ask whether there is a way of creatively extending the evidence base so as to devise an empirical test to break the underdetermination. If we insisted that there is no fact of the matter, we would have to conclude that we are faced with a case of *indeterminacy*.

The main insight to be gained from this philosophy-of-science primer, for present purposes, is that the question of what our evidence for a particular theory is – or even what the maximal body of evidence could be – is fundamentally distinct from, and not to be confused with, the question of what the theory's ontological commitments are.

5 Two kinds of 'revealed preference' approaches

We are now in a position to distinguish more clearly between two kinds of 'revealed preference' approaches to economic theory, and to see whether they commit us to behaviourism (for classic works on revealed preferences, see Samuelson 1938; Richter 1966; and Sen 1971). One kind of approach is defined in terms of an epistemological thesis, the other in terms of an ontological one. As we will see, only one of the two theses – arguably the less plausible one – is genuinely behaviouristic, while the other is fully

[10] Indeed, setting certain structural constraints aside, *any* positive monotonic transform of such a function may arguably serve as a utility representation (see Weymark 1991).

compatible with mentalism.[11]

An epistemological 'revealed preference' thesis: Our body of evidence for any theory in economics is restricted to agents' choice behaviour.

An ontological 'revealed preference' thesis: The ontological commitments of any theory in economics – or at least those ontological commitments that we are entitled to take at face value – are restricted to agents' choices and choice-behavioural patterns and therefore exclude mental states.

First consider the epistemological thesis. Although we have already disputed that the evidence base should be fixed as stated by that thesis, some economists might still accept it for stipulative reasons: they might stipulate that what demarcates economics from neighbouring disciplines such as psychology is its reliance on choice-behavioural evidence, rather than richer psychological evidence. This justification for the epistemological thesis may seem *ad hoc*, but it is not incoherent.

The ontological thesis, by contrast, is harder to defend. At least in the technical sense introduced above, the ontological commitments of standard (micro-)economic theories include some 'mental-state constructs' such as preference relations, utility functions, or subjective probability functions. Their role is to rationalize, systematize, or explain observed choice behaviour, but they are formally distinct from that behaviour. Relations or functions that formally play the role of preferences, utilities, or subjective probabilities – for short, which *play the preference-or-belief role* – can be found in virtually every theory in (micro-)economics. Any model of such a theory must then have the relevant relations or functions among its structural elements. Recall that, technically, a theory is committed to everything that is present in all its admissible models, which here includes the relations or functions in question. Of course, this does not yet imply that these correspond to anything real in the world.

To uphold the ontological 'revealed-preference' thesis, one would have to insist that, even when a theory is *technically* committed to such 'mental-state constructs' (by which we simply mean: relations or functions playing a preference-or-belief role), we should not take these at face value. Instead, we should interpret them as instrumentally useful modelling devices, which serve to represent the agent's choice behaviour, but do not correspond to anything real. In short, we would have to be anti-realists with respect to all those elements of our economic models that go beyond choice behaviour – the 'unobservables' of our theory. As we will argue in the next section, however, this would

[11] The taxonomy of different kinds of psychological behaviourism in Moore (2001) also suggests that some more modest, methodological (non-radical) forms of behaviourism are compatible with mentalism.

Figure 2: Possible views about 'revealed preferences'

Evidence **Ontological status of ascribed mental states and processes**	**Restricted to choice behaviour**	**Not so restricted; other psychological evidence admissible**
Not real, but mere auxiliary constructs	Radical behaviourism	Uninteresting position (possibly incoherent)
Real, to be taken at face value	Mentalism with narrow evidence base ('epistemic behaviourism')	Mentalism with broad evidence base

conflict with the naturalistic ontological attitude we have introduced.

Another point is worth noting. Even if we accept the epistemological 'revealed preference' thesis, this does not compel us to accept its ontological counterpart. Economic theories can be committed to certain mental-state constructs, even if we use only choice-behavioural evidence to establish their adequacy. This shows that the epistemological 'revealed preference' thesis does not imply the ontological one, and thus that the epistemological thesis is *compatible* with an ontological commitment to mental states. Indeed, the etymology of the term 'revealed preferences' suggests just this: an agent's behaviour 'reveals' – is evidence for – something other than behaviour, namely the agent's mental states – his or her preferences – which cause the behaviour in question.

In sum, behaviourists and mentalists are divided on two questions: first, whether or not the evidence base of economics should be restricted to choice behaviour, and second, whether the relations or functions playing a preference-or-belief role in economic theories should be treated as mere theoretical constructs or as corresponding to real phenomena. Figure 2 shows the different possible views.[12]

[12]The distinction between radical behaviourism and mentalism with a narrow evidence base is similar to Cozic's (2012) distinction between a stronger and a weaker sense in which conventional models of choice can be 'cognitively mute'.

6 An argument for mentalism

Our objections to behaviourism, especially of the radical, ontological sort, should already be clear from our discussion up to this point. We now wish to state our argument for mentalism more positively. Recall that a radical behaviourist holds the view that even if a theory is formally committed to certain relations or functions playing a preference-or-belief role, these are nothing more than theoretical constructs: they may be instrumentally useful for making sense of behavioural regularities, but they should not be seen as corresponding to anything real. As our philosophy-of-science primer should indicate, however, this view goes against the naturalistic attitude towards ontological questions.[13]

According to that attitude, once something – say, an entity or property – is among the ontological commitments of a well-established scientific theory, we ought to take it at face value, unless we have independent reasons to doubt the theory itself. Our acceptance of the theory, naturalistically speaking, commits us to accepting the existence of that entity or property. To ask whether the entity or property 'really' exists, after it has been established as one of the theory's ontological commitments, is to ask one question too many; or alternatively, it is to express doubts about the theory itself. Therefore, when our best theories of economic decision-making are committed to certain mental-state constructs in the technical sense (i.e., relations or functions playing a preference-or-belief role), we should treat these as corresponding to real features of the world, unless we wish to reject those theories themselves.

Could one still object that although the relevant relations or functions correspond to *something real in the world*, they need not correspond to *mental states* of the agents in question? They could simply correspond to certain functions or structures which somehow exist in the world and are systematically related to the agents' choices, but which need not count as mental states. But this misses the point of what mental states are.

Mental states are, at least in part, states that play a certain role for an agent. Beliefs, for example, play the role of representing certain features of the world from the agent's perspective, and preferences play the role of motivating the agent's actions (see, e.g., List and Pettit 2011: ch. 1). *Functionalism* is the view that what makes something a mental state is simply that it plays the relevant role. On a less demanding view, playing

[13] Although we follow a Quine-inspired approach to ontology, our mentalistic conclusion departs from Quine's own philosophy, since Quine was critical of mental constructs in science. Arguably, Quine's own behaviouristic approach has been superseded, and a broadly Quinean naturalism lends itself well to a defence of a naturalistic form of realism about mental states. On Quine's philosophy, see Hylton (2007).

this role is *one* of the features of a mental state, though perhaps not the only one.[14] So, playing a preference-or-belief role need not be *constitutive* of a mental state; it can merely be *indicative*. This less demanding view is sufficient for our purposes. As noted in the last section, our best economic theories of individual decision making are certainly committed to certain relations or functions that *formally* play preference-or-belief roles. From the standpoint of a naturalistic ontological attitude, these can then be taken to correspond to *real* relations or functions playing such roles. And since playing those roles is indicative of mental states, we should, at least provisionally, treat the 'mental-state relations or functions' posited by our theories as corresponding to real mental states of the agents in question.[15]

It is important to note that taking a theory's mental-state constructs to correspond to real mental states is fully consistent with acknowledging that they depict these mental states in a simplified or idealized way (and perhaps only capture 'structural' rather than 'intrinsic' features of those mental states). This is no different than acknowledging that a physical theory's depiction of a planet or a volcano greatly simplifies or idealizes the details of the real planet or volcano it refers to. Recall that the relationship between the world and its representation by a scientific theory is at best a homomorphic one, which preserves certain structural features but which still abstracts away from many substantive details.

We can now summarize our argument:[16]

Premise 1: Some mental-state constructs – by which we mean: relations or functions playing the role of preferences or beliefs – are technically among the ontological commitments of our best theories of economic decision making.

[14] One question we cannot address here concerns the relationship between the *functional role* of mental states (e.g., the representational or motivational roles of beliefs and preferences) and their *phenomenal or conscious character* (roughly, what it subjectively feels like to have such mental states). Some philosophers, especially dualists, hold that certain mental states have phenomenal (conscious) aspects *above and beyond* their functional aspects. Others consider these phenomenal aspects a *by-product* of their functional aspects. Yet others deny the existence of phenomenal aspects altogether. For an overview of the debate, see Chalmers (2010). Here, we focus only on the functional aspects of mental states, which are most relevant to economics as a positive science. For a discussion of functionalism, see Block (1980).

[15] A particularly strong form of mentalism – a radical limiting case – might be derived from a 'direct realist' attitude towards our theory of an agent's decision making. According to such an attitude, which goes beyond what we are arguing for in this paper, a given scientific theory has only one intended model: namely the world itself or the fragment of the world in which we are interested. When we speak about mental states in our theory, we would then be referring directly to the agent's mental states in the world.

[16] Some of the philosophical ideas underlying this naturalistic argument are developed in List (2014), specifically in the defence of intentional agency as a real (albeit higher-level) phenomenon.

Premise 2: In any normal science, the criterion for whether a theoretically postulated entity, property, or relation should be treated, at least provisionally, as corresponding to a real entity, property, or relation is whether it is among the ontological commitments of our best theory or theories in the relevant area – assuming we have no independent reasons to doubt those theories themselves.

Premise 3: Economics is a normal science.

Conclusion: The mental-state constructs to which our best economic theories are committed should be treated, at least provisionally, as corresponding to real phenomena – assuming we have no independent reasons to doubt those theories themselves.

The argument is clearly valid (i.e., the premises logically entail the conclusion). Whether the argument is also sound depends on whether the premises are all true. Given the nature of practically all our current (micro-)economic theories, ranging from classical rational choice theory to more recent psychologically oriented theories (e.g., Camerer *et al.* 2004), Premise 1 is true in light of our technical definition of an ontological commitment, as already discussed. Premise 2 is also true, since it states a basic principle underlying standard scientific practice: the naturalistic ontological attitude. Premise 3 is a claim that critics of economics might wish to challenge, but scientifically minded economists are unlikely to object to it.

Consequently, the only way to avoid the mentalistic conclusion would be to insist on having doubts about our economic theories themselves, despite their status as our best scientific theories in the relevant area. But those asserting such doubts would then have to explain what evidence underpins them. We suspect that few economists would wish to make their argument against mentalism dependent on a rejection of the adequacy of our best economic theories themselves. We conclude that just as we have strong *prima-facie* reasons to accept the reality of quarks, leptons, and bosons in particle physics, so we have strong *prima-facie* reasons to accept the reality of mental states in economics.

It is worth clarifying how this conclusion differs from the view held by radical behaviourists. We are not suggesting that radical behaviourists such as Gul and Pesendorfer will deny the reality of mental states when they take off their 'hats' as professional economists and adopt a commonsense view of the world, for instance while interacting with other people in their day-to-day lives. What they are committed to denying is that mental states should be part of the scientific ontology of economics.

7 Does the difference between mentalism and behaviourism matter?

One might think that the difference between mentalism and behaviourism is a purely metaphysical matter, which is irrelevant to the practice of economics itself. But this impression is misleading. That the difference matters also in practice can be seen by revisiting the empirical underdetermination problem, the problem that there can exist two or more distinct theories that are empirically equivalent but logically incompatible.

First consider the idealized limiting case of no underdetermination. Take a simple choice problem without risk or uncertainty, where an agent has perfectly well-behaved choice dispositions over some options, satisfying all the standard rationality conditions. The agent's choice behaviour – formally represented by a choice function – can then be uniquely rationalized by a preference ordering over the given options (e.g., Sen 1971; Bossert and Suzumura 2010) (note that the conditions for achieving such a unique rationalization are demanding). Although this rationalization involves, technically speaking, a mental-state construct – namely a binary relation that formally plays a preference role – preference orderings and choice functions stand in a one-to-one correspondence in this case. As long as rationalization of choices is required to take the form of ascribing to the agent a weak ordering, there is no underdetermination of preferences by choice behaviour here: there exists one and only one preference ordering that entails the given choice behaviour.[17] Consequently, there are no logical implications of the mental state ascription that go beyond what is already encoded in the choice function itself, and no issues of indeterminacy arise: there are behaviourally observable facts about everything the theory says. Hence, one might regard the question of what the ontological status of the ascribed preferences is, over and above the agent's choice behaviour, as primarily metaphysical.

Now, however, consider a less idealized case. A much-discussed example is due to Amartya Sen (1993).

The polite dinner-party guest: Given a choice between a large, a medium-sized, and a small apple, a dinner-party guest (who at home would choose larger over smaller apples) chooses the medium-sized apple – in Sen's story, for politeness. If the large apple is no longer available while the medium-sized and small ones still are, the guest chooses the small apple – again for politeness.

[17] If we lift the requirement that rationalization take the form of the ascription of a weak ordering to the agent, and allow other forms of rationalization (e.g., in terms of other mathematical structures), then the underdetermination problem can arise even in the present case of choice without risk or uncertainty.

The agent's choice function violates contraction consistency and cannot be rationalized by a preference ordering over apples. But it would be a bad explanation to suggest that the agent is irrational; this explanation would violate the *principle of charity* in interpretation (see, e.g., Davidson 1973). Rather, the agent is motivated by considerations over and above the sizes of the apples. However, *if* choice behaviour is the only evidence we can go by – for instance, we cannot ask the agent what the reasons for his or her choices were – then we face an underdetermination problem. Several distinct hypotheses will entail the same choice behaviour, ranging from the hypothesis that the agent has complicated (perhaps 'non-consequentialist') preferences over 'extended alternatives' (object-context pairs) to the hypothesis that he or she is governed by various norms of politeness, approval- or esteem-seeking, or other social constraints (e.g., Bhattacharyya *et al.* 2011; Bossert and Suzumura 2009; Suzumura and Xu 2001; Brennan and Pettit 2005). The agent's choice behaviour alone is insufficient to distinguish between these and other rival explanations.

Does this mean that there is no fact of the matter as to what the correct explanation is? Both our psychological understanding and the practices of other cognitive and behavioural sciences suggest that there *can* be a real difference between different rival explanations, despite their choice-behavioural equivalence. First of all, they attribute different internal cognitive mechanisms to the agent and different first-person experiences, which would lead us to predict different introspective reports from him or her, if we could elicit a truthful response. Second, different explanations may also have different repercussions further down the line. Only some but not all explanations may cohere with our explanations of other related phenomena or allow us to predict the agent's choices in other, hitherto unobserved situations. Thus a suitably broadened evidence base may allow us to distinguish between different explanations of an agent's choices. In short, the availability of different choice-behaviourally equivalent explanations does not imply that there is no fact of the matter as to what the correct explanation is.

Wakker (2010: 3; drawing on Harré 1970) distinguishes between *paramorphic* and *homeomorphic* models of decision making. A *paramorphic* model 'describes the empirical phenomena of interest correctly, but the processes underlying the empirical phenomena are not matched by processes in the model'. A *homeomorphic* model, by contrast, has the property that 'not only its empirical phenomena match reality, but also its underlying processes do so'. In outlining a research programme for decision theory, he suggests that we should aim to arrive at homeomorphic models and that this is what prospect theory seeks to do: 'Not only [should] the decisions predicted by the model match the decisions observed, but we also want the theoretical parameters in the model to have plausible

psychological interpretations'.[18]

Sharing this goal, several recent works in decision theory emphasize the importance of 'reasons for choice' or 'psychological states' over and above the choice behaviour induced by them. Some of these works explicitly employ mentalist terminology, such as 'epistemic states', 'knowledge', and 'beliefs' in epistemic game theory (e.g., Aumann and Brandenburger 1995); 'belief-dependent emotions' in psychological games (Geanakoplos and Pearce 1989); 'emotions' such as 'anger' or 'fear' (Elster 1998; Loewenstein 2000); 'thinking' and 'feeling' (Romer 2000); 'intrinsic' and 'extrinsic motivations', 'ego boosting' and 'ego bashing' (Bénabou and Tirole 2003); 'rationales' (Manzini and Mariotti 2007; Cherepanov *et al.* 2013); 'moods' and 'mindsets' (Manzini and Mariotti 2012); 'motivating reasons' and 'weighing of reasons' (Dietrich and List 2013a; 2013b; 2015); 'experiences' (Dietrich 2012); and 'the minds of checklist users' (Mandler *et al.* 2012).

In sum, since choice behaviour routinely underdetermines its theoretical explanation, good scientific practice requires us to consider all the different rival explanations and then creatively to identify an enriched evidence base, and more advanced empirical designs, to determine which explanation is most adequate – in particular, which is most homeomorphic and not merely paramorphic. The enriched base might include novel choice situations, psychological data over and above choice behaviour, verbal reports, related social phenomena, and occasionally (for plausibility checks) even introspection. However, even if we fail to find a purely empirical criterion for picking out a unique correct theory, Occam's razor principle would tell us to choose a theory which is ontologically not too rich, but also not too sparse, to explain our observations parsimoniously.

8 Can economics be reduced to neuroscience?

Some neuroscientists hope to dispense with traditional psychological theories by explaining psychological phenomena in terms of neurophysiological processes in the brain (for a critical discussion, see Bennett *et al.* 2007). Similarly, some of the most radical neuroeconomists hope to dispense with traditional economic theories by explaining

[18] Wakker (2010: 3) also stresses that the evidence base and domain of economic explanations should not be considered fixed: '[Milton] Friedman's arguments in favor of paramorphic models are legitimate if all that is desired is to explain and predict a prespecified and limited domain of phenomena. It is, however, usually desirable if concepts are broadly applicable, also for future and as yet unforeseen developments in research. Homeomorphic models are best suited for this purpose. In recent years, economics has been opening up to introspective and neuro-imaging data. It is to be expected that the concepts of prospect theory, in view of their sound psychological basis, will be well suited for such future developments and for connections with such domains of research.'

economic behaviour in terms of agents' brain processes (for discussion, see Camerer *et al.* 2005). At first sight, one might think that scientific progress is inexorably headed in this direction, and many advances in science seem to confirm this picture. Throughout the sciences, we are gaining a better understanding of the 'lower-level' mechanisms underlying 'higher-level' phenomena, for instance the biochemical mechanisms underlying the functioning of cells, the cellular mechanisms underlying the life of organisms, and the individual-level mechanisms underlying larger social processes. The search for micro-foundations of macroscopic phenomena, with a view to replacing less fundamental theories (at a higher level) with more fundamental ones (at a lower level), is *en vogue*.

Yet, there is a common misconception underlying many of these attempts at theory reduction. The misconception can be termed the 'supervenience implies explanatory reducibility' fallacy. To explain this fallacy, let us consider a familiar argument for theory reduction. Its (correct) premise is that the world is fundamentally made up of elementary particles, atoms, and molecules, which stand in various physical and chemical relations to each other and whose interaction underlies all more complex phenomena, including the functioning and behaviour of organisms. More formally:

The supervenience thesis: The totality of 'lower-level', physical facts about the world determines all 'higher-level' facts, such as facts about organisms and their behaviour.

It is then argued that, because everything in the world 'supervenes' on the physical, the best explanation of any phenomenon must also be a physical one.

The explanatory-reducibility thesis: Any phenomenon in the world can and should ideally be explained in terms of underlying physical mechanisms. Any non-physical explanations – such as psychological or social explanations – are at best provisional and reflect a lack of understanding of underlying mechanisms.

The claim that psychology can be reduced to neuroscience is sometimes defended in just this way. Psychological phenomena are surely the result of underlying neurophysiological brain processes, and 'so', the reasoning goes, our most fundamental explanations of them should also be given at the neurophysiological level.

But does supervenience really imply explanatory reducibility? A large body of work in philosophy challenges this view, beginning with Jerry Fodor's (1974) and Hilary Putnam's (1975) classic arguments that the sheer combinatorial complexity of the relationship between the physical states of a person's brain and the psychological states of his or her mind rules out the effective reducibility of psychological 'natural kinds' (which are

the relata of regularities that we are interested in) to purely neurophysiological ones.[19] What makes 'higher-level' mental states, such as beliefs and desires, more explanatorily useful than 'lower-level' patterns of neural activity is precisely that they abstract away from a large number of physical details that are irrelevant, even detrimental, to the explanatory purposes at hand. Supervenience, in short, does not imply explanatory reducibility (for a defence of this anti-reductionistic view, see List and Menzies 2009; for a structurally similar discussion in relation to the individualism-holism debate in the philosophy of social science, see List and Spiekermann 2013).

Consider, for example, how you would explain a cat's appearance in the kitchen when the owner is preparing some food. You could either try (and in reality fail) to understand the cat's neurophysiological processes which begin with (i) some sensory stimuli, then (ii) trigger some complicated neural responses, and finally (iii) activate the cat's muscles so as to put it on a trajectory towards the kitchen. Or you could ascribe to the cat (i) the belief that there is food available in the kitchen, and (ii) the desire to eat, so that (iii) it is rational for the cat to go to the kitchen. It should be evident that the second explantion is both simpler and more illuminating, offering much greater predictive power. The belief-desire explanation can easily be adjusted, for example, if conditions change. If you give the cat some visible or smellable evidence that food will be available in the living room rather than the kitchen, you can predict that it will update its beliefs and go to the living room instead. By contrast, one cannot even begin to imagine the informational overload that would be involved in adjusting the neurophysiological explanation to accommodate this change.

Good explanations – ones that are parsimonious and predictively successful – should identify the most functionally relevant regularities, while leaving out extraneous details. Functionally relevant regularities, in turn, need not be found at the most fine-grained level of description. It is an empirical question at which level of description any given system exhibits the most tractable regularities. There is no reason, for example, why a good theory of forest ecology should refer to quantum-mechanical effects inside the individual atoms in each tree. Similarly, if you want to explain why Microsoft Windows crashes if you install a particular software package, you should first look at possible programme errors or incorrect system parameters before trying to give a detailed account of the flow of individual electrons in the computer's micro-processor and memory chips.

[19] More technically, the inverse image, with respect to the relevant supervenience function from physical brain states to psychological states, of any set of psychological states forming a 'natural kind' at the psychological level need not be a set of physical brain states forming a 'natural kind' at the physical level. For a critical discussion of multiple-realizability arguments against reductionism, see Sober (1999).

As Daniel Dennett (1987) has argued, we explain the behaviour of certain organisms in terms of their mental states and not in terms of complicated physical processes – thereby taking an 'intentional' rather than 'physical stance' towards those organisms – precisely because this is the level of explanation most suited for the explanatory purpose at hand. A doctor who wishes to treat a brain hemorrhage or a tumor may well take a physical stance towards the patient, at least during the medical intervention, but it is far from clear how much economists can gain from trying to explain socio-economic behaviour by looking at people's brains, rather than interpreting their minds.

All of this is consistent, of course, with the idea of enriching the evidence base of economics when this helps us to distinguish between different rival theories, and this could certainly include some neuroeconomic evidence. But it should be clear that neither the focus on behaviour alone, nor the focus on brain physiology alone, can deliver satisfactory economic theories. (For a nuanced defence of a 'non-revolutionary' approach to neuroeconomics, see Craver and Alexandrova 2008.)

9 Concluding remarks

We have offered an argument for mentalism, and against behaviourism, in economics. We have not only responded to the central epistemological and ontological claims made by behaviourists, but also distinguished mentalism from the more radical neuroeconomic view that economic behaviour should be explained in terms of agents' brain processes, as distinct from their mental states. Gul and Pesendorfer (2008) seem to miss this distinction, frequently equating the mental with the neural and treating what might charitably be understood as a case for a 'brainless economics' (i.e., for an economic science separate from, and not reducible to, neuroscience) as a case for a 'mindless economics' instead (i.e., for an economic science free from mental-state ascriptions).

Our present critique of behaviourism differs from other, more familiar critiques of behaviourism and 'revealed preference' approaches (see, among others, Sen 1977; Hausman 2000; Kőszegi and Rabin 2007; Craver and Alexandrova 2008; and Bermudez 2009). The behaviouristic account of preferences (and other mental states such as beliefs) is often criticized for what it fails to deliver: (i) it fails to say anything about human psychology, motivation, and the mechanisms of decision making, all of which it leaves out of the picture; (ii) it fails to provide adequate foundations for normative economics, as it gives at most an impoverished account of human well-being, says nothing about fundamental desires and interests, and renders interpersonal comparisons of utility impossible (all of which may matter for policy-making); (iii) it fails to 'explain' behaviour in a non-

circular way, since behaviour is 'explained' by preferences (or other attributes) that are in turn defined in terms of behaviour; and (iv) it 'insulate[s] economics from possible contributions from other disciplines', by rendering findings from fields such as psychology irrelevant to economics by methodological design (Craver and Alexandrova 2008: 396).

While such arguments are important and can be (indeed have been) made, we have taken a different approach here. Those earlier arguments construe economics as a discipline that should deliver more than a theory of (economic) choice (providing an account of, e.g., some psychological features of agents, normatively relevant features beyond revealed preferences, or non-circular explanations of choice). This premise is not shared by those economists who, when pressed, are prepared to 'define' (micro-)economics as a science of choice behaviour. Such a science should be as free as possible from normative assumptions and play no 'therapeutic' role, in Gul and Pesendorfer's terms. Critics of behaviourism who presuppose a broader definition of the discipline have little hope of convincing those who endorse the narrow, choice-centered definition. By contrast, our critique should convince also those who view economics as a science of choice behaviour alone, devoid of any further psychological or normative goals. Our naturalistic argument shows that even if one is not interested in mental states *as such*, one's theory of choice may well have to take them on board. A theory *of choice* may have to be a theory *about more than choice*.

Acknowledgements

This paper was presented at several occasions, including the LSE Choice Group workshop on 'Rationalizability and Choice', July 2011, the D-TEA workshop, Paris, July 2012, the EIPE seminar, Rotterdam, September 2012, the Paris Seminar on Economics and Philosophy, July 2013, and the Rational Choice & Philosophy Conference, Vanderbilt University, May 2014. We are grateful to the participants and especially Nick Baigent, Walter Bossert, Richard Bradley, Mikaël Cozic, Eddie Dekel, Ido Erev, Itzhak Gilboa, Conrad Heilmann, Johannes Himmelreich, Thomas Westhoff Holaday, Marco Mariotti, Friederike Mengel, Samir Okasha, Clemens Puppe, Larry Samuelson, David Schmeidler, Asli Selim, Daniel Stoljar, Kotaro Suzumura, Peter Wakker, and John Weymark for comments and discussion. We also thank two anonymous reviewers for their detailed and helpful suggestions. Franz Dietrich was supported by a Ludwig Lachmann Fellowship at the LSE and the French Agence Nationale de la Recherche (ANR-12-INEG-0006-01). Christian List was supported by a Leverhulme Major Research Fellowship and the Franco-Swedish Program in Philosophy and Economics.

References

Aumann, R. and A. Brandenburger. 1995. Epistemic Conditions for Nash Equilibrium. *Econometrica* 63(5): 1161–1180.

Bénabou, R. and J. Tirole. 2003. Intrinsic and Extrinsic Motivation. *Review of Economic Studies* 70: 489–520.

Bennett, M., D. Dennett, P. Hacker and J. Searle. 2007. *Neuroscience and Philosophy.* New York: Columbia University Press.

Bermudez, J. L. 2009. *Decision Theory and Rationality.* Oxford: Oxford University Press.

Bhattacharyya, A., P. K. Pattanaik and Y. Xu. 2011. Choice, Internal Consistency and Rationality. *Economics and Philosophy* 27(2): 123–149.

Block, N. 1980. What is Functionalism? In *Readings in Philosophy of Psychology*, vol. 1, ed. N. Block, 171–184. Cambridge/MA: Harvard University Press.

Bossert, W. and K. Suzumura. 2009. External Norms and Rationality of Choice. *Economics and Philosophy* 25: 139-152.

Bossert, W. and K. Suzumura. 2010. *Consistency, Choice, and Rationality.* Cambridge/MA: Harvard University Press.

Brennan, G. and P. Pettit. 2005. *The Economy of Esteem.* Oxford: Oxford University Press.

Bricker, P. 2014. Ontological Commitment. *Stanford Encyclopedia of Philosophy (Winter 2014 Edition)*, E. N. Zalta (ed.), URL: <http://plato.stanford.edu/archives/win2014/entries/ontological-commitment/>.

Camerer, C. F., G. Loewenstein and D. Prelec. 2005. Neuroeconomics: How Neuroscience Can Inform Economics. *Journal of Economic Literature* 43(1): 9–64.

Camerer, C. F., G. Loewenstein and M. Rabin. 2004. *Advances in Behavioral Economics.* Princeton: Princeton University Press.

Caplin, A. and A. Schotter, eds. 2008. *The Foundations of Positive and Normative Economics.* Oxford / New York: Oxford University Press.

Chakravartty, A. 2014. Scientific Realism. *Stanford Encyclopedia of Philosophy (Spring 2014 Edition)*, E. N. Zalta (ed.), URL: <http://plato.stanford.edu/archives/spr2014/entries/scientific-realism/>.

Chalmers, D. 2010. *The Character of Consciousness*. Oxford: Oxford University Press.

Cherepanov, V., T. Feddersen and A. Sandroni. 2013. Rationalization. *Theoretical Economics* 8: 775–800.

Chomsky, N. 1959. A Review of B. F. Skinner's *Verbal Behavior*. *Language* 35(1): 26–58.

Conradt, L. and C. List, eds. 2009. Group decision making in humans and animals. Theme issue of *Philosophical Tansactions of the Royal Society* B 364: 717–852.

Cozic, M. 2012. Economie 'sans esprit' et données cognitives. Working paper, Institut d'Histoire et de Philosophie des Sciences et des Techniques, Paris.

Craver, C. F. and A. Alexandrova. 2008. No Revolution Necessary: Neural Mechanisms for Economics. *Economics and Philosophy* 24(3): 381–406.

Davidson, D. 1973. Radical Interpretation. *Dialectica* 27(3–4): 313–328.

Dennett, D. 1987. *The Intentional Stance*. Cambridge/MA: MIT Press.

Dietrich, F. 2012. Modelling change in individual characteristics: an axiomatic framework. *Games and Economic Behavior* 76: 471–494.

Dietrich, F. and C. List. 2013a. A reason-based theory of rational choice. *Nous* 47(1): 104–134.

Dietrich, F. and C. List. 2013b. Where do preferences come from? *International Journal of Game Theory* 42(3): 613–637.

Dietrich, F. and C. List. 2015. Reason-based choice and context-dependence: An explanatory framework. Working paper, LSE.

Edwards, J. M. 2008. On Behaviorism, Introspection, Psychology and Economics. Working paper, University of Paris 1, Panthéon-Sorbonne.

Elster, J. 1998. Emotions and Economic Theory. *Journal of Economic Literature* 36(1): 47–74.

Fine, A. 1984. The Natural Ontological Attitude. In *Philosophy of Science*, ed. J. Leplin, 261–277. Berkeley: University of California Press.

Fodor, J. A. 1974. Special sciences (or: The disunity of science as a working hypothesis. *Synthese* 28(2): 97–115.

Fodor, J. A. 1975. *The Language of Thought*. Cambridge/MA: Harvard University Press.

Geanakoplos, J. and D. Pearce. 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1(1): 60–79.

Gigerenzer, G., P. M. Todd and the ABC Research Group. 2000. *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Graham, G. 2010. Behaviorism. *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, E. N. Zalta (ed.), URL: <http://plato.stanford.edu/archives/fall2010/entries/behaviorism/>.

Gul, F. and W. Pesendorfer. 2008. The Case for Mindless Economics. In *The Foundations of Positive and Normative Economics*, ed. A. Caplin and A. Schotter, 3–39. Oxford / New York: Oxford University Press.

Harré, R. 1970. *The principles of scientific thinking*. London: Macmillan.

Harrison, G. W. 2008. Neuroeconomics: A Critical Reconsideration. *Economics and Philosophy* 24(3): 303–344.

Harsanyi, J. C. 1955. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy* 63(4): 309–321.

Hausman, D. 1998. Problems with Realism in Economics. *Economics and Philosophy* 14: 185–213.

Hausman, D. 2000. Revealed Preference, Belief, and Game Theory. *Economics and Philosophy* 16: 99–115.

Hausman, D. 2008. Mindless or Mindful Economics: A Methodological Evaluation. In *The Foundations of Positive and Normative Economics*, ed. A. Caplin and A. Schotter, 125–151. Oxford / New York: Oxford University Press.

Hylton, P. 2007. *Quine*. Abingdon: Routledge.

Katz, J. J. 1964. Mentalism in Linguistics. *Language* 40(2): 124–137.

Kőszegi, B. and M. Rabin. 2007. Mistakes in Choice-Based Welfare Analysis. *American Economic Review* 97(2): 477–481.

Langendoen, D. T. 1998. Bloomfield. In *The MIT Encyclopedia of Cognitive Science*, ed. R. A. Wilson and F. C. Keil, 90–91. Cambridge/MA: MIT Press.

Lewis, D. 1983. *Philosophical Papers Volume I*. Oxford: Oxford University Press.

List, C. 1999. Craig's Theorem and the Empirical Underdetermination Thesis Reassessed. *Disputatio* 7: 28–39.

List, C. 2003. Are Interpersonal Comparisons of Utility Indeterminate? *Erkenntnis* 58: 229–260.

List, C. and P. Menzies. 2009. Non-reductive physicalism and the limits of the exclusion principle. *Journal of Philosophy* 106(9): 475–502.

List, C. 2014. Free will, determinism, and the possibility of doing otherwise. *Nous* 48(1): 156–178.

List, C. and P. Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

List, C. and K. Spiekermann. 2013. Methodological Individualism and Holism in Political Science: A Reconciliation. *American Political Science Review* 107(4): 629–643.

Loewenstein, G. 2000. Emotions in Economic Theory and Economic Behavior. *American Economic Review* 90(2): 426–432.

Mandler, M., P. Manzini and M. Mariotti. 2012. A million answers to twenty questions: Choosing by checklist. *Journal of Economic Theory* 147: 71–92.

Manzini, P. and M. Mariotti. 2007. Sequentially Rationalizable Choice', *American Economic Review* 97(5): 1824–1839.

Manzini, P. and M. Mariotti. 2012. Moody choice. Working paper, University of St Andrews.

Maxwell, G. 1962. On the Ontological Status of Theoretical Entities. In *Scientific Explanation, Space, and Time; Minnesota Studies in the Philosophy of Science*, Volume III, ed. H. Feigl and G. Maxwell, 3–27. Minneapolis: University of Minnesota Press.

Mongin, P. 2011. La théorie de la décision et la psychologie du sens commun. *Social Science Information* 50(3–4): 351–374.

Monton, B. and C. Mohler. 2014. Constructive Empiricism. *Stanford Encyclopedia of Philosophy (Spring 2014 Edition)*, E. N. Zalta (ed.), URL: <http://plato.stanford.edu/archives/spr2014/entries/constructive-empiricism/>.

Moore, J. 2001. On Distinguishing Methodological from Radical Behaviorism. *European Journal of Behavior Analysis* 2: 221–244.

Musgrave, A. 1989. Noa's Ark – Fine for Realism. *The Philosophical Quarterly* 39(157): 383–398.

Okasha, S. Forthcoming. On the Interpretation of Decision Theory. Manuscript, Bristol University.

Pettit, P. 1991. Decision Theory and Folk Psychology. In *Foundations of Decision Theory: Issues and Advances*, ed. M. Bacharach and S. Hurley, 147–175. Oxford: Blackwell.

Pinker, S. 1994. *The Language Instinct: How the Mind Creates Language*. New York: William Morrow.

Psillos, S. 1999. *Scientific Realism: How Science Tracks Truth*. Abingdon: Routledge.

Putnam, H. 1975. Philosophy and our mental life. In *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Quine, W. V. 1948. On What There Is. *Review of Metaphysics* 2: 21–38.

Quine, W. V. 1960. *Word and Object*. Cambridge/MA: MIT Press.

Quine, W. V. 1975. On Empirically Equivalent Systems of the World. *Erkenntnis* 9: 313–328.

Richter, M. K. 1966. Revealed Preference Theory. *Econometrica* 34(3): 635–645.

Romer, P. M. 2000. Thinking and Feeling. *American Economic Review* 90(2): 439–443.

Samuelson, P. 1938. A Note on the Pure Theory of Consumer's Behaviour. *Economica* (New Series) 5(17): 61–71.

Sen, A. K. 1971. Choice Functions and Revealed Preference. *Review of Economic Studies* 38(3): 307–317.

Sen, A. K. 1977. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs* 6(4): 317–344.

Sen, A. K. 1993. Internal Consistency of Choice. *Econometrica* 61(3): 495–521.

Shapere, D. 1982. The Concept of Observation in Science and Philosophy. *Philosophy of Science* 49(4): 485–525.

Simon, H. A. 1956. Rational choice and the structure of the environment. *Psychological Review* 63(2): 129–138.

Sober, E. 1999. The Multiple Realizability Argument against Reductionism. *Philosophy of Science* 66(4): 542–564.

Suzumura, K. and Y. Xu. 2001. Characterizations of Consequentialism and Nonconsequentialism. *Journal of Economic Theory* 101(2): 423–436.

Tomasello, M. 1995. Language is Not an Instinct. *Cognitive Development* 10: 131–156.

van Fraassen, B. C. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Wakker, P. 2010. *Prospect Theory: For Risk and Ambiguity*. Cambridge: Cambridge University Press.

Weymark, J. A. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-being*, ed. J. Elster and J. E. Roemer, 255–320. Cambridge: Cambridge University Press.

Woodward, J. 2011. Scientific Explanation. *Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*, E. N. Zalta (ed.), URL: <http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/>.