The Specter of Revealed Preference Theory

Lukas Beck (L.Beck1@lse.ac.uk)

London School of Economics and Political Science (LSE)

Abstract

My aim in this paper is to argue that the recent philosophical defenses of revealed preference theory do not withstand scrutiny. Towards this aim, I will first outline revealed preference theory. I will then briefly present the two most common arguments that the received view offers against it. Afterwards, I will outline three argumentative strategies for rehabilitating revealed preference theory, and successively rebut each of them.

Keywords: preference, philosophy of economics, revealed preference theory, behaviorism

1. Introduction

A specter is haunting economics—the specter of revealed preference theory. Many philosophers of old have entered into an alliance to exorcise this specter; Sen (1977) and Hausman (2012), Dietrich and List (2016), and Guala (2012; 2019). In the face of the trenchant critique it has faced, the longevity of revealed preference theory is quite surprising. While it still holds considerable power among economists, in recent years also philosophers have begun to offer novel arguments in its defense (e.g., Vredenburgh 2020; Clarke 2020; Thoma 2021a; 2021b). At its core, revealed preference theory can be stated as the view that preferences are just patterns in choice-behavior.

My aim in this paper is to argue against the revival of revealed preference theory. Towards this end, I will first outline the different facets of revealed preference theory (Section 2). I will then briefly present the two most common arguments that philosophers of economics have offered against it. In particular, I will look at the argument from belief and the argument from causality (Section 3). Afterward, I will outline three argumentative strategies for rehabilitating revealed preference theory. The first strategy is to argue that RPT is the only game in town because of the advantages it realizes and the problems of its rivals (Section 4). The second strategy aims to dispel the argument from belief (Section 5). The third strategy is to dispel the argument from causality. For each of these strategies, I will offer reasons for resisting them.

2. Revealed Preference Theory

Revealed preference theory (RPT) has become a buzzword in philosophical debates about economics. It, therefore, makes sense to distinguish between different meanings of revealed preference theory in order to make clear what this paper is concerned with: (i) RPT as a research program, (ii) RPT as an

epistemological thesis, (iii) RPT as an ontological thesis. To see more clearly how the different guises of RPT relate to each other, I will now outline each of them in turn.

As a research program, RPT started with Paul Samuelson's 1938 paper that aimed at "eliminating the last vestiges of utility analysis" from economics. At the time, mainly due to the work of Pareto (1927/2014) and Hicks and Allen (1934), ordinal utility theory (OUT)—which takes preferences to be the primary concept for analyzing choice-behavior—had become the main modeling framework in economics. Yet, Samuelson, who took himself to be concerned with operationalizing economics, wanted to eliminate any reference to supposedly unobservable entities like preferences or utility from economics. Arguably, the primary motivation for this was that utility was widely considered to be unmeasurable. This, according to Samuelson, implied that economics was built on shaky empirical foundations (see Hands 2013; Moscati 2018). Therefore, he wanted to derive the restrictions OUT puts on demand-functions from observable restrictions on choice-behavior alone. While Samuelson could not derive all of OUT's restrictions in that way, Houthakker (1950) later demonstrated that all of them could be derived by putting further restrictions on observable choice-behavior.[1] Houthakker, thereby, showed that OUT and RPT were equivalent in their implications for demand analysis. This means that for any agent whose choice-behavior satisfies the restrictions of Samuelson and Houthakker—and thereby implies certain restrictions on the agent's demand-function—there exists a utility-function from which we can derive a demand-function that satisfies the same restrictions. The next significant contribution to this literature was Afriat (1967), which went beyond mere existence proofs and demonstrated how to construct utility-functions from finite sets of choice data. Nowadays, economists build on these results to project consistent patterns in choice-behavior via a utility-function from observed contexts to new contexts. Bernheim and Rangel (2008: 159) provide a very clear statement of this approach:

> Usually, choice data are not available for all elements of X, but rather for elements of some restricted set $X^D \subset X$. The objective of positive economic analysis is to extend the choice correspondence C from observations on X^D to the entire set X. This task is usually accomplished by defining a parameterized set of utility functions (preferences) defined over X, estimating the utility parameters with choice data for the opportunity

[1] In particular, Samuelson's so-called weak axiom of revealed preference theory only allowed us to derive (i) Negative Substitution Terms and (ii) Negative Semi-Definiteness of the Slutsky Substitution Matrix. Yet, demand-functions generated from an ordinal utility-function also imply (iii) Symmetry of the Slutsky Substitution Matrix. Samuelson (1938: 68) was not concerned with this and claimed that he "cannot see that it is really an important problem, particularly if we are willing to dispense with the utility concept, and its vestigial remnants.'' Yet, when Houthakker showed that his strong axiom of revealed preference also allows us to derive (iii), Samuelson (1950) declared that he was always interested in finding a theory with the same full empirical implications as ordinal utility theory, but with more solid epistemological foundations in observable phenomena (see Hands 2014).

sets in X^D, and using these estimated utility functions to infer choice for opportunity sets in X/X^D (by maximizing that function for each $x \in X/X^D$).

So, RPT as a research program set out to eliminate any reference to utility and preference from economics and ended up developing tools for deriving utility-functions from choice data (see Wong 2006; Hands 2014).

As an epistemological thesis, RPT is the claim that the evidential base of economics is restricted to choice data. That is, the only evidence that economics should consider is evidence concerning people's choices. To see why this epistemological thesis is usually seen as problematic, consider that economic models refer, at least at face value, to non-choice entities. That is, they seem to represent beliefs via credence-functions and preferences via utility-functions. Some have even argued that economics provides us with a regimentation of folk or common-sense psychological explanations that trade in terms of beliefs and desires (see Hausman 2012). However, evidence for such mental states does not seem to exhaust itself in evidence about choice-behavior. Psychological and neuroscientific evidence, as well as people's verbal reports, so the thought goes, can also help us to determine people's preferences and beliefs. In fact, one of the motivations behind behavioral economics (Rabin 2002; Angner 2019) and neuroeconomics (Camerer 2008) is precisely to extend the evidential base of economics to psychological and neuroscientific evidence. Hence, RPT, understood as an epistemological thesis, appears problematic. Yet, RPT, as an ontological thesis, is trying to come to the rescue.

According to the ontological version of RPT, preference relations in economics are nothing more than concise descriptions of patterns in choice-behavior. For example, in economics, ascribing a preference X over Y to some agent is just to say that she would choose X in some circumstances, where Y is an available alternative (Clarke 2016). As an ontological thesis, RPT is a version of behaviorism. In particular, it is a version of analytic behaviorism, according to which ascribing a mental state to an agent is just to say that the agent will behave in specific ways in certain circumstances (see Graham 2010). Yet, it is a narrow version of behaviorism because it does not hold that analytic behaviorism is true about mental states in general. All it states is that preferences in economics refer to patterns in choice-behavior. Hence, there is nothing incoherent if an (ontological) revealed preference theorist tells a waiter that she prefers red wine over white wine and, thereby, takes herself to reveal her mental representation of the value of these two options if she speaks in her role as a private citizen. As long as she is aware that her technical use of the term on which she relies in her job is different from her everyday use, there is no problem. In other words, narrow behaviorists *about economics* make no ontological commitments apart from the fact that a particular choice occurs in certain circumstances

when using preferences *in the context of an economic model.* If ontological RPT were correct, it could be argued that it is still possible to restrict the evidential base of economics to just choice-behavior (i.e., save RPT as epistemological thesis), because, contrary to initial appearance, economic models do not a provide us with a regimentation of folk or common-sense psychological explanations that refer to mental states.

From here on, we will be concerned with assessing RPT as an ontological thesis as it is the most contested version of RPT and the claim that several philosophers have recently attempted to revive. Moreover, it is also often viewed as supporting RPT as an epistemological thesis, which, in turn, favors an economics that heavily relies on the tools developed by RPT as a research program. To avoid verbosity, I will use the term 'behaviorism' to refer to RPT as an ontological thesis. Bear in mind that, in what follows, we are only concerned with a narrow version of behaviorism only meant to apply to economics.

3. The Received View

Despite the fact that many economists endorse behaviorism in economics (for the most infamous endorsement, see Gul & Pesendorfer 2008), most *philosophers* of economics tended to reject it based on the following two arguments (see Beck 2022).[2] The first argument concerns the importance of beliefs for inferring preferences from choices (3.1.). The second argument concerns the role of preferences in causal explanations (3.2.).[3]

3.1. The Argument from Belief

The first argument points out that in order to (correctly) infer preferences from choice-behavior, we also need information on the agent's beliefs. Consequently, it makes no sense to identify preference relations with patterns in choice. If preferences were identical with choice, information on an agent's choice alone should be sufficient to inform us about the agent's preferences (see Rosenberg 1992; Hausman 2000). For example, my choice for the vegetarian option on the menu is insufficiently captured by my preference for vegetarian options unless we combine it with my belief that the option I chose (let us say falafel) is indeed a vegetarian option. If I had the belief that the option is not vegetarian, my choice would have been different even though my preferences stayed the same. To put

[2] Saying that behaviorism holds considerable sway among economists is not to deny that there is also an anti-behaviorist camp in economics. In fact, as one anonymous reviewer correctly noted, much of the current interest in the topic can be explained as a reaction to the recent inclusion of behavioral economics into the mainstream of the discipline—and many prominent behavioral economists are critics of behaviorism. Nevertheless, what I would like to address here is the support that behaviorism has lately gained from philosophers.

[3] There is also a third argument stating that economics cannot support the assumption that preferences are stable and context-independent if we adopt behaviorism (see Bruni & Sugden 2007). As the recent literature has not directly targeted this argument, I will bracket it here.

it the other way around, if we were to ascribe beliefs about the available options to the agent that do not match how the agent conceptualizes these choices, our preference ascriptions will go wrong. Consider, for instance, that we are observing Sam, who eats a full spoon of wasabi paste. Yet, Sam mistakenly believed that the wasabi cream was avocado cream (see Thoma 2021a). Hence, if we ascribed a preference for eating a spoon of wasabi to Sam, predictions about her future behavior based on this preference would likely go wrong. Because of these considerations, preferences in economics cannot be identical with choice-behavior. Information on an agent's preferences will not provide us with all the information on an agent's choices, and we cannot infer preference without having knowledge of the agent's beliefs.

To add to this, consider that beliefs and preferences are often *both* given explicit representations in economic models in the form of utility- and credence-functions. In such cases, just considering the representation of an agent's preferences (i.e., the utility-function) will not provide us with all the information about an agent's choices that is contained in the model. If preferences in economics were identical with choices, utility-functions should also directly represent the agent's choices. So, the idea that choices are identical with preference seems to conflict with those economic models that explicitly include credence-functions.[4]

3.2. The Argument from Causality

The second argument states that preferences being identical with choice-behavior would prevent them from figuring into causal explanations of choice-behavior because an event cannot cause itself. Yet, economists often refer to preferences when attempting to give causal explanations of choice-behavior (see, for example, Guala 2012). For example, the behaviorist cannot explain why I chose the vegetarian option with recourse to my preference for the vegetarian option. After all, under behaviorism, this would amount to saying that I chose the vegetarian option because I chose the vegetarian option (see Vredenburgh 2020 for a particularly clear presentation of this argument). As a result, the behavioristic conception fails to capture specific uses of the term preference in economics because it cannot make sense of the fact that economists use preferences in causal explanations of choice-behavior.

4. The Motivation for Behaviorism

[4] Models in consumer choice-theory—the birthplace of RPT—usually do not explicitly represent people's beliefs, as they assume that agents have all the relevant information about the available option. In other words, all their beliefs are assumed to be correct.

In light of the two outlined arguments and their widespread acceptance by philosophers, it may seem surprising that many economists (at least at face value) are still attracted to behaviorism. One can, of course, offer a sociological explanation by pointing out that philosophy is not taken very seriously by most economists. However, I do not want to go down this road here. Instead, in this section, I look at the motivations for holding onto behaviorisms in economics. I will try to assess whether these factors indeed make behaviorism an attractive position for economists.

Towards this aim, I first outline and evaluate some of the alleged advantages of behaviorism that have been explicated in the recent literature (4.1.). I shall argue that these alleged advantages do not single out behaviorism as compared to alternative conceptions of preferences. I then look at a recent argument by Thoma (2021a) that tries to establish that there are no good alternatives to behaviorism in economics, as there are severe obstacles to identifying preferences with (folk psychological) mental states (4.2.). I will point out that Thoma's argument fails as a defense of behaviorism because it rests on problematic assumptions about preferences as functionally individuated (mental) states.

4.1. Behaviorism and Its Alleged Advantages

We have already encountered the biggest sources of motivations behind revealed preference theory, namely that the main base of evidence that economists have available is choice data (Thoma 2021b; see also Dietrich & List 2016). In light of this, the behaviorist can argue that identifying preferences with anything but choice-behavior will introduce an additional inferential step into the process of assigning preferences to an agent. Yet, if preferences are choice-behavior, no further inference is required, and our ascription of preferences is, at least on a conceptual level, more secure.

However, the fact that inferences become more secure from a conceptual viewpoint is—on its own—of little help if those inferences do not help achieve our epistemic goals. For instance, if we are interested in inferring preference in order to provide causal explanations of choice-behavior, but preferences understood as patterns in choice cannot deliver such explanations, a more secure inference to preferences does not get us very far. Similarly, if we are interested in making projections to unobserved choice-behavior based on preferences, but the correctness of those projections crucially depends on having the correct information about people's beliefs, a more secure inference to preferences would be of little worth if behaviorism would also prevent us from acquiring correct information about beliefs. All of this is just to say that whether one sees the fact that behaviorism eliminates an inferential step in economics as an advantage crucially depends on what one thinks of the two arguments outlined in the previous sections. Consequently, I hold that the possibility of drawing more direct inferences to preferences by itself offers little motivation for behaviorism.

Yet, we also find other alleged advantages of behaviorism in the literature. Thoma (2021b) mentions three of them, which she calls advantages from black-boxing. In particular, the first alleged advantage is allowing economists to retain a more transparent disciplinary boundary to neuroscience, psychology, and related disciplines. The second alleged advantage is that behaviorism looks attractive in the face of skepticism about specific psychological processes. The third alleged advantage is that behaviorism can help economic theory to achieve greater generality because, under behaviorism, the applicability of economic theory does not depend on agents' decisions resulting from specific processes.

However, I now argue that these alleged advantages are also possessed by other conceptions of preference. One example of such a conception is functionalism, a version of the position that construes preferences in economics as mental states (see Dietrich & List 2016). Functionalism holds that (intentional) mental states should not be individuated by their intrinsic properties but by the causal role they play in a certain system of inputs, outputs, and other mental states. Any set of entities that occupies these causal roles is, therefore, said to realize a mental state. Importantly, if different sets of entities occupy the same causal role, they all realize the same type of mental state. As long as we are sure that a particular part of a system or agent fulfills the causal role that defines a certain mental state, we do not require any detailed knowledge of this part's internal composition, location, or organization. Hence, a functionalist conception of preferences also allows for a considerable amount of black-boxing.

Consequently, I hold that functionalism can also deliver the alleged advantages Thoma attributes to behaviorism. For instance, Dietrich and List (2016) note that understanding preferences as functionally individuated mental states does not imply that economics is reducible to psychology or neuroscience. However, they note that not being reducible to a certain discipline does not imply that those disciplines cannot aid economics by providing it with further evidence (see also Craver & Alexandrova 2008 for a nuanced defense of a mechanistic, non-reductive approach to neuroeconomics). Nevertheless, functionalism allows us to leave the disciplinary borders between economics and its neighboring disciplines in place (for whatever this is worth).

Turning to Thoma's second advantage from black-boxing, functionalism allows us to abstract away from the precise assumptions of specific psychological theories of decision making. Hence, like behaviorism, it looks attractive in the face of skepticism about specific psychological processes. Finally, functionalism can also secure the third alleged advantage mentioned by Thoma because different processes can realize the same mental states if they all occupy the relevant causal role. Hence, similar to behaviorism, functionalism can secure the generality of choice-theory as the ascription of preferences under functionalism does not depend on the presence of specific decision processes.

To sum up, the alleged advantages identified in the recent literature do not force us to adopt behaviorism. Whether we should see the fact that behaviorism would eliminate an inferential step as advantageous depends on whether we accept that behaviorism allows us to realize epistemic aims like explanations and predictions. Moreover, the other three advantages identified by Thoma can be realized by other conceptions of preferences than behaviorism. I have illustrated this with functionalism. Hence, we have so far seen little that would motivate a rehabilitation of behaviorism. Next, I will assess another argument by Thoma that aims at establishing that there are no good alternatives to behaviorism because of the fine-grainedness of preferences in choice-theoretic models.

4.2. Against Alternatives to Behaviorism

Thoma (2021a: 914) provides us with the example of an agent who "drank coffee on the first day because she prefers tasting coffee to tasting tea, and she knew that she would taste coffee if she drank coffee." Yet, the agent "drank tea on the second day, because she wanted to keep her nerves down for her important meeting, and believed the tea would keep her less nervous than the coffee."

She argues that to capture this type of behavior with a consistent preference relation, we have to describe the objects of choice at a very fine-grained level that includes descriptions of the various combinations of circumstances that can affect the agent's choice (e.g., tea *before an important meeting*, coffee *on a relaxed day*). She points out that preferences and desires in folk psychological or mentalistic explanations are usually very coarse-grained. For instance, we may say that an agent prefers tasting coffee over tasting tea in a folk psychological explanation to indicate that the agent has a desire for coffee. Yet, we do not refer to such fine-grained descriptions of options in folk psychological explanations as we require them for choice-theory.

Now, Thoma takes folk psychology to fix the causal roles that mental states are meant to occupy according to functionalist accounts of mental states. In particular, in line with arguments by Dietrich and List (2016), she takes the causal roles that preferences would need to occupy to count as mental states to be fixed by our folk psychological concept of desires. Yet, as her example is meant to show, preferences in economics do not seem to play desire-like roles because their content is far more fine-grained than the content of (folk psychological) desires. She, therefore, concludes that functionalism about preferences is unpersuasive as a preference in economics "does not, like the functionalist claims, play a desire-like role in folk psychological explanation" (Thoma 2021a: 913). Hence, according to Thoma, functionalism is not an option in economics.

She, thus, considers more "substantive kinds of mentalism" like the judgmentalism of Bradley (2017) and Hausman's (2012) view that preferences are total subjective comparative evaluations.[5] Thoma treats these views as basically stating that preferences are consciously accessible, fine-grained attitudes that we form during deliberation out of our more coarse-grained, folk psychological desires. The problem then, so Thoma argues, is that there is very little evidence that we usually form such attitudes before making decisions. This is meant to be supported by evidence generated by our best psychological theories of decision making as well as the evidence we gain from introspection.

To illustrate this position, Thoma (2021a: 923) tells us that, from an introspective perspective, an agent who does not have an important meeting on a particular day may simply conclude that she drank coffee on that day because she likes the taste of coffee. There is no need for her to form a conscious attitude of the kind: '*I prefer coffee on a day where I have no important meetings and tea on a day where I have an important meeting*.'

That is, the agent can choose coffee without forming this attitude even if the agent would have chosen differently on a day of an important meeting. Yet, if the agent had an important meeting, something would have triggered her desire to keep her nerves so that she would have become aware of it. She would then have based her decision (at least partly) on this desire and opted for tea. In light of cases like this one, Thoma also holds that more substantive forms of mentalism like judgmentalism are unsuccessful because they would restrict choice-theory's applications to the rare cases in which we really form such fine-grained attitudes. In sum, Thoma concludes that the idea that preferences in economics are mental states cannot be sustained because of the fine-grainedness of preferences in choice-theory. This, according to Thoma, supports behaviorism.

Against this, I hold that her argument rests on an artificial distinction between functionalist views and what Thoma calls substantive views of mentalism. Thoma argues that the latter usually assumes that mental states are consciously accessible. Yet, she acknowledges that conscious accessibility is not necessary for mental states under the functionalist picture. Hence, she dismisses functionalism by saying that preferences in economics do not play the roles desires play in folk psychology and then deals with the substantive view by saying that there is no evidence that we form consciously accessible attitudes of the kind that substantive versions of mentalism would identify with preferences.

Yet, it remains unclear why functionalism forces us to submit to the view that preferences play the role of folk psychological desires. So-called psycho-functionalist (e.g., Fodor 1968; Quilty-Dunn &

[5] According to Hausman, a total evaluation is a ranking to the effect that some option X is better than another option Y that takes all relevant considerations into account. Those evaluations are comparative because having a preference requires you to *weigh* some option X with another option Y. They are subjective in the sense that they are states that (in combination with other mental states like beliefs, and constraints) cause our choice-behavior (see Angner 2018).

Mandelbaum 2018) usually identify mental states as entities defined by their role in a cognitive psychological theory. The functional role that a mental state needs to occupy to count as a particular type of mental state would, according to this view, be identified by our best psychological theories.

In light of this, why would it be illegitimate to say that preferences in economics are identified by the role they play in economic theory? An investigation of this role may lead us to the conclusion that preferences in economics occupy causal roles similar to the causal processes that make the agent in Thoma's example choose coffee on a day at which she has no important meeting and tea on the day at which she has an important meeting. Identifying this whole mechanism with the agent's preference would preempt Thoma's objection from fine-grainedness precisely because the mechanism is sensitive to all the various factors that influence our choices. That only parts of this mechanism are consciously accessible would not matter for the functionalist. Finally, that not all parts of this mechanism are triggered on a day on which the agent does not have an important meeting would also not matter for the functionalist.

Of course, Thoma could insist that if we are free to define the causal role of preferences how we like, functionalism ceases to be a plausible position. Therefore, we have to import the relevant causal roles from folk psychology. However, it is pretty standard for functionalists to claim that folk psychology can at best be a start for fixing the causal roles mental states are meant to occupy and that our (best) scientific theories also have an essential role to play here (for the textbook version of this argument see Braddon-Mitchell & Jackson 2007).[6] Moreover, it is also unclear why desires, as Thoma understands them, are the relevant folk psychological posit for economics. As Thoma notes, many authors have already argued that there is a relevant difference between folk psychological desires and choice-theoretic preferences (e.g., Hausman 2012; Bradley 2017). So why not have the causal role of preferences under functionalism fixed by the role occupied by what Hausman's would call total subjective comparative evaluations?[7] Furthermore, Clarke (2020) points out that folk psychological concepts like desiring are usually too vaguely specified to serve as the basis for identifying causal roles. For instance, what are the causal roles that folk psychology would associate with the desire that 'the government' should do more about climate change?

Hence, I hold that Thoma's argument against understanding preferences as functionally individuated states is unsuccessful. Her claim that preferences in economics would need to play desire-like roles in order to count as functionally individuated mental states is undermotivated. Instead, preferences

[6] For a more general framework for evaluating the legitimacy of functionally individuated states and processes in model-based social sciences, see Beck and Grayot (2021).

[7] In fact, I hold that one can plausibly interpret Hausman's (2012) view on preferences as stating that a preference is simply a process that occupies the function of selecting between two options by taking all relevant factors into account.

could, for instance, be seen as playing total-subjective-comparative-evaluation-like roles. This suffices to classify them as mental states, according to a version of functionalism that takes economic theory as its basis. This, in turn, seriously undermines Thoma's argument in favor of behaviorism.

However, Thoma (2021a: 926) appears to anticipate this objection. She argues that there is a sense in which we can identify the whole psychological process she describes in her coffee example as the preference of the agent even if not all parts of it are consciously accessible. Yet, she thinks that this would lead to no explanatory gain at all and that, therefore, "parsimony seems to demand we do away with mentalistic preference and stick to a behavioral interpretation of preference." The idea seems to be that if we identify preferences with the more complex causal process in her example, all we can say in the end is that an agent chooses coffee because she prefers coffee. Hence, Thoma seems to think that no explanatory value is gained from identifying the whole processes with a preference. Therefore, we should stick with the more parsimonious behaviorism, which refrains from claims about the underlying mechanisms.

However, to have any force against functionalism specifically, such an argument from parsimony must assume that an explanation that cites a consciously accessible, fine-grained substantive mental state (whatever this may be) offers explanatory power. On top of this, we need to assume that an appeal to the whole process that could be viewed as occupying the preference role offers no explanatory power. Otherwise, such an argument would amount to outright denying the explanatory power of preferences even in cases where agents would deliberately form a consciously accessible, fine-grained substantive mental state. For instance, one would need to assume that an explanation of the kind: *'an agent chooses coffee because she deliberately formed a total subjective comparative evaluation for coffee*' offers explanatory value and an appeal to the functionally individuated causal processes in Thoma's example offers none. In other words, Thoma's argument from parsimony cannot serve as an argument against a particular conception of preference if it licenses skepticism about the explanatory value of the preference concept in general.

To highlight that Thoma's position is indeed threatened to collapse into such skepticism, consider also that explanations referring to preferences as fine-grained attitudes under substantive versions of mentalism appear to be rather shallow if we just look at one particular choice of a single agent. To see this, consider Hausman's (2012) view that preferences are total subjective comparative evaluations. What is crucial for my purposes here is that under this view, there is no counter-preferential choice as preferences are rankings to the effect that some option X is better than another option Y that take all relevant considerations into account.

Hence, for Hausman (2012), it appears to be almost true by definition (as far as economics is concerned) that an agent chooses based on total subjective comparative evaluations.[8] Yet, if this is the case, how could it ever be more than a very shallow explanation to state that an agent chooses based on her preferences? Rather than speaking against the explanatory power of preferences in general, I take this to indicate that there is something wrong with the argument that, construed as more complex functionally individuated states, preferences offer no explanatory power.

Instead, what we need to consider is that economic models usually do not attempt to offer explanations of single choices of a single agent. In this regard, I take it that economic explanations will always lose out against more detailed psychological explanations for single choices of a single agent even if the agent in question deliberately forms a preference understood as a consciously accessible, fine-grained attitude. In other words, explaining single choices of single agents is not the strong suit of economics. Rather, economics seems to do better in explaining patterns in large sets of choices.[9]

Economic models try to come up with utility- (and credence-)functions that allow us to derive implications for sets of choices.[10] I shall now argue that utility-functions (as representations of sets of preferences) in combination with credence-functions (as representations of beliefs) that account for large sets of choices already put considerable restrictions on the agent's decision mechanism. As a result, the explanations we gain for those sets of choices are far from shallow. Let me make this point again with the help of a functionalist conception of preference. Under this reading, what we commit ourselves to when assigning preferences and credences is that the agent's underlying decision mechanisms are such that they can generate the whole profile of choices implied by these preferences and credences. While we abstract away from details and internal organization, we nonetheless put considerable restrictions on the agent's decision mechanism. For instance, we rule out mechanisms that could only produce some but not all of the choices.

[8] I assume here that the agent has true beliefs.

[9] To be clear, Thoma (2021a: 925) acknowledges that the mentalist could appeal to patterns in choice to defend mentalism against her arguments. Yet, she considers this reply only in the context of what she calls substantive versions of mentalism.

[10] Of course, another purpose of theoretical models in economics like the market-of-lemons (Akerlof 1970) or the Hotelling model (Stokes 1963) is to show that the same pattern in choice-behavior can arise from the same principles (e.g., information asymmetries) across many different circumstances. For this project, black-boxing or abstracting away from the underlying decision mechanism can be highly advantageous because different processes can underlie the same patterns in choice in different circumstances or for different agents. What is important to note here, though, is that those models do not attempt to capture the preferences of any particular agent. Instead, they try to isolate causal mechanisms at a level that abstracts away from the (psychological) idiosyncrasies of different agents (Mäki 2009). In other words, those models do not employ the tools of RPT as a research program that allow us to derive utility-functions from observed choice-behavior. Instead, they hypothesize utility-functions and credence-functions that can stand in as an abstract description of the agent's decision mechanisms and demonstrate how they—in combination with other postulated features—produce the choice patterns (or stylized facts) that they are meant to account for. In fact, this type of theoretical modeling in economics has been around far longer than RPT as a research program.

To further illustrate this, recall that the agent in Thoma's example does not choose coffee on the day on which she has an important meeting. Imagine now that we observe the agent only on days on which she does not have an important meeting. Hence, we assign her a utility-function that always attaches a higher utility to 'coffee' than to 'tea.' Because we never observed the agent on the day of an important meeting, her choices so far are still rationalizable by the utility-function that we came up with. Therefore, understood as a mere representation of her actual pattern in choice, this utility-function seems to be perfectly adequate. However, suppose we now understand this utility-function as a representation of a set of functionally individuated states realizable by a particular set of decision mechanisms. In that case, it would not be adequate to assign this function because it would rule out that the choices of the agent are indeed produced by the more complex mechanism, according to which a looming meeting will impact the choice between coffee and tea. To put things differently, in assigning a utility-function that always attaches a higher utility to 'coffee', we commit ourselves to the position that the underlying decision mechanism is not the more complex one that Thoma presupposes in her example. Hence, utility- and credence-functions put restrictions on the processes that cause our choices via narrowing down the set of possible causal histories of your choices. In doing so, they provide us with information on the causal processes underlying the agent's choices. While the resulting type of explanations may look shallow when it comes to explaining single choices of a single agent, this is no longer true once we focus on (large) sets of choices.

My argument here is in line with what Jackson and Pettit (1990) call program explanations. The general idea behind such explanations is that the presence of a higher-level property (e.g., a preference) "programs" for the realization of one of several lower-level properties that causes the event in question. Hence, by citing the relevant higher-level property we provide information on the causal history of the event in question via narrowing down the set of its possible causal histories (for a more detailed take on the application of program explanations to preferences in economics, see Holmes 2022).

Importantly for my purposes, the behaviorist cannot avail herself of program explanations, as according to her conception of preferences, preferences refer directly to choice. By refusing to say anything about the mechanisms that underlie our choices, behaviorism forgoes our ability to offer program explanations because it is not introducing higher-level restrictions that would "program" for the realization of one of several lower-level properties. Hence, Thoma's argument that parsimony would favor behaviorism over other conceptions of preferences seems unwarranted because program explanations are an epistemic good that behaviorism cannot deliver.

To conclude, this section aimed to show that behaviorism is neither favored by the alleged advantages identified in the literature nor do behaviorists provide any principled obstacles to alternative interpretations of preferences. I, therefore, find that the project of rehabilitating behaviorism is insufficiently motivated. In other words, the behaviorist cannot play for a win here. Yet, she may still aim at a draw. If she could dispel the arguments that speak against behaviorism, she could still try to claim that whether someone opts to be a behaviorist or goes for a different conception is a matter of philosophical taste (e.g., for or against different theories of explanations like program explanations, for or against metaphysical jargon, etc.) (cf. Clarke 2020). However, in the following two sections, I will argue that behaviorists cannot convincingly deal with these arguments.

5. Preferences and Beliefs

Recall that the first argument states that preferences cannot be identical with choice-behavior because of the role that beliefs play in choice-theory. To start my discussion of this argument, I will outline a potential reply by the behaviorist. The behaviorist could reply that this argument assumes a common sense understanding of preferences according to which their relation to choice-behavior is mediated via belief. Yet, the term 'preference' in economics is a technical term (see Clarke 2016). Assuming that we need to construe the relation between preference and beliefs in economics analogously to how they are related according to common sense begs the questions against behaviorism.

The behaviorist could then emphasize that, according to her, preferences are identical with choices in particular circumstances. Consequently, beliefs can enter the picture as part of the circumstances in which our choices take place. We could then, for example, say that a preference for falafel over kebab simply means that I will choose falafel over kebab in those circumstances that include me believing that the falafel in front of me is vegetarian and indeed a falafel. This response would basically admit that we need beliefs to infer preference from choice.

Guala (2019: 386) briefly suggests such a response on behalf of the behaviorist. Yet, he dismisses it as incompatible with the goals of behaviorism because the behaviorist "would make use of the sort of psychological concepts that she intended to eliminate in the first place."[11] That is, he holds that reinterpreting economic models in such a way that beliefs are merely viewed as part of the circumstances of choice-behavior would betray the goals of RPT as an epistemological thesis. The behaviorist could no longer argue that the evidential base of economics consists only of choice data

[11] The behaviorist could also try to give a behavioristic interpretation of belief. Yet, it is well known that it is only possible to derive beliefs from behavior if we already know the agent's preferences (Rosenberg 1992). The behaviorist needs to identify one of the two concepts with something other than behavior to be able to identify the other concept with behavior. Similar arguments have also been put forward in the philosophy of cognitive science (see Braddon-Mitchell and & Jackson 2007).

because preference assignments would inevitably entail commitments about agents' beliefs. In other words, understanding beliefs as part of the circumstances could only save the behavioristic interpretation at the cost of betraying its original motivation. So, pursuing this strategy would amount to a pyrrhic victory. However, Thoma (2021b) has recently put forward an updated version of this response.

In order to provide the context for this argument, consider that predictions based on preferences will often go wrong if we do not individuate an agent's options correctly. Recall that we have already seen a version of this problem in Section 3 (i.e., the wasabi example). Here the problem was that preference ascriptions will go wrong if we do not individuate options in line with how the agent conceptualizes them herself. Yet, if it is important to 'correctly' individuate choice options, how should we go about it?

Thoma (2021b: 172) argues that we can solely focus on an agent's beliefs about the options when it comes to individuating choice options. Therefore, she holds that we can ignore the "motivating factors behind choice" when it comes to individuating choice options. This, in turn, would allow us to retain behaviorism. In particular, she proposes what she calls a limited version of mentalism about beliefs according to which descriptions of choice options should meet the following two conditions (Thoma 2021b: 173):

> 1. The description of options should be consistent with the agent's beliefs about the nature and consequences of the actions open to her, provided the agent's relevant beliefs are mutually consistent.
> 2. A perceived feature of a choice situation should be included in the description of the agent's options whenever that feature affects the agent's choice-behavior, that is, when there are choice situations where the agent would make a different choice when she believes that feature is present or absent respectively.

As the individuation of choice options, under this view, only depends on information about the agent's beliefs and not her preferences, the behaviorist can still claim to be able to avoid any commitments about the causes of choice. Moreover, as it only requires us to consider some of the features the agent believes about the options, we arrive at a rather limited version of mentalism about belief, that is, we do not need to respect all the features in virtue of which an agent conceptualizes an option but only some of them. In other words, Thoma (2021b: 184) thinks that her proposal ensures that no commitment to the "motivating factors that bring about choice" is necessary when doing economics.

She also argues that her proposal for individuating choice options already matches the practices of economists.

My objection to this argument is that it lacks an explanation for why the second step in the individuation strategy matters. In other words, it remains unclear why a "feature of a choice situation should be included in the description of the agent's options whenever [. . .] the agent would make a different choice when she believes that feature is present or absent respectively" (Thoma 2021b: 173). In this regard, it is not clear how we should envision the relationship between beliefs and choice-behavior. Of course, Thoma seems to hold that beliefs are just part of the background conditions of choice. Yet, this is hardly an explanation for why we should include certain features that the agent believes about the option into the description of the option, while we can ignore others. As I will argue now, once we attempt to explicate this relationship, Thoma's proposal turns out to make use of the kind of information she aims to avoid.

To see this, consider that there is a straightforward explanation for why an "agent would make a different choice when she believes that a feature is present or absent respectively", namely that this feature is also tracked by the agent's preference (or motivating factors as Thoma calls it). Hence, while we can ignore those features the agent believes about the options that are not tracked by her motivating factors, we have to include those that are tracked by agent's motivating factors. However, if we accept this explanation, we can no longer maintain that Thoma's account eliminates information about motivating factors (or preferences). After all, her account would rely on information about features of an option that are tracked by the agent's motivating factors in order to individuate choice options. They would merely get a different label.

To get a better grip on the problem for Thoma's proposal here, consider that there are already various proposals for individuating choice options based on motivating factors or preferences (e.g., Dreier 1996). A more recent version of this idea can be found in Fumagalli (2020). In the context of addressing several worries against the so-called re-individuation strategy for accommodating violations of choice-theoretic axioms (i.e., the idea of re-describing choice options in such a way that an agent's preferences still satisfy the relevant axioms), Fumagalli proposes that the legitimacy of incorporating certain factors into the description of choice options depends on whether the relevant factors can be shown to be tracked by agents' preferences.

Now, my point is simply that if the features that are "tracked by preferences" (and, therefore, relevant in Fumagalli's account) are exactly the same features in virtue of which the "agent would make a different choice when she believes that a feature is present or absent respectively" (and, therefore relevant according to Thoma's account), there would only be a terminological but not a

substantive difference between the two proposals. That is, both accounts would recommend seeking out exactly the same information to decide how we should individuate choice options.

If this is correct, Thoma's account cannot claim to be a belief-based alternative to already existing preference-based proposals for individuating choice options. Instead, it would simply be a different formulation of the preference-based proposals.

Yet, offering reformulations of preference-based proposals, in which terms like 'preference' or 'motivating factors' no longer appear, cannot provide a reason for behaviorism. More specifically, it does not rehabilitate behaviorism in the face of the received view that points out that behaviorism does not capture how beliefs and preferences interact in economic models to produce choices. Even if we tried to save behaviorism by interpreting beliefs as just a part of the background condition of choice, the individuation of choice option would ultimately force us to use information about agents' preferences or motivating factors to get things right.

6. Preference and Causal Explanations

In this section, I will examine attempts to address the argument that behaviorism cannot account for the fact that preferences are employed in causal explanations. I will first assess the proposal of using an interventionist account of causation to defend that behaviorism can account for the employment of preferences in causal explanations (6.1.). I will then assess the prospects of saving this idea by reinterpreting causal claims about preferences as causal claims about the context of choice (6.2.).

6.1. Causal Explanations and EUT

Recall that the second argument of the received view states that economists often refer to preferences when attempting to give causal explanations of choice-behavior. Yet, being identical with choice-behavior would prevent preferences from figuring into causal explanations of choice-behavior because an event cannot cause itself (Guala 2012). As a result, the behavioristic conception would fail to capture certain uses of the term 'preference' in economics. However, Craver and Alexandrova (2008) and, more recently, Vredenburgh (2020) have suggested that under an interventionist account of causation, preference understood along the lines of behaviorism can explain choice-behavior. I want to give this line of thought a more thorough treatment in what follows.

Roughly speaking, the interventionist account states that X is a cause of Y iff there is a possible intervention on X that changes Y (see Woodward 2005). An intervention is a surgical manipulation of the cause that changes only the value of the cause and severs it from all other causal factors that would determine its value in the absence of the intervention. It is assumed that such interventions are

possible in some very general sense. Yet, this does not imply that any human being must be able to perform the intervention.

Note that, under this account of causation, there is, of course, a trivial sense in which choices can cause choices. For example, consider the following claim: my choice to turn on the gas of my stove caused my choice to ignite the lighter. This would qualify as a causal claim under an interventionist account. If I had not chosen to turn on the gas, I would not have chosen to ignite the lighter. Consequently, an intervention that alters my choice to turn on the gas (imagine blocking my path to the stove *or* informing me that the gas is not connected) would have changed the event of me igniting the lighter.

However, those who want to argue that the interventionist account can help defend behaviorism need to go much further. To see this, consider that preferences in economics are sometimes assumed to depend on each other. For an example, consider the Independence Axiom of EUT (Mas-Colell et al. 1995: 171):

$$L \succcurlyeq L' \text{ if and only if } pL + (1\text{-}p)L'' \succcurlyeq pL'+(1\text{-}p)L''.$$

The axiom states that you prefer a lottery L over another lottery L' iff you would also prefer getting L with probability p (and L'' otherwise) to a lottery that gives you L' with probability p (and L'' otherwise). For example, you prefer coffee instead of tea just in case you also prefer getting coffee if a coin comes up heads (and nothing otherwise) to getting tea if the coin comes up heads (and nothing otherwise). The axiom needs to be satisfied for there to be a cardinal utility-function consistent with EUT.

Hausman (2012) reads the axiom as stating that preferences over simple options determine preferences over lotteries. However, Angner (2018) points out that some economists treat the axiom as stating that preferences over lotteries determine preferences over simple options. Nevertheless, under both readings the axiom postulates a causal relationship between different preferences. These causal readings of the Independence Axiom will help me to illustrate what is wrong with the interventionist defense of behaviorism.

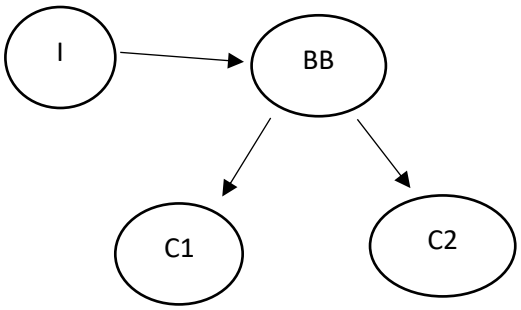
Under Hausman's reading, a preference for pL+(1-p)L'' over pL'+(1-p)L'' is determined by a preference $L \succcurlyeq L'$. The behaviorist may now wish to utilize the interventionist account to recast this reading and argue that what the Independence Axiom really commits us to is that a choice between L and L' *causes* a choice between pL+(1-p)L'' and pL'+(1-p)L''. To illustrate this, consider the claim that if I had not chosen coffee instead of tea, I would not choose coffee if a coin comes up heads (and nothing otherwise) instead of tea if the coin comes up heads (and nothing otherwise).

Of course, under a non-behavioristic interpretation of preferences, we could explain this in terms of the underlying preferences that cause both choices. Yet, the behaviorist who relies on the interventionist account will want to argue that (in those cases in which the Independence Axiom holds) an intervention on the first choice will also alter the second choice, that is, making it such that I do not choose coffee will also alter my choice over the lotteries that include the chance of getting nothing. At first glance, this might tempt us to conclude that the interventionist account enables economics to give non-trivial causal explanations of choices even if we understand preferences in terms of choice-behavior.

Yet, the main problem with this idea is that it does not take seriously the notion of a surgical intervention that is central to the interventionist account (see Woodward 2015a). According to Woodward, interventionism is an account of how we should reason about causality, that is, under which conditions we should accept causal claims or how we should go about explicating the meaning of our causal claims. In this regard, one of the main functions of interventionism is to help us disambiguate our causal claims. In other words, interventionism prescribes us to spell out the details of the intervention and, thereby, forces us to commit ourselves to what Woodward calls an ideal experiment and its outcome. Thereby, it enables us to explicate the meaning of our causal claims.

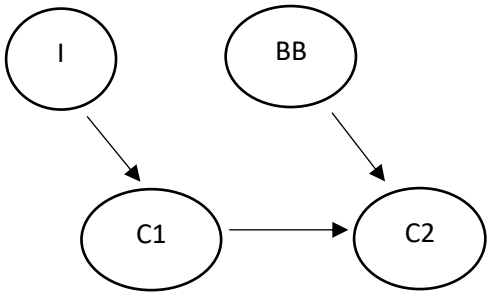
Consequently, in the case of intervening on preferences, we must ask how the envisioned intervention would look in detail in order to take the interventionist account seriously. Hence, we need to disambiguate between different interventions that could be entailed by the claim that '*if one's choice had been different, other choices would have been different as well*.' For example, consider again the first choice between tea and coffee and a second choice between lotteries that would give us tea and coffee with the same probability (and nothing otherwise). The Independence Axiom seems to commit us to an ideal experiment in which intervening on the agent's preference to produce a choice of tea instead of coffee also produces a change in the choice between the lotteries. Yet, the conception of preferences employed here is unlikely to be behaviorism. The intervention will not be an intervention on just the choice-behavior under any plausible construal of the ideal experiment. To illustrate this, consider the following graph, where BB stands for the Blackbox that includes all the factors that produce choice-behavior, and C1 stands for the choice-behavior in the first choice-scenario between coffee and tea, and C2 stands for the choice-behavior in the second-choice scenario between lotteries, I stands for the intervention, and the arrows denote *causal* connections and their directions:

Figure 1: Causal Graph 1



In contrast to this plausible (yet vaguely specified) experiment, the behaviorist would have to commit herself to an alternative intervention in which we only change the actual choice-behavior in one of the choice situations (e.g., the words she utters or her hand-movement) while severing it from all other factors that would influence it in the absence of the intervention. Only this would constitute an intervention on preferences as they are properly understood according to behaviorism. In other words, in order to make sense of the fact that economists cite preferences in causal explanations, the behaviorist would need to argue that there is a direct causal link between the choice in the first choice situation and the choice in the second choice situation. Consequently, the behaviorist would be committed to the following picture in which there is a direct causal link from one choice to another:

Figure 2: Causal Graph 2

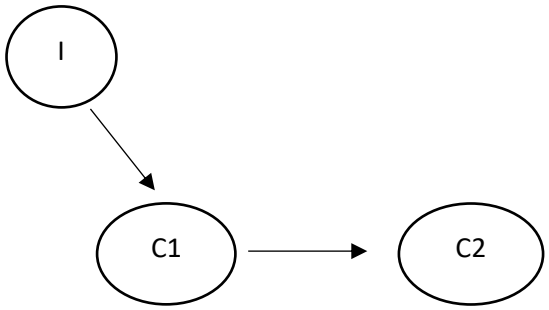


However, the problem that arises is that C2 is now determined by BB (the Blackbox) and C1 (the first choice). Yet, BB and C1 can have countervailing implications on C2. Assuming that the Independence Axiom holds, the behaviorist has to argue that changing the choice in C1 to coffee is (ceteris paribus) sufficient to causally produce a choice for the lottery containing coffee in C2. After all, a choice for coffee in C1 *entails* a choice for the lottery containing coffee in C2 according to the Independence Axiom.

Yet, given that our intervention on the choice in C1 only severs C1 from all other factors that would otherwise influence it (i.e., BB), C2 is still influenced by BB. Now, because BB is unaffected by the intervention, it still impacts C2 (i.e., it has the same impact that it had before the intervention). Yet,

the impact of BB on C2 is incompatible with C1 being sufficient to causally produce choice-behavior in C2. For example, if the original choice in C1 was coffee and we now intervene in order to change it to tea, then the causal connection between C1 and C2 would imply that the lottery containing tea is chosen in C2. Yet, as BB is unaffected by the intervention, the causal connection between BB and C2 would still imply that the lottery containing coffee is chosen in C2. Hence, we get contradictory results for C2. To avoid this, one of the two causal connections must go. The only available option for the behaviorist is to deny that there is an underlying mechanism that causes choice-behavior. Yet, this is obviously an untenable position. Nevertheless, it would give us the following picture in which C1 is the only variable causally influencing C2.

Figure 3: Causal Graph 3



Assuming that there is no underlying mechanism that causes choice-behavior and that the first choice is the only variable that causes the second choice obviously solves the problem of getting contradictory results for C2. However, it would commit us to a highly implausible picture of the world. Hence, I take this response to be unconvincing. Instead, we should resolve the problem by committing ourselves to the non-behavioristic picture represented in Figure 1.

Before I proceed, let me make the following clarification. One might think that what I have outlined here is simply an instance of the so-called exclusion problem in the context of interventionism that is extensively discussed in the literature (e.g., Baumgartner 2010; Woodward 2015b). The debate concerns what it is appropriate to control for in assessing causal claims in contexts that involve relations of supervenience (or other non-causal relations like definitional or mathematical connections). As Woodward (2015b) persuasively argues, it not appropriate to control for supervenience bases when assessing causal connections, nor does a coherent version of interventionism require such control. In line with this, if one reads the relationship between BB and C1 (or C2) as C1 (or C2) supervening on BB, it would not be correct to sever C1 from all other factors that would otherwise influence it (i.e., BB) when intervening on it. Instead, we should think of the intervention on C1 as also an intervention on BB (for details see Woodward 2015b: 331–32). If we do so, we will get no contradictory result for C2. However, here I assume that an intervention on C1 is

possible without a change in BB and that there is a causal connection—as opposed to supervenience relation—between the two. While I think such interventions are (sometimes) possible (e.g., think of a situation akin the scene in the Marx Brothers' movie *Duck Soup* where the lemonade vendor tries to reach for his hat, but repeatedly ends up with another one), my argument does not hinge on this. What is rather important is that the behaviorist would not have gained much if they committed to the view that an intervention on C1 is also an intervention on BB in virtue of a non-causal connection that holds between the two. The question would then arise what the real difference between their position and, for instance, functionalism about preferences is.[12]

6.2. Reinterpreting Causal Claims about Preferences

A more plausible route for the behaviorist would be to argue that ascribing a preference under behaviorism is to commit oneself to a set of claims of the form '*in context K the agent chooses x*.' The behaviorist could argue that if such claims hold, an intervention on the context K will affect choice x (see Vredenburgh 2020: 153). She could then argue that causal explanations involving preferences should really be understood as claims that causally link context and choice-behavior. In other words, we need to reinterpret causal claims about preferences as causal claims about context and choice.

The claim that different contexts have different causal influences on choice is, of course, perfectly acceptable. In fact, I take it to be so uncontroversial that we do not even need to employ an interventionist framework to convince us of this fact. Yet, how does it fare in relation to causal claims that economists make about preferences?

In order to asses this, let us look at an example that could easily work in the behaviorist's favor. When game theorists study social norms, they are interested in showing how changing people's preferences or beliefs would affect the stability of those norms (Lewis 1969; Young 1996; Bicchieri 2016). According to Bicchieri's highly influential analysis of social norms, three conditions must be present for the existence of a social norm (e.g., Bicchieri & Chavez 2010): (i) individuals must believe that the norm exists, (ii) individuals must have a conditional preference for following the norm (such conditional preferences are sensitive to the empirical expectation that others comply with the norm and the normative expectation that others expect them to obey and may sanction for noncompliance), and (iii) individuals must actually possess the empirical and normative expectations.

At first glance, this definition explicitly specifies two features on which the agent's conditional preference depends, that is, her empirical and normative expectation. In this regard, it does not seem

[12] I thank an anonymous reviewer for urging me to clarify this issue.

to treat preferences as mere patterns in choice, but tells us something about their determinants. Yet, the behaviorist can argue that Bicchieri's conditions (ii) and (iii) can be easily reformulated such that they solely focus on context and choice. In particular, she can restate these conditions as *'individuals have a preference for/choose following the norm only in contexts in which they hold the empirical and normative expectations.*' Hence, intervening on this context would alter whether the agent chooses to comply or not. This would basically collapse the two conditions into a claim about a simple input-output relation (see Clarke 2020). The input is defined in terms of context (i.e., expectations), and the output is choice. Hence, there are cases in which claims about preferences are easily restated as claims about context and choice.

However, this reinterpretation seems unmotivated. As we have already seen, by including states like expectations into the context, the behaviorist "would make use of the sort of psychological concepts that she intended to eliminate in the first place" (Guala 2019: 386). In particular, to intervene on the relevant context (i.e., the expectations) in a way that is precise enough for an interventionist account, one would need to have some grasp of what influences people's expectations. Yet, to acquire this information, one would have to make the kind of assumptions about an agent's psychology that most behaviorists are trying to avoid. Hence, to argue that causal claims about preferences should be recast as causal claims about context and choice, the behaviorist would have to adopt a notion of context that betrays her original motivations.[13] Considering this, there appears to be little basis on which the behaviorist could demand that scholars like Bicchieri are required to adopt the behaviorist's outlook on preferences and amend their definitions accordingly.[14]

[13] The behaviorist could, of course, try to make a move here similar to the one that Thoma (2021a) makes in the context of individuation problem and maintain that we can adopt a "limited version of mentalism" for specifying the relevant contexts. However, this move would run into the same problems that I outline above.

[14] To provide an arguably even more problematic example for the behaviorist, consider a Prisoners Dilemma in normal-form. In such a game, agents can choose between cooperating and defecting. In the scenario where both agents cooperate, they will both be rewarded with R. If they both choose to defect, they will both receive punishment P. However, if one decides to cooperate while the other defects, the cooperator will receive a sucker's payoff S while the defector will receive a temptation payoff T. Each agent's preferences follow the ranking $T \succ R \succ P \succ S$. This information already allows us to explain why both agents will defect. Roughly, each agent prefers $T \succ R$. Consequently, if they believed that the other player would play 'cooperate,' they would play 'defect.' Similarly, both players prefer $P \succ S$. Hence, if they believed that the other player would play 'defect,' they would play 'defect.' Therefore, they will always choose to defect independently of what they believe the other player will do. In sum, the preferences over the individual outcomes allows us to explain the choice of the agents to defect. Can we reinterpret this explanation as a claim of the form 'in context K the agent chooses x?' It would hardly work for the behaviorist to say that in a context in which the agent chooses T over R and P over S, she will choose defect over cooperate. T, R, P, and S are not options actually available to the agent. Similarly, also the claim that in contexts in which the agent prefers $T \succ R$ and $P \succ S$, she will choose defect over cooperate seems of no help to the behaviorist. If 'prefers' here denotes a non-behavioristic concept that is supposed to be part of the relevant context of choice, one can no longer endorse that preferences are merely patterns in choice. This seems even more problematic than putting beliefs and expectations in the context of choice (see Hausman 2000). Finally, even if the behaviorist could find a clever way to repackage information about the agent's preferences for outcomes as merely constituting the context of choice, the question would remain what the advantage of such a repacking is.

On top of this, I hold that there are clear heuristic advantages that favor Bicchieri's original formulation of her conditions for social norm over the behavioristic reformulation. Bicchieri (2014) herself insists that social norms cannot be analyzed just in terms of observable behavior to highlight the importance of expectations for conceptualizing them. To illustrate why behaviorism threatens to obscure this vital feature in the game-theoretic study of conventions and social norms, consider the following example by Guala (2019: 386):

> Tony [prefers] not to contribute to the organization of the Christmas fair because he believes that no one in the neighbourhood will give any money, although he regrets that the fair will not take place. Vince, in contrast, [prefers] not [to] contribute because he hates Christmas fairs and wouldn't give any money even if everyone else did.

The preferences of these two agents are very different. Even though, in the actual world, none of them would contribute to the organization of the Christmas fair, Tony is facing the kind of coordination game that underlies social norms under Bicchieri's view, while Vince does not. The most straightforward explanation for why they are engaged in different games is that, even though they face the same context, their preferences are relevantly different. Tony's preference are conditional preferences that depend on certain expectations where Vince's do not. Moreover, we can be confident in this explanation because we have evidence that Tony feels regret and Vince does not. Under a non-behavioristic interpretation, it becomes immediately clear why Tony's regret can serve as evidence for the hypothesis that only Tony is facing a type of coordination game, that is, Tony's regret can inform us about the determinants of his preferences.

But what is the story the behaviorist can tell us for why we can connect the evidence of Tony's regret with the conclusion that Tony is facing a coordination game? Of course, she can hypothesize, after observing a change in neighborhood contribution and a change in Tony's behavior but not in Vince's, that only Tony is facing a coordination game. Yet, the behaviorist appears to lack a plain explanation for why Tony's regret is a relevant piece of evidence in virtue of which the hypothesis could have been formed before making these observations.[15]

Moreover, when dealing with conventions and social norms we often would like to know why compliance differs between agents that we intuitively frame to be in the same context. In this regard,

[15] The behaviorist could try to make inferences about hypothetical choices by appealing to the fact that Tony felt regret for his actual choice. She could then, in turn, infer that Tony and Vince are facing different games based on their different hypothesized choices. Yet, it is again hard to see how the inference from Tony's regret to hypothetical choices is supported unless we endorse the assumption that Tony's regret informs us about the motivational factors underlying his choices. In this regard, I take it that identifying preferences with those underlying factors makes it more apparent why certain information, e.g., Tony's regret, can serve as evidence for what kind of game the agents are facing than identifying preferences merely with choices in a particular context.

it seems natural to say that two agents make different choices in the same context (e.g., a context in which they have to choose whether to contribute to the organization of the Christmas fair or not) because of their different expectations and the resulting preferences. However, the behaviorist who also admits factors like expectations into the relevant context of choice seems to be committed to the somewhat more cumbersome claim that the two agents are making different choices because they are in different contexts (e.g., in virtue of their different expectations). I do not think that this is the best way to frame the issue given the kind of questions we are interested in when studying social norms.[16]

All in all, I, therefore, take it that Bicchieri's conditions are more straightforwardly construed as claims about what influences certain preferences under a non-behavioristic interpretation than as claims about context and choice. More generally, I conclude that reinterpreting causal claims involving preferences in terms of context and choice, while (sometimes) technically possible, is insufficient to reestablish the attractiveness of behaviorism.

7. Conclusion

I have argued that there are good reasons for rejecting RPT as an ontological thesis. In other words, I hold that we should reject narrow behaviorism about preferences in economics and stick with the received view among philosophers of economics that preference are non-choice states that cause choices. Not only is there insufficient motivation for adopting behaviorism in economics, but recent attempts to defend it do not dismantle the common worries that speak against behaviorism. To be clear, this does not necessarily speak against RPT as a research program. Supplemented with other evidence, the tools developed by RPT as a research program can still be usefully employed, for example, for making inferences to preferences. Yet, the problems of RPT as an ontological thesis highlight that RPT as an epistemological thesis is implausible. In light of this, we clearly cannot and should not restrict the evidential base of economics to choice-behavior. Instead, we should banish the specter of revealed preference theory once and for all.

Acknowledgements

[16] It could be argued that it would be better to provide an example concerning a more conventional research topic of economics than social norms, e.g., the effect of a tax on commodities. I hold that the fact that scholars like Bicchieri (2014) explicitly state that social norms cannot be analyzed in terms of observable behavior makes research on social norms a particularly important case for the behaviorist to grapple with. Nevertheless, there are also more conventional research topics that could serve as examples. For instance, Mattauch et al. (2022) represent the influence of policies (e.g., a tax) on consumers' preferences for high- or low-carbon goods via an appreciation parameter. Hence, they make preferences effectively conditional on people's apperception, which summarizes various motivational factors like "values or tastes." While they are not as straightforward as Bicchieri regarding their intended use of 'preference', various passages throughout the paper suggest that they do not view appreciation as merely a feature of the choice-context. For instance, they tell us that they assume "that a single parameter can be used to translate relevant aspects of preferences into 'appreciation' for low-carbon consumption" (Mattauch et al. 2022: 13). Behaviorists may wish to regiment such talk. However, like in the case of social norms such over-regulation seems under-motivated. I thank an anonymous reviewer for raising this concern.

Thanks to Bele Wollesen, Marcel Jahn, Isaac Kean, Anna Alexandrova, Francesco Guala, Christopher Clarke, and the reviewers for their comments and support.

References

Afriat, Sydney N. (1967). The Construction of Utility Functions from Expenditure Data. *International Economic Review*, *8(1),* 67–77.

Akerlof, George A. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics, 84(3*), 488–500.

Angner, Erik (2018). What Preferences Really Are. *Philosophy of Science, 85(4),* 660–68.

Angner, Erik (2019). We're All Behavioral Economists Now. *Journal of Economic Methodology*, *26(3*), 195–207.

Baumgartner, Michael (2010). Interventionism and Epiphenomenalism. *Canadian Journal of Philosophy*, *40(3),* 359–83.

Beck, Lukas (2022). Why We Need to Talk About Preferences: Economic Experiments and the Where-Question. *Erkenntnis*, Advance online publication.

Beck, Lukas and James D. Grayot (2021). New Functionalism and the Social and Behavioral Sciences. *European Journal for Philosophy of Science, 11(4*), 1–28.

Bernheim, B. Douglas and Antonio Rangel (2008). Choice-Theoretic Foundations for Behavioral Welfare Economics. In Andrew Caplin and Andrew Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (155–92). Oxford University Press.

Bicchieri, Cristina (2014). Norms, Conventions, and the Power of Expectations. In N. Cartwright and E. Montuschi (Eds.), *Philosophy of Social Science: A New Introduction* (208–32). Oxford University Press.

Bicchieri, Cristina (2016). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Bicchieri, Cristina and Alex Chavez (2010). Behaving as Expected: Public Information and Fairness Norms. *Journal of Behavioral Decision Making*, *23*(2), 161–78.

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge University Press.

Braddon-Mitchell, David and Frank Jackson (2007). *Philosophy Of Mind And Cognition: An Introduction*. Blackwell.

Bruni, Luigino and Robert Sugden (2007). The Road Not Taken: How Psychology was Removed from Economics, and How It Might Be Brought Back. *The Economic Journal*, *117(516),* 146–73.

Camerer, Colin F. (2008). The Case for Mindful Economics. In Andrew Caplin and Andrew Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (43–69). Oxford University Press

Clarke, Christopher (2016). Preferences and Positivist Methodology in Economics. *Philosophy of Science*, *83(2),* 192–212.

Clarke, Christopher (2020). Functionalism and the Role of Psychology in Economics. *Journal of Economic Methodology*, *27(4),* 292–310.

Craver, Carl F. and Anna Alexandrova (2008). No Revolution Necessary: Neural Mechanisms for Economics. *Economics and Philosophy*, *24(3),* 381-406.

Dietrich, Franz and Christian List (2016). Mentalism versus Behaviourism in Economics: A Philosophy-of Science Perspective. *Economics & Philosophy*, *32(2),* 249–81.

Dreier, James (1996). Rational Preference: Decision Theory as a Theory of Practical Rationality. *Theory and Decision*, *40(3),* 249–76.

Fodor, Jerry A. (1968). *Psychological Explanation: An Introduction to the Philosophy of Psychology.* Crown Publishing Group/Random House.

Fumagalli, Roberto (2020). On the Individuation of Choice Options. *Philosophy of the Social Sciences*, *50(4),* 338–65.

Graham, George (2010). Behaviorism. In Edward N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*.

Guala, Francesco (2012). Are Preferences for Real? Choice Theory, Folk Psychology, and the Hard Case for Commonsensible Realism. In Aki Lehtinen, Jaakko Kuorikoski, and Petri Ylikoski (Eds.), *Economics for Real: Uskali Mäki and the Place of Truth in Economics* (137–55). Routledge.

Guala, Francesco (2019). Preferences: Neither Behavioural Nor Mental. *Economics and Philosophy*, *35(3),* 383–401.

Gul, Faruk and Wolfgang Pesendorfer (2008). The Case for Mindless Economics. In Andrew Caplin and Andrew Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (3–42). Oxford University Press.

Hands, D. Wade (2013). Foundations of Contemporary Revealed Preference Theory. *Erkenntnis*, *78(5),* 1081–108.

Hands, D. Wade (2014). Paul Samuelson and Revealed Preference Theory. *History of Political Economy*, *46(1),* 85–116.

Hausman, Daniel M. (2000). Revealed Preference, Belief, and Game Theory. *Economics & Philosophy*, *16(1),* 99–115.

Hausman, Daniel M. (2012). *Preference, Value, Choice, and Welfare*. Cambridge University Press.

Hicks, John R. and Roy G. D. Allen (1934). A Reconsideration of the Theory of Value. *Economica, 1*, 52–76.

Holmes, Travis (2022). How Revealed Preference Theory Can Be Explanatory. *Studies in History and Philosophy of Science Part A*, *91*, 20–27.

Houthakker, Hendrik S. (1950). Revealed Preference and the Utility Function. *Economica*, *17(66),* 159–74.

Jackson, Frank and Philip Pettit (1990). Program Explanation: A General Perspective. *Analysis*, *50(2),* 107–17.

Lewis, David K. (1969). *Convention: A Philosophical Study.* Blackwell.

Mäki, Uskali (2009). MISSing the World. Models as Isolations and Credible Surrogate Systems. *Erkenntnis*, *70(1),* 29–43.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green (1995). *Microeconomic Theory*. Oxford University Press.

Mattauch, Linus, Cameron Hepburn, Fiona Spuler, and Nicholas Stern (2022). The Economics of Climate Change with Endogenous Preferences. *Resource and Energy Economics*, *69*, 101312

Moscati, Ivan (2018). *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. Oxford University press.

Pareto, Vilfredo (2014). *Manual of political economy: a critical and variorum edition*. Oxford University Press. (Original work published 1927)

Quilty-Dunn, Jake and Eric Mandelbaum (2018). Against Dispositionalism: Belief in Cognitive Science. *Philosophical Studies*, *175(9),* 2353–72.

Rabin, Matthew (2002). A Perspective on Psychology and Economics. *European Economic Review*, *46(4–5),* 657–85.

Rosenberg, Alexander (1992). *Economics: Mathematical Politics or Science of Diminishing Returns?* University of Chicago Press.

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

Samuelson, Paul A. (1950). The Problem of Integrability in Utility Theory. *Economica*, *17(68),* 355–85.

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

Stokes, Donald E. (1963). Spatial Models of Party Competition. *American Political Science Review*, *57(2),* 368–77.

Thoma, Johanna (2021a). Folk Psychology and the Interpretation of Decision Theory. *Ergo*, *7,* 904–36.

Thoma, Johanna (2021b). In Defence of Revealed Preference Theory. *Economics & Philosophy*, *37(2),* 163–87.

Vredenburgh, Kate (2020). A Unificationist Defence of Revealed Preferences. *Economics & Philosophy*, *36(1),* 149–69.

Wong, Stanley (2006). *Foundations of Paul Samuelson's Revealed Preference Theory: A Study by the Method of Rational Reconstruction*. Routledge.

Woodward, James (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, James (2015a). Methodology, Ontology, and Interventionism. *Synthese*, *192(11),* 3577–99.

Woodward, James (2015b). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91(2),* 303–47.

Young, H. Peyton (1996). The Economics of Convention. *Journal of Economic Perspectives*, *10(2),* 105–22.