Philosophy of Science (2023), **00**, 1–19
doi:10.1017/psa.2023.112

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Causal Explanation and Revealed Preferences

Kate Vredenburgh

Department of Philosophy, Logic, and Scientific Method, the London School of Economics and Political Science, London, UK
Email: K.Vredenburgh@lse.ac.uk

(Received 27 May 2022; revised 04 May 2023; accepted 22 August 2023)

Abstract

This article tackles the objection that revealed preferences cannot causally explain. I mount a causal explanatory defense by drawing out three conditions under which such preferences can explain well, using an example of a successful explanation employing behavioral preferences. When behavioral preferences are multiple realizable, they can causally explain behavior well. Behavioral preferences also explain when agential preferences cannot be analytically separated from the environment that produces the relevant behavior (Condition 2) and when the environment is a significant causal factor (Condition 3). Thus, there are not causal explanatory grounds to completely bar revealed preference explanations from social science.

1. Introduction

Sometimes, social scientists aim to describe what happened: What percentage of male, college-educated Americans voted for Donald Trump in 2016? Are gender differences in alcohol consumption robust across cultures? Other times, social scientists seek to predict—in a merely forecasting sense—what will happen. Are we due for another financial crisis? Which labor inputs will increase productivity?[1] Often, however, social scientists aim to understand why a type of phenomenon occurs. Why are some nations rich and others poor? What are the effects of mandatory K–12 schooling on intergenerational wealth inequality?[2]

In economics, causal inference and psychologically enriched models have replaced psychologically noncommittal theory as the preferred means to understand how the world works (Alexandrova et al. 2021). Randomized controlled trials are held up as the "gold standard" for causal inference (Gelman 2011, 956), and social scientists working

[1] Purely statistical machine-learning methods are increasingly used in the social sciences for such predictive tasks (see Chalfin et al. 2016 for an example use of machine learning to predict which workers will be the most productive).

[2] The division of the tasks of social science into descriptive versus causal is taken from Gelman (2011, 955).

with observational data treat them—best case!—as if they were generated by a conditionally randomized experiment (Hernán and Robins 2018, ch. 3). One of the main aims of econometrics over the last three decades has been to estimate causal effects across a wide variety of economic outcomes in different populations, which has become particularly important for the evaluation of policies.

Complimenting the focus on causal inference in econometrics and applied microeconomics is the use of experimental results from psychology by behavioral economists to improve the psychological realism of economic models. Insights from psychology are leveraged to create a new general model of rational human decision making, such as prospect theory, or new macroeconomic models of mispricing in the stock market (Froot and Dabora 1999), saving for retirement (Banks et al. 1998; Laibson 1997), and so on.

The privileging of causal inference and psychological accuracy make rational choice models of the behavior of economic agents seem, at best, out of step with the times, none more so than so-called revealed preference approaches. Revealed preference approaches, a set of theoretical and empirical research programs that abstract away from the psychological causes of choice behavior, remain prevalent in theoretical economics, especially in welfare economics, and in empirical research, particularly in the study of consumer demand, and have been defended by economists such as Ken Binmore (2009) and Farak Gul and Wolfgang Pesendorfer (2008).

"Revealed preference" is a somewhat unfortunate misnomer, as the name connotes an inference to individual psychological states from evidence of what people choose. The revealed preference program as defended by Binmore and Gul and Pesendorfer interprets the concept of "preference" in a way that does not represent any psychological states. Here, I will focus on *behavioral preferences*, which interprets the concept of "preference" as a behavioral disposition attributable to an agent (more on this in section 2).

If we take her at her word, the revealed preference theorist's insistence that "preference" ought to be interpreted in entirely behavioral terms seems to remove her from the project of making causal inferences entirely. This tack has been taken by some defenders of revealed preference approaches, such as Binmore (2008), who argue that revealed preference approaches aim at prediction in the merely forecasting sense. This defense leaves revealed preference approaches subject to a serious objection, the causal explanation objection (see section 2). *Ceteris paribus*, a theory that explains a set of phenomena is better than one that does not. And, if they cannot causally explain, one might doubt that revealed preference approaches can successfully claim such predictive advantage over psychologically realistic models as to merit their continued inclusion in social scientific research programs.

In this article, I give a new explanatory defense of revealed preference approaches, one that picks up on a line of defense suggested in Vredenburgh (2020). I argue that if one assumes a counterfactual dependence theory of causal explanation, then behavioral preferences, and thus revealed preferences, can causally explain agents' choice behavior, and be part of a good explanation of those choices (sections 3 and 4). So, the causal explanation objection is false.

In sections 5–7, I then go on to argue the further point that behavioral preferences sometimes explain choice better than explanations that appeal to psychological preferences. I draw out three conditions under which behavioral preferences may

explain choice better than psychological preferences. The first condition is when the behavioral disposition is multiply realized by different psychological mechanisms (section 5). Here the explanatory defense is conditional: If one accepts multiply realizable properties that explain better than their realizers, then one should accept that dispositional preferences can causally explain behavior better than psychological preferences. This argument, however, raises the objection that behavioral preferences really pick out psychological preferences. In section 6, I rebut this objection. This objection also invites us to explore other conditions under which behavioral preferences explain well. Section 7 ends by exploring two further conditions: when the behavioral disposition lies at the intersection of the agent and her environment (Condition 2) and when the agent is highly constrained by her environment (Condition 3).

2. Dispositional preferences and the causal explanation objection

Paul Samuelson's (1938) "A Note on the Pure Theory of Consumer's Behavior" originated the revealed preference program in microeconomics. Samuelson showed that consumer demand can be modeled by demand functions—the amount of each good that will be purchased given a set of prices and the agent's income—and consistency constraints on those demand functions. The key consistency constraint on choices proposed by Samuelson is the so-called Weak Axiom of Revealed Preference (WARP), which states that if a consumer purchases bundle B at price p when bundle C was available, then, if C is chosen at price q, B is not available. The consumer's actions are taken to "reveal" a preference for B over C, where revelation is merely a matter of engaging in a certain pattern of consistent behavior. Samuelson is also commonly read as arguing that consumer demand should be modeled with revealed rather than psychological preferences, as the study of consumer demand would be better off if it were to jettison psychological concepts.[3]

Samuelson's work spawned a fruitful theoretical and empirical research program. Philosophers of economics, such as Dietrich and List (2016a), Hands (2012, 2013), Hausman (2011), Guala (2012), and Thoma (2021) have focused on methodological and interpretational questions around theoretical and empirical revealed preference modeling in microeconomics. This article contributes an explanatory defense of revealed preference approaches to those debates. The defense is very much in the spirit of Guala (2019) and Clarke (2020), who argue that the scientist's explanatory or epistemic task ought to determine the interpretation of "preference" at play.

The foundation of this explanatory defense is an interpretation of revealed preferences as behavioral dispositions to choose.[4] Once one interprets a revealed

[3] According to Hausman (2011), Samuelson is best interpreted as maintaining that revealed preferences are defined in terms of an agent's actual choices. There are interesting interpretational issues regarding Samuelson's interpretation of revealed preferences and whether it was consistent throughout his work (see Hands 2014 for one discussion).

[4] The explanatory defense of revealed preferences in this article is very much inspired by Guala (2019), who defends a view of preferences as belief-dependent dispositions to choose. According to Guala (ibid., 384), these preferences can be realized in different ways depending on the "circumstances of choice and on the characteristics of the decision-maker." This article can be viewed as contributing to his explanatory defense by identifying conditions under which behavioral preferences explain. It is also in the spirit of Cartwright's (1989) defense of capacities as central to science, but it does not assume that

preference as a behavioral disposition to choose, it becomes apparent that revealed preference approaches are continuous with other modeling practices in the social sciences that focus on patterns in agents' behavior under certain conditions, rather than their psychological states.

Accordingly, in this article, I will deal with the general notion of a dispositional preference, which is uncoupled from any particular mathematical representation or rational requirements on preferences. A dispositional preference, as I conceptualize it, is a *simple behavioral rule that results in a choice of the same option across a particular type of context*. A dispositional preference can be represented by different types of mathematical objects, such as a complete and transitive binary relation or a decision rule. And, they may be embedded in different types of models, such as rational choice theoretic models or agent-based models. In other words, one should think of dispositional preference as a genus, of which there may be many species, each defined in part by the specific rational requirements. Common to these species is the fact that each is a behavioral disposition that meets at least the minimal requirement of consistent choice in the same context.

The dialectical upshot of examining questions about behavioral preferences generally is to move beyond the well-developed, theory-backed intuitions about how to interpret "preference" in rational-choice models (either behaviorally or mentally). By examining agent-based models, we may sidestep some entrenched assumptions to revisit the question of whether a behavioral dispositional can explain.

Of course, one might resist this move and object that dispositional preferences, as outlined in the preceding text, are not preferences.[5] Preferences, continues the objection, have a certain kind of rational structure. In particular, preferences are transitive: If I prefer vacations at the seaside to vacations in the mountains, and vacations in the mountains to vacations in the desert, then I ought to prefer vacations at the seaside to vacations in the desert. Transitivity establishes a consistency in one's actions, and that consistency makes one's preferences rational. A simple behavioral rule, by contrast, need not have that rational structure. In particular, a behavioral rule need not be transitive.

There are a number of possible responses here, which are mutually supporting. The first response is that we should take transitivity in one's choice behavior as a contentious theoretical posit that needs further argument, rather than as an obvious truth about rationality. It is not obvious that choosing cake over pie one day and pie over the cake the next is troublingly inconsistent (ruling out preference change, indifference, or other explanations), unlike believing that there is cake on the table and believing there is no cake on the table. So, it is at least not obvious that it is rationally required of agents that their preferences are transitive.[6]

capacities underlie causal structure, nor that economic models are generally successful at isolating capacities (see Alexandrova and Northcott 2013).

[5] Thanks to Richard Bradley for this objection.

[6] The money pump argument is usually taken as one of the strongest normative arguments in favor of a requirement of transitive preferences (see Ramsey 1928 for a canonical statement of the argument); for a canonical argument that preferences aren't required to be transitive, see Quinn's (1990) discussion of the self-torturer case.

The second response is that this article is concerned with descriptive applications of rational choice theory, rather than with normative applications. For predictive and explanatory purposes, transitivity may be too strong a requirement on preference, especially if contexts are only individuated by the (intrinsic) properties of the objects of choice, as is standard in mainstream rational choice theoretic modeling (Dietrich and List 2016b). Of course, it is theoretically desirable that preferences are partially picked out by their satisfaction of certain rationality requirements, to distinguish preferences from mere behavior. But preferences can still have a rational structure under more minimal requirements of consistency (say, pair-wise consistency). Or a different account of rationality may be appropriate for some subspecies of preference. One promising example is so-called ecological rationality, according to which a decision rule would be rational if it helps the agent to achieve a specified goal in light of the structural features of its environment (Simon 1968). Ecological rationality seems a more fitting concept for the type of agential rationality attributable to agents in agent-based models, discussed in the next section.

With the dispositional account of preferences in hand, we can now turn to the causal explanation objection. That objection is an extremely simple one. The first premise of the argument is just a statement of the revealed preference interpretation at issue in this article: Revealed preferences are behavioral dispositions. The second premise of the argument states that behavioral dispositions cannot causally explain choices. And so, the argument concludes, revealed preferences cannot causally explain choices.

This objection strikes many—proponents and critics of revealed preferences alike—as so obvious that its conclusion is often assumed, rather than argued for. After all, Premise 1 is merely a restatement of the interpretation of revealed preferences at issue. So, all the heavy lifting in the argument is done by Premise 2, which seems uncontestable. For example, Binmore (2009, 19–20), a staunch defender of revealed preferences, asserts that:

> In revealed-preference theory, it isn't true that Pandora chooses b rather than a because the utility of b exceeds the utility of a. This is the Causal Utility Fallacy. It isn't even true that Pandora chooses b rather than a because she prefers b to a. On the contrary, it is because Pandora chooses b rather than a that we say that Pandora prefers b to a, and assign b a larger utility.

For Binmore, the representation of an agent as preferring b to a just is a representation of her choice of b over a. In our terms, it is just a representation of a behavioral disposition to choose b over a. So, the disposition cannot explain the choice (Vredenburgh 2020). On the side of the critics, Hausman (2000, 106) also asserts that revealed preferences cannot cause choice behavior, as "it is only preference (not preference* [Hausman's term for the revealed preference interpretation given in section 2]) that combines with belief to determine choice." As we will see in section 6, Hausman (2011) takes it to be conceptually impossible that a dispositional preference causes choice because behavioral dispositions cannot combine with mental states to produce choice.

To rebut the causal explanation objection, then, one must rebut Premise 2, that a behavioral disposition cannot causally explain. Sections 3 and 4 of this article do just that.

3. Revealed preferences and causal explanations: Schelling's spatial proximity model

The case that behavioral preferences can explain is built on the example of a theoretical model from Thomas Schelling (1971). Schelling's model is a highly idealized "toy model": It makes a number of simplifying, false assumptions, and represents a small number of causal factors (Reutlinger et al. 2018). However, building the argument around the example of Schelling's spatial proximity model has three advantages for our purposes. First, it is simple and easy to present. Second, it substantiates one of the claims of section 2, that behavioral preferences are used in models that are not rational choice theoretic models. Third, Schelling's discussion of how to model segregation that accompanies his spatial segregation model supports the behavioral preference interpretation of the model and explanatory claims of this and the next section. Furthermore, despite being a toy model, Schelling's model can be used to generate so-called how possibly explanations in terms of possible causes of segregation. As I will argue in the following text, the model highlights a sparse number of potentially significant causal factors of segregation, such as agents' behavioral tendencies. This claim is all that is needed for the argument, whose aim is to reject the impossibility of causal explanations in terms of behavioral preferences (ibid.). Furthermore, sociologists have gone on to develop more complex versions of Schelling's model, including models of segregation in actual cities (Benenson et al. 2002). Thus, more sophisticated versions of Schelling's model plausibly pick out actual causal structure linking behavioral preferences and choices.

Schelling's (1971) work on segregation is an often-cited, powerful example of how troubling aggregate results can emerge from seemingly innocuous individual preferences. His simple agent-based model generates explanations of patterns of residential segregation, and it spawned one approach to studying residential segregation based on modeling individual preferences.

Schelling's (1971) so-called spatial proximity model aims to model how patterns of segregation are produced by individual choices in responses to incentives. The model contains a population of agents which is divided into two groups (call them "dogs" and "cats"). The agents occupy spaces on a grid, with a maximum of one agent occupying each space. Each dog has a preference tolerance threshold F, which is a real number between 0 and 1 that represents a preference for the percentage of fellow group-members in adjacent cells (and the same for cats). That preference is cashed out in behavioral terms: What it is to have a preference tolerance threshold of .4, for example, is to remain in one's cell if the percentage of fellow dogs in adjacent cells is at or above 40%, and to move otherwise.

The agents in the model therefore choose in accordance with a simple decision rule encoded in the preference tolerance threshold. The agent's preference tolerance threshold summarizes dynamic patterns of behavior that involves two different types of counterfactual scenarios: If a dog were to be surrounded by too few dogs, she would move; otherwise, she would stay put. This decision rule is thus plausibly interpreted as ascribing a behavioral disposition to agents, where there is a robust dependency of an action on features of the surrounding environment.

Let us call agents who occupy a cell where the percentage of fellow group-members in adjacent cells is above their preference tolerance threshold "satisfied,"

and agents where the percentage is below their preference tolerance threshold "dissatisfied." Time is divided into discrete units (call one unit of time a "turn"). The rule for agent movement is that in each turn, satisfied agents remain in their cell, and dissatisfied agents move to the closest empty cell where they will be satisfied.

Different parameters of the model can then be adjusted to see which equilibria are produced. The spatial proximity model is best known for the case in which each group's tolerance threshold is the same and the groups are of roughly equal size. The model has two equilibrium states: When $F < 1/3$, a random pattern emerges; when $F > 1/3$, a segregated pattern emerges. Adjusting the parameters of the model and observing the resulting equilibria reveals dependencies between the final equilibrium state and the values of the initial parameters, especially the agents' tolerance threshold.

There are, of course, a number of causally relevant factors that determine an individual's choice of neighborhood in the actual world, such as the price of housing, the agent's budget constraint, and so on, which are left out of the model. Furthermore, many cases of segregation result from the structure of an institution or organized action by a subset of agents. Schelling is explicit that his model does not model these mechanisms, with the caveat that it is difficult to draw a clean line between economic, institutional, and individual preference-based mechanisms of segregation (1971, 144). Still, there is a good case to be made that Schelling's model explains patterns of segregation when individual discriminatory preferences—that is, preferences to associate with a certain percentage of people who are "like me"—are the primary causal factor.

The generalizations represented by Schelling's spatial segregation model causally explain agents' choices. I contend that the dynamics of all the individuals' movements, as determined by their behavioral dispositions, number of members of each group, and the ability to move freely between rounds according to their behavioral disposition, causally explain patterns of segregation for systems in which there are no other causes of segregation that significantly affect the outcome (such as the institutional and economic causes mentioned in the preceding text).

This claim is supported by the theorist's ability to use the model to ask and answer counterfactual questions about what equilibria result from changes to the model's parameters. Here, I am assuming a broadly counterfactual account of causation and causal explanation (Hall and Paul 2013). For example, we can use the model to answer questions such as: If dogs' and cats' preference tolerance thresholds were lower, would there still be a stable equilibrium pattern of segregation? As was mentioned in the preceding text, the spatial proximity model gives a cutoff point ($F < 1/3$) below which individuals' behavioral dispositions do not produce a stable equilibrium pattern of segregation (assuming roughly equal group size). Furthermore, there is a range of interventions that change dogs' and cats' preference tolerance thresholds and thereby change the cutoff point. This indicates that we can use the model to understand how patterns of segregation counterfactually vary according to different dispositional preferences, holding all else fixed.

The ability to ask and answer such counterfactual questions on the basis of interventions on variables in the model indicates an underlying causal model that is driving the dynamics modeled by the spatial segregation model. Thus, we have a causal explanation of an aggregate outcome—namely, segregated residential

housing—in terms of a disposition of agents to behave in a certain way under different conditions, which is efficiently summarized by a decision rule. In other words, a dispositional preference explanation.

Say that Schelling's model does indeed make a compelling case that dispositional preferences can, in principle, causally explain choices. And yet, the causal explanation objection of section 2 has struck many as extremely compelling. Where does that objection go wrong?

My diagnosis is that some of the disagreement rests on an ambiguity in "choices." In the case of individual choices of particular options, the revealed preference theorist may grant that revealed preferences approaches do not allow the theorist to construct good causal explanations. After all, there often is easily accessible information about the agent's psychological states that more robustly explains her choices. For example, a psychological explanation of why an agent chose what she did in terms of the reasons for her choice can be used to explain choices of new options, whereas a revealed preference explanation cannot be so used (Dietrich and List 2016b). However, when "choices" is understood as the aggregate pattern of individuals' choices, then dispositional preferences do sometimes explain. More precisely, a set of dispositional preferences either causally explain the resulting states that are partially constituted by a set of individuals' choices, or states that causally depend on those choices.

Schelling's model thus offers support for the possibility of causal explanations that feature dispositional preferences. If the argument bears out, this undermines a powerful argument in the philosophy of economics literature, according to which revealed preferences cannot causally explain choices (Hausman 2000, 2011; see section 6 for further discussion). However, it does not support the continued inclusion of dispositional preferences in the modeler's toolkit. In addition, we ought to ask whether dispositional preferences can be part of good explanations. In particular, any defense of the explanatory power of dispositional preferences has to address the claim that psychological preferences always explain better. The rest of the article defends the explanatory power of dispositional preferences against such a threat.

4. Are dispositional preference explanations good explanations?

Causal explanations are easy to come by. For example, say that Edith asks Federica why a running car moves forward at a certain speed. Federica explains that a running car moves forward because it's not in reverse and because the person driving pressed the gas pedal down a certain amount. According to counterfactual dependence accounts of causation and causal explanation, Federica's explanation is indeed an explanation—the speed of the car depends on how much the gas pedal is depressed. But it is, intuitively, not a very good explanation, especially when contrasted with an explanation of the car's forward movement that gives information about the internal workings of the engine.[7] More generally, scientists, engineers, and other generators and users of explanations also care about whether the relevant explanations are *good*

[7] Hitchcock and Woodward (2003, 184) discuss the relative explanatory power explanations of plant growth that appeal to an explanatory generalization relating water and fertilizer to plant height versus explanations that appeal to plant physiology.

explanations. We thus need to now query whether revealed preference explanations are sometimes good explanations.

To address this question, I will draw on a common criterion of explanatory goodness. This criterion is *generality*. An idea going back to Hume is that causal attributions should be general, that is, a description of a cause should represent those properties of the cause that are shared by tokens of that type-level cause. If I observe a number of differently colored bowling balls knock over some bowling pins, which fall in the same manner regardless of the color of the impacting bowling ball, I should not attribute "being blue" as a causally relevant property of the bowling ball. In other words, causal attributions should abstract away from irrelevant details. A similar principle holds for explanatory causal generalizations. A causal generalization is more explanatory in virtue of being more general.[8] This latter point is the point that I will rely on in the defense of dispositional preference explanations.

To make the case that revealed preference explanations because they are more general, it will be helpful to have a particular criterion of explanatory generality on the table. Here I will look at *invariance*, or stability under certain types of changes to the system or object that is the target of explanation (Mitchell 2009; Hitchcock and Woodward 2003; Woodward 2003). Hitchcock and Woodward (2003) discuss a number of ways that a generalization can be invariant. For present purposes, I will focus on invariance under testing interventions.

Invariance under testing interventions is the key interventionist measure of explanatory depth, according to Woodward (2003, 248) and Hitchcock and Woodward (2003). Testing interventions are interventions that produce a change in the outcome variable. An invariant explanatory generalization is one that holds (approximately) under a large range or change in kind of testing interventions. Consider an explanation of plant growth in terms of water and fertilizer. Such a generalization is not very invariant, because testing interventions that change how the watered is delivered, say, break the explanatory generalization. (Think about the explanatory generalization "lemon trees grown indoor in pots grow inversely to the amount of water and fertilizer" fares under the testing intervention of delivering all of the water at the beginning or end of the growth time, for example.)

Let's move to the argument that Schelling's spatial proximity model can provide good causal explanations, when judged against the criterion of whether the explanations generated a model are invariant under a range of possible testing interventions.

Here we quickly run into a potential problem. If "possibility" is understood as physical possibility, as is common in the social sciences (Holland 1986), then there do not seem to be any possible testing interventions on a behavioral disposition. A testing intervention changes the value of the variable of interest while the values of

[8] There is a vast literature on generality and explanation, especially as related to abstraction and to explanatory depth (e.g., Godfrey-Smith 2009; Strevens 2011; Thomson-Jones 2005; Woodward 2003; Hitchcock and Woodward 2000), and as related to questions about reduction in the sciences (e.g., Fodor 1975; Jackson and Pettit 1992; Putnam 1975; and Weslake 2010). Generality is often cashed out in terms of *abstraction*, or the number of possible objects or systems to which the generalization applies. The invariance criteria that I use in this section, however, embodies a different conception of generality than abstraction (Hitchcock and Woodward 2003). Thank you to one of the anonymous reviewers for pushing me to clarify this point.

all other variables are held fixed. But, the objection goes, how would one do so for a behavioral disposition? Many of the interventions on dispositional preferences, the objection continues, are interventions on agents' psychological states. However, it is difficult to specify a psychological intervention that is nonbacktracking, holds the values of other variables fixed, and changes all the agents' behavioral disposition to the same new behavioral disposition. One intervention, for example, could be a comprehensive education program early in life that encourages positive associations with outgroup members. An educational program, however, is extremely unlikely to change all agents' behavioral dispositions to the same alternative and hold all else fixed. It could change some group member's beliefs, lowering their preference tolerance threshold; it could cement a negative association in others, raising their preference tolerance threshold.

In response to this objection, I will adopt Woodward's (2003, 52–53) account of the notion of possibility at play: logical possibility. Woodward argues that this restriction to physically possibly interventions excludes many claims that, on the face of it, seem causal, such as an explanation of the effect of the position of the moon on the tides. He argues that interventions need to be merely logically possible, and that the theory can still generate well-defined interventions, supported by our ability to use the laws of nature to reason about what would follow from logically possible interventions.

With the relevant interpretation of possibility in mind, we can now ask whether the generalizations established by the spatial proximity model are invariant under testing interventions. The answer is a fairly straightforward yes. How and why new individuals from one group move into the neighborhood does not matter for the explanatory generalization that relates the percentage of out-group members (cause) and individual behavior to move or to stay in the neighborhood (effect). Thus, this generalization is invariant under a range of testing interventions. The behavioral disposition is also invariant under a range of testing interventions. We can imagine holding other facts about agents fixed (their income level, liquidity constraints, etc.), and consider a range of testing interventions that changes the agent's behavioral dispositions: conditioning them with positive rewards to raise their preference tolerance threshold; implanting a new belief that it is desirable to live around a certain percentage of cats; implanting a desire to live around a certain percentage of out-group members; and so on. Because they are invariant under a range of testing interventions, the explanations of segregation generated by Schelling's spatial proximity model are good explanations, according to the criterion of invariance under testing interventions.[9]

By probing the model with counterfactual theories of explanation and explanatory goodness in mind, we thereby get support for the claim that the model can be used to generate good explanations of aggregate patterns of choices and their downstream effects. This strategy of focusing on a model, however, also comes with a risk, as it raises a serious objection. The model may seem to generate explanations but may not

[9] This argumentative strategy is unlikely to convince someone who is skeptical of the existence of robust causal relations in the social world that explain (e.g., Cartwright 1989). But, even for such a skeptic, generalizations featuring behavioral preferences may explain just as well as generalizations featuring psychological preferences, i.e., not very well at all. But the arguments here would still establish the claim that they are on the same explanatory footing.

in fact do so. Instead, it may provide a framework for formulating causal hypotheses, as it posits causal relationships that may hold in only some background circumstances, a fact that may or may not be specified by the model. A causal hypothesis is then needed that specifies some background circumstances to explain (Alexandrova 2008; Alexandrova and Northcott 2009). Models, the objection goes, can provide some of the raw materials for explanations, but that more causal knowledge is needed to construct explanations, knowledge that the models do not provide.

There are two ways to expand this objection. One might claim that Schelling's model does not on its own provide causal explanations, but that the conditions under which a dispositional preference causes a certain equilibrium pattern of segregation can be further specified to generate a causal explanation. This does not challenge the arguments for the explanatory power of dispositional preferences; it instead challenges the possibility to generate explanations from the model alone. But this latter challenge does not impugn the argument that dispositional preferences can explain well because the explanations are more general.

Another way to further specify the objection does challenge the arguments of sections 3 and 4. According to this objection, Schelling's model does not yet pick out a causal explanatory relationship between preferences and aggregate outcomes. It instead provides a schema that must be filled in by a psychological preference, namely, a preference that causes the agent to move in line with the decision rule. More generally, the objection states that the relationship of counterfactual dependence between an agent's behavioral disposition and outcomes cannot be read back into the world. To convince this objector, we need to give them reason to think that actual agents have robust behavioral dispositions that ground such an invariant generalization and are themselves explanatory.

5. Difference making and multiple realizability

In the following text, I will argue that it does not always improve an explanation to substitute a psychological preference for a dispositional preference when the behavioral disposition is multiply realizable by different underlying psychological mechanisms. And indeed, explanations that use behavioral preferences are better than explanations in terms of the underlying realizers, either the realizers of the behavioral disposition or the realizers of the causal relation.

The argument of this section is inspired by the methodological discussion in Schelling (1971). Before laying out his model, Schelling sets up his modeling task: He states what he aims to explain, which determines which features of the world he will focus on and motivates some of his modeling choices in terms of properties of the target systems. Schelling states that he is interested in "aggregate results that the individual neither intends nor needs to be aware of, results that sometimes have no recognizable counterpart at the level of the individual" (1971, 145).

Of course, it is the intentional actions of agents that gives rise to these aggregate results—Schelling, and most other economists, are not metaphysical emergentists. However, these aggregate results are better explained by dispositional than psychological preferences because agents' behavioral dispositions are multiply realized by the underlying psychological states (Putnam 1975; Fodor 1975). As the epigraph states (Schelling 1971, 146):

> What goes on in the "hearts and minds" of small savers has little to do with whether or not they cause a depression. The hearts and minds and motives and habits of millions of people who participate in a segregated society may or may not bear close correspondence with the massive results that collectively they can generate.

It is easy to dismiss Schelling as slightly confused here. Of course, what goes on the "hearts and minds" of small savers has something to do with whether there is a financial crisis, or whether people live in segregated communities. Unless one is a strong emergentist, individual actions collectively produce these aggregate results. However, I take Schelling's point here to be one about multiple realizability. The behavioral dispositions that produce individual choices in certain conditions—such as living in a neighborhood surrounded by others that you identify as unlike you—are multiply realizable by different types of psychological states. And the particular causal relations are also "micro-realization robust," for example, the causal relations are robust to changes in the realizers (List and Spiekermann 2013). Together, these two claims establish the superiority of high-level explanations in terms of a dispositional preference, as it is at this level that the causes make a difference to the effect.

Consider the first claim, that behavioral dispositions are multiply realizable by underlying psychological realizers. This claim is plausible for the social systems picked out by Schelling's spatial proximity model. The agents may have racist preferences; they may all prefer to live in a diverse neighborhood where the majority of their neighbors are not group members, as long as at least 40% are; they may feel a negative valence toward non-group members; and so forth. What makes a difference to the pattern of segregation is not whether agents have any of those particular mental states. Instead, the higher-level, behavioral preference is what makes a difference.

Multiple realizability may be a more intuitive claim than micro-realization robustness. Wouldn't changes in the underlying realizers change the causal structure? Here I will draw on an insight from Satz and Ferejohn (1994) to substantiate the claim of microrealization robustness. They remark that "rational-choice explanations are most plausible in settings in which individual action is severely constrained, and thus where the theory gets its explanatory power from structure-generated interests and not from actual individual psychology" (ibid. 72). In other words, individual "preferences" are derived from the agent's location in a social structure. In competitive markets, for example, a firm's preference to maximize profit is determined by competitive market demands, the relationship to its creditors, the behavior of other firms, and so on. In such contexts, it better captures the relevant causal structure to interpret the preference as a structure-generated behavioral disposition that robustly produces the right kinds of behavior for the firm to survive in a competitive market.

With this point in mind, we can now ask: Why might the behavioral disposition to move when surrounded by a certain proportion of out-group neighbors be microrealization robust? Schelling was writing in a society where individuals' preferences about where to live were strongly shaped by group membership and prevailing ideologies. Even if patterns of psychological dispositions changed, the behavioral dispositions and attendant causal mechanisms that produce segregation

would likely be robust against those changes, if the changes were caused by a stable ideology and group identity (Cohen 2001; Mills 1997). This point is bolstered by sociological approaches to studying discrimination, which focus on behavior and do not assume a particular psychological mechanism (Pager and Shepherd 2008, 182).

So, we have one condition under which dispositional preference explanations are good explanations: The disposition is multiply realizable, and the causal relation between the disposition and the outcome of interest is microrealization robust. A sociologist that observes patterns of racial segregation in a neighborhood, for example, may find that the various individual choices to move neighborhoods are made when individuals are surrounded by very few in group members. By contrast, the individual psychological states that preceded agents' choices to move may vary. So, the more general, and thus the better, causal explanation gives information about the agents' behavioral dispositions. And reflection on the constraints that social structures place on individuals, as well as the attendant ways that ideologies shape their mental states, allows us to rebut the objection that the behavioral disposition must be substituted by a psychological state to causally explain choice well.

This defense of dispositional explanations, though, raises the specter of another objection. Why think that the higher-level realizer is a *behavioral* preference, rather than a coarse-grained psychological state? The next section will deal with this objection.

6. Psychological or behavioral dispositions?

There have been a number of important recent arguments against revealed preference interpretations and in favor of a psychological interpretation of preferences. One of the most important objections was posed by Hausman (2011), who argued against the conceptual possibility of revealed preferences. Such an argument needs to be countered if the previously mentioned strategy has any hope of working.

Hausman (2011) argues against three different argumentative strategies that proponents of revealed preferences might adopt. I will deal with one argumentative strategy here, which challenges the interpretation of revealed preferences as behavioral dispositions. Hausman argues that revealed preference explanations often, if not almost always, make background assumptions about the agent's belief states. Take, for example, the Schelling spatial proximity model. for the model to be predictively and explanatorily useful, the agents must be taken to have beliefs about whether each of their neighbors is like them or not. These beliefs are part of the background conditions that activate one of the disjuncts of the decision rule: Move if the percentage of similar neighbors does not meet some threshold, or do not move it the percentage does so. Agents who had no beliefs about the similarity of their neighbors would not move or stay, and so the analysis of segregation in terms of agent responses to the similarity of their neighbors would be false of these agents.

The second step of Hausman's argument was mentioned in section 3. This second step asserts that only psychological states, such as psychological preferences, combine with beliefs to produce choice. One might worry that this assumption is objectionably *a priori*. Shouldn't we base our theorizing on what science teaches us about the mental?

Here, though, we run into a follow-up objection. The objector may grant this point, but object that, according to a prominent theory of mind, dispositional preferences are really coarse-grained mental states.[10] Call this the *functionalist objection*. A dispositional preference reliably produces an output behavior and mental state in response to certain sensory inputs or environmental cues, as well as mental states.[11]

Functionalism, however, is an uneasy basis on which to make a strong case that dispositional preferences are psychological. One compelling objection to functionalism is that it attributes mental states to agents that, intuitively, do not have mental states (Block 1980; Schwitzgebel 2015).[12] If functionalism is the best, or only, theory of mind to account for behavioral preferences, this fact suggests that behavioral preferences may not be mental states after all.

The second response is that even if one grants that agents with multiply realizable behavioral preferences have a coarse-grained psychological disposition that causes choice, this interpretation will not be plausible of all models utilizing behavioral preferences to explain causally. In the final section, I argue for two further conditions under which behavioral preferences explain choice: (1) that the disposition lies at the interface of the agent and the environment, and (2) that the environment is highly constraining.

7. Two further explanatory conditions

In some models, behavioral preferences refer to properties at the boundary of the agent and the environment; thus, preferences cannot be cashed out in terms of the agent's psychology. This is the second condition under which behavioral preferences explain. And sometimes the environment is so constraining that it is the major determinant of choice; this is the third condition.

To motivate both conditions, I want to start us out with a parable from Herbert Simon of an ant walking across a beach:

> We watch an ant make his laborious way across a wind- and wave-molded beach. He moves ahead, angles to the right to ease his climb up a steep dunelet, detours around a pebble, stops for a moment to exchange information with a compatriot. Thus he makes his weaving, halting way back to his home . . . . It is a sequence of irregular, angular segments—not quite a random walk, for it has an underlying sense of direction, of aiming toward a goal . . . . Viewed as a geometric figure, the ant's path is irregular, complex, hard to describe. But its complexity is really a complexity in the surface of the beach, not a complexity in the ant. (Simon 1968, 51)

Why would Simon claim that the complexity is "really" a complexity in the surface of the beach, not a complexity in the ant? Here there are two readings of the claim, neither of which is friendly to a psychological account of behavioral preferences.

[10] Thanks to Christian List for this objection.

[11] To make sense of the environmental cues of the Schelling model, the view seems to require so-called long arm functional theories; see Block (1990).

[12] Schwitzgebel targets materialism, but his arguments apply equally well to functionalism. But note that some functionalists may be willing to bite the bullet here (e.g., Pettit and List 2011).

The first reading of the claim is that it is difficult for the scientist to locate the relevant property as definitively within or outside of the agent. Simulations are used in the social sciences when scientists cannot analytically separate the properties of the agent from properties of the environment. One such example are agent-based models of financial markets, which make minimal assumptions about human behavior to, for example, explain financial bubbles and crashes (Farmer and Foley 2009). One account of why minimal behavioral rules are apt to model market dynamics is that the information is "in the market," and prices drive individual behavior by acting as signals that do not require agents to represent all the aggregated information to act on it (Hayek 1948, ch. 4; Satz and Ferejohn 1994). Of course, the agents have to have experience in the environment, such that a behavioral response is reliably cued by the environment (Kahneman and Klein 2009). In such cases, scientists tend to attribute the relevant behavioral cause to the intersection of the agent and the environment. This indicates that it is a behavioral disposition, not individual psychology, that causes the phenomena of interest.

A second reading of the claim that the complexity is "really" complexity in the beach is that the environment constrains the individual to such a degree that their individual psychology does not matter causally, given that the agent has certain goals. If the ant, for example, has certain navigational capabilities and familiarity with the environment, as well as the goal to arrive home safely, then we can explain its behavior in terms of the causal properties of the environment. We can felicitously describe the agent as having a robust behavioral disposition that allows her to achieve those goals, in light of the constraining environment. Psychology, again, is beside the point, as long as it does not alter the agent's goal or undermine her capacity to interact with her environment.[13]

Under what conditions do either of those interpretations of Herbert's parable make sense? For example, why is it felicitous to idealize away the decision procedures of agents, or focus on the causal powers of the environment? Here we can draw on another insight of Simon's: that our environments are designed. In other words, our environments are often the product of human intentions to produce external conditions that facilitate the achievement of some goal. Simons further held that the behavior of agents is largely determined by the environment, if their cognitive system—or, more generally, "inner" system—is adapted to that environment. If the latter assumption is true of designed systems, this vindicates a focus on modeling the causal powers of the environment. It also may explain why it is difficult to analytically separate properties of the agent's cognition and the environment: The latter has been designed to facilitate agential activity in pursuit of a goal, and the activity can often be easier to model than that cognition itself. Thus, sometimes, what the modeler is interested in studying is at the border of the agent and her environment, and they can capture system dynamics well by ignoring cognitive complexity.

Thus, not all cases of modeling with behavioral dispositions are cases of multiple realizability. So, not all dispositional preferences can be interpreted as coarse-grained psychological states.

[13] This defense is similar in spirit to Haslanger's (2016) contrastive defense of structural explanations. When a contrastive explanation assumes that agents have certain mental states, then the explanation may explain behavior entirely in terms of social structures.

8. Conclusion

This article mounted a defense against the causal explanation objection to revealed preference approaches. I argued that behavioral dispositions are sometimes causes, according to counterfactual accounts of causation. Furthermore, behavioral preferences can be part of good explanations—indeed, in cases of multiple realizability, they may sometimes be better than explanations that appeal to psychological preferences.

This general lesson fits well with a common defense of classical rational choice theory, premised on the grounds that economists are often interested in explaining aggregate outcomes that emerge from human interaction, such as stock market crashes and segregation. A common style of such a defense is to argue that explanations of these aggregate outcomes assume preferences; so, they do not explain. Instead, the explanatory work is done by social facts, such as the agent's budget constraints and incentives. This defense accepts the causal explanation objection and denies that behavioral preferences causally explain agents' choices. By contrast, a better response is to grant that social facts are important to explain aggregate outcomes that result from human interaction but push back on the assumption that behavioral preferences do not play some role in the explanation. In the Schelling model, social facts about an agent's environment, particularly the composition of one's neighborhood, are important in explaining aggregate outcomes. But individuals' behavioral dispositions to move under certain conditions are also important to produce aggregate patterns of segregation. Examples such as Schelling's spatial segregation model thus show that individual-level facts are indeed important in causally explaining an aggregate outcome, but they are facts about individual-level behavior, rather than mental states.

The arguments of this article also fit into a larger discussion of a seeming tension within economic methodology. On the one hand, economists put a lot of weight on the demand for microfoundations in terms of individual beliefs and preferences. On the other hand, economists also insist that they aim to explain aggregate patterns that are compatible with a wide variety of irrational or random behavior by individuals, and that structural factors, such as budget constraints, are explanatorily more important than individual preferences. Like Lehtenin and Kuorikoski (2007), I think that the resolution of this tension sometimes involves understanding the demand for microfoundations in terms of patterns in individual behavior, rather than psychological preferences.

Acknowledgments. The author wishes to thank Richard Bradley, Liam Kofi Bright, Ned Hall, Christian List, Susanna Rinard, Lucas Stanczyk, Michal Strevens, Johanna Thoma, and the two anonymous reviewers at *Philosophy of Science* for helpful feedback, as well as audiences at the London School of Economics' Popper Seminar, the UNC philosophy department, Brown University's Political Theory Workshop, and Cambridge University's History and Philosophy of Science Seminar.

References

Alexandrova, Anna. 2008. "Making Models Count." *Philosophy of Science* 75 (3):383–404. https://doi.org/10.1086/592952

Alexandrova, Anna, and Robert Northcott. 2009. "Progress in Economics: Lessons from Spectrum Auctions." In *The Oxford Handbook of the Philosophy of Economics*, edited by Harold Kincaid and Don Ross, 306–37. Oxford: Oxford University Press.

Alexandrova, Anna, and Robert Northcott. 2013. "It's Just a Feeling: Why Economic Models Do Not Explain." *The Journal of Economic Methodology* 20 (3):262–67. https://doi.org/10.1080/1350178X.2013.828873

Alexandrova, Anna, Robert Northcott, and Jack Wright. 2021. "Back to the Big Picture." *Journal of Economic Methodology* 28 (1):54–59. https://doi.org/10.1080/1350178X.2020.1868772

Banks, James, Richard Blundell, and Sarah Tanner. 1998. "Is There a Retirement-Savings Puzzle?" *The American Economic Review* 88 (4):769–88.

Benenson, Itzhak, Itzhak Omar, and Erez Hatna. 2002. "Entity-Based Modeling of Urban Residential Dynamics: The Case of Yaffo, Tel Aviv." *Environment and Planning B: Urban Analytics and City Science* 29 (4):491–512.

Binmore, K. 2009. *Rational Decisions*. Princeton, NJ: Princeton University Press.

Block, Ned. 1980. "Troubles with Functionalism." In *Readings in the Philosophy of Psychology, Volumes 1 and 2*, edited by Ned Block, 268–305. Cambridge, MA: Harvard University Press.

Block, Ned. 1990. "Inverted Earth." In *Philosophical Perspectives 4*, edited by J. Tomberlin, 52–79. Atascadero, CA: Ridgeview Press.

Cartwright, Nancy. 1989. *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review: Papers and Proceedings* 106 (5):124–27. https://doi.org/10.1257/aer.p20161029

Clarke, Christopher. 2020. "Functionalism and the Role of Psychology in Economics." *Journal of Economic Methodology* 27 (4):292–310. https://doi.org/10.1080/1350178X.2020.1798016

Cohen, Gerry A. 2001. *Karl Marx's Theory of History: A Defense*, exp. edition. Princeton, NJ: Princeton University Press.

Dietrich, Franz, and Christian List. 2016a. "Mentalism vs. Behaviorism in Economics: A Philosophy of Science Perspective." *Economics and Philosophy* 32 (2):249–81. https://doi.org/10.1017/S0266267115000462

Dietrich, Franz, and Christian List. 2016b. "Reason-Based Choice and Context-Dependence: An Explanatory Framework." *Economics and Philosophy* 32 (2):175–229. https://doi.org/10.1017/S0266267115000474

Farmer, J. Doyne, and Duncan Foley. 2009. "The Economy Needs Agent-Based Modeling." *Nature* 460:285–86. https://doi.org/10.1038/460685a

Fodor, Jerry. 1975. *The Language of Thought*. New York: Thomas Cromwell.

Froot, Kenneth, and Emil M. Dabora. 1999. "How Are Stock Prices Affected by the Location of Trade?" *Journal of Financial Economics* 53 (2):189–216.

Gelman, Andrew. 2011. "Review Essay: Causality and Statistical Learning." *American Journal of Sociology* 117 (3):955–66. https://doi.org/10.1086/662659

Godfrey-Smith, Peter. 2009. "Abstractions, Idealizations, and Evolutionary Biology." *Mapping the Future of Biology: Evolving Concepts and Theories*, edited by Anouk Barberousse, Michel Morange, and Thomas Pradeu, 47–56. New York City: Springer.

Guala, Francesco. 2012. "Are Preferences for Real? Choice Theory, Folk Psychology, and the Hard Case for Commonsense Realism." In *Economics for Real: Uskali Mäki and the Place of Truth in Economics*, edited by Aki Lehtenin, Jaako Kuorikoski, and Petri Ylikoski, 137–55. New York: Routledge.

Guala, Francesco. 2019. "Preferences: Neither Behavioural nor Mental." *Economics and Philosophy* 35 (3):383–401. https://doi.org/10.1017/S0266267118000512

Gul, Faruk, and Wolfgang Pesendorfer. 2008. "The Case for Mindless Economics." In *The Foundations of Positive and Normative Economics: A Handbook*, edited by Andrew Caplan and Andrew Schotter, 2–40. Oxford: Oxford University Press.

Hall, Ned, and Laurie Paul. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Hands, D. Wade. 2013. "Foundations of Contemporary Revealed Preference Theory." *Erkenntnis* 78:1081–1108. https://doi.org/10.1007/s10670-012-9395-2

Hands, D. Wade. 2014. "Paul Samuelson and Revealed Preference Theory." *History of Political Economy* 46:85–116. https://doi.org/10.1215/00182702-2398939

Hands, D. Wade. 2012. "Realism, Commonsensibles, and Economics: The Case of Contemporary Revealed Preference Theory." In *Economics for Real: Uskali Mäki and the Place of Truth in Economics*, edited by Aki Lehtenin, Jaako Kuorikoski, and Petri Ylikoski, 156–78. New York: Routledge.

Haslanger, Sally. 2016. "What Is a (Social) Structural Explanation?" *Philosophical Studies* 173 (1):113–30. https://doi.org/10.1007/s11098-014-0434-5

Hausman, Dan. 2000. "Revealed Preference, Belief, and Game Theory." *Economics and Philosophy* 16 (1):99–115. https://doi.org/10.1017/S0266267100000158

Hausman, Dan. 2011. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Hayek, F. von. 1948. "The Use of Knowledge in Society," in *Individualism and Economic Order*, Chicago: University of Chicago Press, Chapter 4.

Hernán, Michael, and Jamie Robins. 2018. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC.

Hitchcock, Christopher, and James Woodward. 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Noûs* 37 (2):181–99.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396):945–960. https://doi.org/10.1080/01621459.1986.10478354

Jackson, Frank, and Philip Pettit. 1992. "Structural Explanation and Social Theory." *Reduction, Explanation, and Realism*, edited by David Charles and Kathleen Lennon, 97–132. Oxford: Oxford University Press.

Kahneman, Daniel, and Gary Klein. 2009. "Conditions for Intuitive Expertise: A Failure to Disagree." *American Psychologist* 64 (6):515–26. https://doi.org/10.1037/a0016755

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics* 112 (2):443–77.

Lehtenin Aki, and Jaako Kuorikoski. 2007. "Unrealistic Assumptions in Rational Choice Theory." *Philosophy of the Social Sciences* 37 (2):115–38. https://doi.org/10.1177/0048393107299684

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

Mills, Charles. 1997. *The Racial Contract*. New York: Cornell University Press.

Mitchell, Sandra. 2009. "Complexity and Explanation in the Social Sciences." In *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, edited by Chrysostomos Mantzavinos, 130–45. Cambridge: Cambridge University Press.

Pager, Devah, and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34:181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740

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

Putnam, Hilary. 1975. "Philosophy and Our Mental Life." *In Mind, Language, and Reality, 291–303*. Cambridge: Cambridge University Press.

Quinn, Warren S. 1990. "The Puzzle of the Self-Torturer." *Philosophical Studies* 59:79–90.

Ramsey, Frank. 1928. "Truth and Probability." In *The Foundations of Mathematics and Other Logical Essays*, edited by Richard Bevan Braithwaite, 156–98. London: Routledge & Kegan Paul.

Reutlinger, Alex, Dominik Hangleiter, and Stephan Hartmann. 2018. "Understanding with (Toy) Models." *British Journal of the Philosophy of Science* 69 (4):1069–99. https://doi.org/10.1093/bjps/axx005

Samuelson, P. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5:61–71.

Satz, D., and Ferejohn, J. 1994. "Rational Choice and Social Theory." *The Journal of Philosophy*, 91 (2):71–87. https://doi.org/10.2307/2940928

Schelling, Thomas. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1:143–86.

Schwitzgebel, Eric. 2015. "If Materialism Is True, the United States Is Probably Conscious." *Philosophical Studies* 172:1697–1721. http://www.jstor.org/stable/24704177

Simon, Herbert. 1968. *The Sciences of the Artificial*. Cambridge, MA: MIT University Press.

Strevens, Michael. 2011. *Depth: An Account of Scientific Explanation*. Cambridge: Harvard University Press.

Thoma, Johanna. 2021. "In Defence of Revealed Preference Theory." *Economics and Philosophy* 37 (2):163–87. https://doi.org/10.1017/S0266267120000073

Thomson-Jones, Martin R. 2005. "Idealization and Abstraction: A Framework." In *Correcting the Model: Idealization and Abstraction in the Sciences*, edited by Martin Jones and Nancy Cartwright, 173–218. Amsterdam: Rodopi.

Vredenburgh, Kate. 2020. "A Unificationist Defense of Revealed Preferences." *Economics and Philosophy* 36:149–69. https://doi.org/10.1017/S0266267118000524
Weslake, Brad. 2010. "Explanatory Depth." *Philosophy of Science* 77 (2): 273–94. https://doi.org/10.1086/651316
Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Cite this article: Vredenburgh, Kate. 2023. "Causal Explanation and Revealed Preferences." *Philosophy of Science*. https://doi.org/10.1017/psa.2023.112