DOI:
10.1016/j.shpsa.2020.03.003

Document Version
Peer reviewed version

How Thin Rational Choice Theory Explains Choices

Abstract

The critics of rational choice theory (RCT) frequently build on the contrast between so-called thick and thin applications of RCT to argue that thin RCT lacks the potential to explain the choices of real-world agents. In this paper, I draw on often-cited RCT applications in several decision sciences to demonstrate that despite this prominent critique there are at least two different senses in which thin RCT can explain real-world agents' choices. I then defend this thesis against the most influential objections put forward by the critics of RCT. In doing so, I explicate the implications of my thesis for the ongoing philosophical debate concerning the explanatory potential of RCT and the comparative merits of widely endorsed accounts of explanation.

Keywords: Rational Choice Theory; Explanation; Rationality; Scientific Modelling; Decision-Making.

Word Count:10654

1. Introduction

Over the last few decades, intense debates have taken place concerning the explanatory potential of rational choice theory (henceforth, RCT). The proponents of RCT take this theory to provide informative explanations of observed choices and claim that "[no] alternative approach [has comparable] explanatory power" (Becker, 1996, 4; also Ferejohn, 2002, Ferejohn and Satz, 1995, Lazear, 2000). For their part, the critics complain that on entrenched accounts of explanation, RCT merely describes (as opposed to explains) the choices of real-world agents (e.g. Alexandrova and Northcott, 2013, Hodgson, 2012, Morgan, 2001 and 2006, Sen, 1985 and 1987, Sugden, 1991 and 2011). These authors' critique builds on the often-made contrast between so-called *thick* and *thin* applications of RCT to argue that RCT falls prey to the following dilemma. On the one hand, there are thick applications of RCT, which regard choices as the outcome of a process of instrumental reasoning and rest on empirical assumptions about the neuro-psychological substrates of choice. On the other hand, we find thin applications of RCT, which provide a formal axiomatic representation of consistent choices and make no empirical assumptions about the neuro-psychological substrates of choice. However, the critique goes, neither thick not thin applications of RCT provide adequate explanations of real-world agents' choices. More specifically, thick RCT can be used to explain choices, but derives its explanatory power (if any) from neuro-psychological findings and is inconsistent with empirical evidence from neuro-psychology. Conversely, thin RCT is insulated from falsifying empirical evidence from neuro-psychology, but lacks the potential to explain real-world agents' choices.

If correct, this critique would have far-reaching implications for scientific theorists and practitioners, since RCT applications figure prominently in a vast range of disciplines

(e.g. Boudon, 2003, on sociology, Green and Shapiro, 1994, on political science, Hausman, 2012, on economics). In this paper, I draw on often-cited applications of RCT to demonstrate that despite such critique there are at least two different senses in which thin RCT can explain real-world agents' choices. My argument proceeds as follows. Thin RCT abstracts away from all information concerning the neuro-psychological substrates of choice. This precludes thin RCT applications from being explanatory under various accounts of explanation (e.g. Craver, 2006, on the mechanistic account). Nonetheless, some of the implications derivable from thin RCT's axioms can count as explanatory under at least two widely endorsed accounts of explanation. First, they demarcate what classes of agents can make choices with the structural patterns (e.g. transitivity) defined by thin RCT's axioms, enabling rational choice theorists (henceforth, RCTs) to explicate why agents who differ radically in their neuro-psychological makeup can make choices with analogous structural patterns (*unificationist explanations*). And second, they disclose structural dependences between thin RCT's axioms and agents' choices, enabling RCTs to determine in what respects and to what degree the choices of real-world agents deviate from the choices of the agents posited by thin RCT across a wide range of actual and counterfactual conditions (*counterfactual explanations*). My thesis that thin RCT can explain real-world agents' choices in these two senses implies neither that thin RCT is *in general* more explanatory than thick RCT nor that empirical findings from neuro-psychology are *irrelevant* for explaining choices. Still, if my thesis is correct, the often-made claim that thin RCT cannot explain real-world agents' choices is mistaken, and thin RCT has a greater explanatory potential than its critics maintain.

The paper is organized as follows. In Section 2, I clarify the distinction between thick and thin applications of RCT, providing some illustrations. In Section 3, I draw on the specialized literature on scientific explanation and often-cited RCT applications to

explicate the two senses in which I take thin RCT to be explanatory. In Sections 4-7, I defend my thesis that thin RCT can explain real-world agents' choices against the most influential objections put forward by the critics of RCT. More specifically, I address in turn: the objection from *mere representations* (e.g. Hodgson, 2012, Saatsi, 2016); the objection from *causal/mechanistic explanations* (e.g. Alexandrova and Northcott, 2013, Craver, 2006); the objection from *axioms' untenability* (e.g. Sen, 1985, Sugden, 1991); and the objection from *interdisciplinary consilience* (e.g. Craver and Alexandrova, 2008, Quartz, 2008). I shall argue that these objections cast doubt on the explanatory import of several thin RCT applications, but do not undermine my thesis that thin RCT can explain real-world agents' choices. In doing so, I differentiate my thesis from other authors' accounts of how models that abstract away from empirical information concerning their targets can be explanatory (e.g. Pincock, 2015, on abstract models in physics, Rice, 2012, on abstract models in biology, Ross, 2015, on abstract models in neuroscience).[1]

Before proceeding, three preliminary remarks are in order. First, I use the expression 'neuro-psychological substrates' broadly to encompass a vast set of entities, processes and events, ranging from agents' psychological motivations to the activation patterns of particular neural areas. In doing so, I group neural and psychological substrates together because several critics and proponents of RCT alike contrast RCT applications that rest on some empirical assumptions about neural and/or psychological substrates and RCT applications that do not rest on any empirical assumption about either neural or

[1] The term 'abstraction' has been used to indicate a wide variety of representational techniques, including both idealizations that leave putatively irrelevant features of one's targets unspecified and omissions that eliminate such features altogether (e.g. Mäki, 1992, McMullin, 1985, Weisberg, 2007). Below I use 'abstraction' to encompass both of these representational techniques unless stated otherwise. This broad use of 'abstraction' is sufficiently precise to evaluate the explanatory potential of thin RCT and sufficiently general to cover most of the representational techniques discussed in the literature on abstract explanations.

psychological substrates (Sections 2-3). Second, the two accounts of explanation I examine respectively build on the unificationist and the counterfactual accounts, yet drop various controversial tenets of these accounts' original formulations (e.g. footnotes no.9-12). As I illustrate in Sections 3-7, unificationist and counterfactual accounts are not without critics, but are widely endorsed by scientific theorists and practitioners across different disciplines. Hence, showing that some thin RCT applications can count as explanatory under such accounts would have significant relevance for both the critics and the proponents of RCT. Finally, my paper aims to provide at least three contributions of general interest to the ongoing philosophical debate concerning the explanatory potential of RCT (e.g. Hausman, 2008 and 2012, Sugden, 2011 and 2013, Thoma, 2019) and the comparative merits of widely endorsed accounts of explanation (e.g. Kuorikoski and Ylikoski, 2010 and 2015, Mäki, 2001 and 2013, Vredenburgh, 2019), namely: address often-made criticisms of RCT that build on prima facie plausible empirical and methodological presuppositions; critically assess the requirements for adequate explanation that leading authors advocate in the specialized philosophical, economic and neuro-psychological literatures; and articulate a systematic evaluation of thin RCT's explanatory potential under two entrenched accounts of explanation.[2]

2. Thin versus Thick RCT

RCTs have put forward several applications of RCT across different decision sciences (e.g. Fumagalli, 2013, Guala, 2019, Hausman, 2012). The distinction between thick and

[2] My evaluation focuses on 'accounts' of explanation, which specify what conditions explanations must satisfy to be successful, rather than 'conceptions' of explanation, which specify what explanation itself is (e.g. Saatsi, 2016). One may consistently advocate the same conception of explanation and endorse different accounts of explanation, and vice versa (e.g. Rohwer and Rice, 2016). I gloss over the debate concerning the interrelations between various accounts and conceptions of explanation (e.g. Roche and Sober, 2017) for the purpose of my evaluation.

thin applications categorizes the proffered applications of RCT into two exclusive sets. More specifically, an application of RCT counts as *thick* if and only if this application rests on *some* empirical assumptions about the psychological and/or neural substrates of choice, i.e. such application is justifiably rejected as implausible if these assumptions prove to be false or inaccurate. Conversely, an application of RCT counts as *thin* if and only if it does not rest on *any* empirical assumption about either neural or psychological substrates, i.e. the falsity or inaccuracy of assumptions about neuro-psychological substrates (if any) does not *per se* undermine the plausibility of such application.[3]

A variety of increasingly thick applications of RCT may be differentiated in the RCT literature. For instance, Loomes and Sugden's (1982) regret theory, which aims to account for a wide range of intransitive choices by incorporating reference to psychological feelings of regret and rejoicing into the description of choice options, only rests on empirical assumptions about psychological substrates. For its part, Glimcher's (2011, ch.6-8) hard expected utility theory, which aims to account for real-world individuals' choices by showing that individuals behave the way they do because their neuro-psychological architecture encodes choice options' desirability in a continuous monotonic fashion, rests on empirical assumptions about both psychological and neural substrates. Still, as I illustrate in Sections 3-7, several applications of RCT are plausibly regarded as thin (rather than thick) in the sense I explicated. For the mere fact that an

[3] The expressions thin RCT and thick RCT are occasionally used to indicate different sets of RCT applications than the two exclusive sets I target (e.g. Sen, 1985, 109, for a contrast between 'thin' RCT applications and RCT applications that assume agents' "unfailing pursuit of self-interest"; also Guala, 2012, 151, for a contrast between behaviourist and psychological interpretations of RCT). I shall expand on different uses of the expressions thin RCT and thick RCT when the differences between such uses and my use of these expressions bear on the evaluation of thin RCT's explanatory potential (e.g. footnote no.10). Also, the authors debating over RCT's explanatory potential endorse different positions concerning what conditions empirical assumptions must satisfy to qualify as true and/or accurate (e.g. Mäki, 1992 and 2011). I do not expand on these differences since my defence of thin RCT's explanatory potential does not directly rest on such differences.

application of RCT is inspired or heuristically motivated by neuro-psychological findings does not *per se* make this application thick. To give one example, Machina's (1982) generalized expected utility theory, which aims to account for observed discrepancies between expected utility theory's implications and real-world agents' choices by modifying or even relinquishing specific axiomatic requirements (e.g. independence), is partly inspired by psychological findings, but counts as thin since it does not rest on any empirical assumption about neuro-psychological substrates. Similarly, the mere fact that the axioms figuring in some RCT applications could be given a neuro-psychological interpretation does not *per se* make these applications thick (rather than thin) unless one actually gives a neuro-psychological interpretation to such axioms. For instance, completeness is often taken to ascribe sophisticated psychological abilities to agents, yet this ascription is not implied by RCT's mathematical formalism, so an application of RCT may assume complete preferences and still count as thin. That is to say, thin RCT is a highly *abstract* mathematical formalism, but is not an *uninterpreted* mathematical formalism, i.e. "there is always an element of interpretation" in both thick and thin RCT applications (Gibbard and Varian, 1978, 666; also Guala, 2006). Let us focus on thin RCT applications.[4]

Thin RCT applications build on specific axiomatic requirements on the modelled agents' preferences, together with the implications derivable from such requirements. So-called *representation theorems* figure prominently among these implications. Such theorems

[4] The fact that an element of interpretation figures in both thin and thick RCT applications may complicate attempts to establish whether specific RCT applications are most plausibly classified as thin or thick. Still, the necessary and sufficient conditions specified at p.6 above provide clear criteria for demarcating thin and thick RCT applications across a wide range of choice contexts. That is to say, the distinction between thin and thick RCT applications may occasionally be difficult to draw in concrete cases, but RCTs are able to identify several uncontroversial cases of thin (as opposed to thick) RCT applications (e.g. the examples provided at p.6-7 above and in Sections 3-7 below).

demonstrate that if an agent's preferences satisfy specific axiomatic requirements, then this agent's choices can be represented as solutions to a constrained optimization problem. For instance, von Neumann and Morgenstern (1947) famously demonstrate that if an agent's preferences satisfy completeness, transitivity, continuity, and independence, then there exists a utility function unique up to positive linear transformations, such that for any two options, the one preferred is assigned higher expected utility. To put it differently, an agent whose preferences satisfy von Neumann and Morgenstern's axiomatic requirements necessarily behaves 'as if' she maximizes expected utility. In this context, the notion of utility does not refer to any substantive neuro-psychological magnitude, but rather refers to a mathematical representation of the agent's preferences (e.g. Fumagalli, 2013, Okasha, 2016).[5]

Thin RCT applications typically proceed in three stages.[6] The first stage specifies what set of *axiomatic requirements* are imposed on the preferences of the modelled agents. Over the last few decades, different sets of axiomatic requirements have been imposed on agents' preferences (e.g. Bhattacharyya et al., 2011, Machina, 2008, for reviews). Yet, the axiomatic requirements underlying thin RCT applications do not rest on any empirical assumption concerning the neuro-psychological substrates of choice. That is to say, nothing in thin RCT's mathematical formalism requires or implies that this theory

[5] The proponents of thin RCT commonly regard the preferences that figure in thin RCT's representation theorems as indexes of choices (e.g. Bradley, 2017, ch.2). This does not commit the proponents of thin RCT to the further claim (often associated with revealed preference theory) that preferences in general are reducible to or identical with choices (e.g. Vredenburgh, 2019). In fact, several conceptions of preference reject this purported equivalence between preferences and choices (e.g. Guala, 2012, Hands, 2013). For a critical appraisal of revealed preference theory, e.g. Dietrich and List, 2016, Hausman, 2000. For a detailed discussion of the relation between the preferences figuring in thin RCT and other conceptions of preferences, e.g. Hausman, 2011, Thoma, 2019.

[6] I say 'typically' because my three-stage reconstruction covers a wide range of thin RCT applications in several decision sciences, but allows for the possibility that some thin RCT applications do not explicitly distinguish between the three stages I mention or only focus on some of those stages.

accurately (or even approximately) represents such substrates (e.g. Fumagalli, 2019, Ross, 2014). The second stage *derives the implications* of the axiomatic requirements specified in the first stage to determine how the modelled agents' choices would vary across a range of actual and counterfactual conditions. Some of the implications derived in this second stage only demarcate a set of choices consistent with the employed axiomatic requirements (e.g. Samuelson, 1938). Other derivations, instead, enable RCTs to identify what choices the modelled agents would make in specific choice settings (e.g. Hands, 2006). The third stage aims to *determine* whether the implications derived in the second stage *hold* for the choices of the targeted real-world agents. Two kinds of contributions are provided by thin RCT applications at this third stage.

On the one hand, some contributions determine which of the auxiliary assumptions that are not part of thin RCT's mathematical formalism can be modified or removed without altering thin RCT's implications. For instance, Lehtinen and Kuorikoski (2007) document the robustness of thin RCT's implications (e.g. specific optimality results) to a wide range of variations in auxiliary assumptions concerning agents' degree of self-interest. Similarly, Mandler et al. (2012) document the robustness of thin RCT's implications to a wide range of variations in auxiliary assumptions concerning agents' computational abilities and decision-making procedures (e.g. sequential choices through axiomatizable checklists of desirable properties). On the other hand, other contributions determine how thin RCT's implications vary when one modifies or removes some of the axiomatic requirements that thin RCT imposes on agents' preferences. For instance, Machina (2008) demonstrates how thin RCT's implications vary for a vast range of axiomatic modifications (e.g. various modifications of independence) implemented in generalized expected utility analyses. Similarly, Gaertner and Xu (1999) provide a detailed comparison showing how divergences between the axiomatic requirements that

respectively underlie expected utility maximization and other choice rules (e.g. choice of the second largest element, choice of the median element) lead to divergent behavioural implications. I shall expand in Section 3 on how these two kinds of contributions are implemented to provide unificationist and counterfactual explanations of choices. For now, I note that some of the implications derivable from thin RCT's axioms explain choices not so much *in spite of* the fact that thin RCT abstracts away from all information concerning neuro-psychological substrates, but rather *because* this theory abstracts away from such information. That is to say, the axiomatic requirements underlying some thin RCT applications cannot be modified or removed without hampering these applications' potential to explain choices.[7]

3. How Thin RCT Explains Choices

In this section, I draw on the specialized literature on scientific explanation and often-cited RCT applications to explicate the two senses in which I take thin RCT to explain real-world agents' choices. More specifically, I argue that some of the implications derivable from thin RCT's axioms can count as explanatory under at least two widely endorsed accounts of explanation, namely the unificationist and the counterfactual accounts. The unificationist account targets the degree to which thin RCT's implications hold under variations in auxiliary assumptions that are not part of thin RCT's mathematical formalism. The counterfactual account, instead, targets the variability of

[7] In recent years, axiomatic requirements have been imposed on several entities other than the preferences figuring in thin RCT (e.g. Sugden, 1993, for an axiomatic foundation of regret theory where utility is defined in terms of a psychological experience of pleasure; also Caplin et al., 2010, for a representation theorem concerning the so-called reward prediction error hypothesis that dopaminergic activations encode the difference between experienced and predicted rewards). Below I focus on the axiomatic requirements imposed on the preferences figuring in thin RCT unless indicated otherwise.

thin RCT's implications to changes in the axiomatic requirements that thin RCT imposes on agents' preferences. Let us consider these two accounts in turn.

According to the *unificationist account*, explanation obtains when one discloses connections between phenomena that were formerly regarded as independent (e.g. Friedman, 1974, Kitcher, 1981). On this account, a theory is explanatory to the extent that it demonstrates that "apparently independent and diverse phenomena [are] manifestations [of a common set] of entities, powers, and processes" (Mäki, 2001, 498). The idea is to "derive [accurate] descriptions of many phenomena" from few derivation patterns and "reduce the number of types of facts we have to accept as [brute]" (Kitcher, 1989, 432; also Friedman, 1974, 15). Several thin RCT applications succeed in deriving accurate descriptions of many different choice patterns by building on few axiomatic requirements (e.g. Ferejohn and Satz, 1995, Fumagalli, 2014, on both individual and aggregate choice patterns). These applications count as explanatory under the unificationist account because they demarcate what classes of agents can make choices with the structural patterns (e.g. transitivity) defined by thin RCT's axioms, enabling RCTs to explicate why agents who differ radically in their neuro-psychological makeup can make choices with analogous structural patterns (e.g. Guala, 2019, Ross, 2014, on how agents as diverse as human organizations and non-human animals can make choices with the structural patterns defined by thin RCT's axioms).[8]

To be sure, *not all* thin RCT applications which succeed in deriving accurate descriptions of many different choice patterns count as explanatory under the unificationist account.

[8] The entities, powers, and processes figuring in the unificationist account encompass both empirical and theoretical entities, powers, and processes. Hence, both the preferences figuring in thin RCT's representation theorems and the axiomatic requirements imposed on such preferences can in principle serve as basis for explaining choices under the unificationist account.

For this account holds that only *stringent* theories can explain, and several thin RCT applications fail to place stringent limitations on what set of choices belong to thin RCT's domain of applicability (e.g. Reiss, 2012, 57-59; also Green and Shapiro, 1994, 34, for the claim that "it is not obvious what [choice patterns cannot] be explained by some variant of [thin RCT]"). Still, these stringency-related considerations do not exclude that some thin RCT applications can count as explanatory under the unificationist account. For although thin RCT's mathematical formalism may be used to represent a wide variety of choice patterns, the axiomatic requirements underlying *specific* thin RCT applications often impose stringent limitations on what set of choices can be plausibly taken to belong to thin RCT's domain of applicability (e.g. Guala, 2006, Hands, 2006). In this respect, it is telling that the critics of RCT frequently criticize thin RCT's axiomatic requirements for imposing excessively (rather than insufficiently) stringent limitations on such set (e.g. Sections 6-7).[9]

To illustrate how thin RCT applications can provide unificationist explanations of real-world agents' choices, consider again the influential thin RCT applications cited in Section 2, which aim to show that few axiomatic requirements suffice to generate specific implications for a wide range of variations in auxiliary assumptions about the neuro-psychological substrates of choice (e.g. Lehtinen and Kuorikoski, 2007, on the robustness of some optimality results to a wide range of variations in auxiliary assumptions concerning agents' degree of self-interest; also Mandler et al., 2012, 73, on models where

[9] Several tenets of Friedman's and Kitcher's formulations of this account have been subjected to criticisms (e.g. Roche and Sober, 2017, Vredenburgh, 2019, for recent discussion). I mention these criticisms in passing both because my evaluation of thin RCT's explanatory potential does not directly rest on those tenets and because many proponents and critics of RCT alike concur that RCT applications are explanatory to the extent that they are unifying as per the unificationist account (e.g. Reiss, 2012; also Mäki, 2001, 490, for the claim that "many developments in economics are celebrated because they are regarded as advancing explanatory unification").

agents who choose sequentially through axiomatizable checklists of desirable properties make choices that fit thin RCT's implications "whatever goes on in the [agents'] minds"). These thin RCT applications illustrate that in many choice settings "it is not the agents' [neuro]psychologies that primarily explain their [choices]" (Satz and Ferejohn, 1994, 74), but rather specific axiomatic requirements together with information about empirical facts that are neither neural nor psychological in character (e.g. Guala, 2019, on information about environmental and institutional factors). The availability of these thin RCT applications does not *per se* exclude that some thick RCT applications grounded in neuro-psychological detail may also count as explanatory under the unificationist account (e.g. Hausman, 2012, on thick RCT applications grounded in psychological detail; Craver and Alexandrova, 2008, on thick RCT applications grounded in neural detail). Still, those thin RCT applications enable RCTs to identify several choice patterns that are invariant to a wide range of changes in assumptions concerning agents' neuro-psychological makeup, thereby pointing to unificationist explanations grounded in formal properties of agents' preferences and in information about empirical facts that are neither neural nor psychological in character.[10]

[10] One may object that thin RCT applications that make no empirical assumptions whatsoever about their target agents "allow for derivational unification without a concomitant underlying ontological unity in the causal basis of the behavior" and therefore fail to provide unificationist explanations of real-world agents' choices (Kuorikoski and Lethinen, 2010, 356; also Mäki, 2001, for the claim that providing unificationist explanations requires one to point to ontological - as opposed to merely derivational - unifications). This objection casts doubt on the explanatory import of thin RCT applications that make no empirical assumptions whatsoever about their target agents, but does not cast general doubt on the explanatory import of thin RCT applications. For as noted in the main text, the fact that thin RCT applications make no empirical assumptions about neuro-psychological substrates does not *per se* exclude that thin RCT applications may make empirical assumptions about factors that are neither neural nor psychological in character (e.g. Guala, 2019, on thin RCT applications that make empirical assumptions about environmental and institutional factors). Hence, one may consistently grant that thin RCT applications that make no empirical assumptions whatsoever about their target agents fail to provide unificationist explanations of real-world agents' choices, yet hold that thin RCT applications that make empirical assumptions about factors that are neither neural nor psychological in character can provide such explanations.

According to the *counterfactual account*, explanation obtains when one identifies patterns of counterfactual dependence between one's explanandum variable and explanans variables (e.g. Reutlinger, 2016, Woodward, 2003). On this account, a theory is explanatory to the extent that it accurately answers what-if-things-had-been-different questions regarding how actual and counterfactual variations in the value of the explanans variables affect the value of the explanandum variable (e.g. Woodward, 2003, 210-221; also Woodward and Hitchcock, 2003). The idea is to establish whether specific hypotheses concerning one's phenomena of interest would continue to hold under various changes in the variables figuring in such hypotheses (e.g. Hitchcock and Woodward, 2003; also Ylikoski and Kuorikoski, 2010). RCTs often examine how thin RCT's implications vary when one modifies or removes some of this theory's axiomatic requirements (Section 2). This process of modifying and removing axiomatic requirements does not *per se* provide well-confirmed explanations of real-world agents' choices in the absence of empirical information about the examined choice settings (e.g. Fumagalli, 2016, on information about agents' payoffs). Still, such process can disclose several structural dependences between thin RCT's axiomatic requirements and real-world agents' choices without having to draw on any empirical information about neuro-psychological substrates (e.g. the comparisons of different thin RCT applications provided by Machina, 2008, Gaertner and Xu, 1999, cited in Section 2).

To be sure, *not all* thin RCT applications that disclose structural dependences between thin RCT's axiomatic requirements and real-world agents' choices count as explanatory under the counterfactual account. For on this account, thin RCT applications have to *accurately* answer what-if-things-had-been-different questions about the choices of both modelled and real-world agents to count as explanatory, and several thin RCT applications fail to accurately answer such questions (e.g. Pollak, 2003, against Becker's,

1976, thin RCT models of family relations; also Northcott and Alexandrova, 2015, against various thin RCT applications to prisoner's dilemma type of situations). Even so, disclosing structural dependences between thin RCT's axiomatic requirements and real-world agents' choices frequently enables RCTs to determine in what respects and to what degree the choices of real-world agents deviate from the choices of the agents posited by thin RCT across a wide range of actual and counterfactual conditions (e.g. Hindriks, 2013, Ylikoski and Aydinonat, 2014, on cases where highly abstract thin RCT models are used with less abstract models to answer why observed choices deviate from thin RCT models' implications). These contributions, in turn, count as explanatory under the counterfactual account because they enable RCTs to accurately answer many what-if-things-had-been-different questions about the choices of both modelled and real-world agents by pointing to interrelated variations in thin RCT's axiomatic requirements and thin RCT's implications.[11]

To illustrate how thin RCT applications can provide counterfactual explanations of real-world agents' choices, consider influential thin RCT applications aiming to explain how these choices vary in situations of strategic interaction. Thin RCT applications abstract away from a number of factors that are known to causally affect choices in situations of strategic interaction (e.g. Guala, 2006, on psychological factors). Moreover, RCTs can model a wide range of strategic interactions by ascribing an ordinal (rather than cardinal) interpretation to payoffs and by imposing only minimal consistency requirements on preferences. In fact, the equilibrium of various types of games (e.g. defection by both

[11] The counterfactual account is often taken to target causal patterns of counterfactual dependence (e.g. Woodward, 2003), but can be plausibly taken to target also non-causal patterns of counterfactual dependence (e.g. Baron et al., 2017, Bokulich, 2012, Povich, 2018, Saatsi and Pexton, 2013). I expand on the suitability of the counterfactual account to target both causal and non-causal patterns of counterfactual dependence in Section 5 below.

players in one-shot prisoner's dilemmas) can often be rationalized by dominance considerations without having to assume that the involved agents possess any specific beliefs (e.g. Nowak, 2006). In this context, RCTs can accurately answer a wide range of what-if-things-had-been-different questions about real-world agents' choices in situations of strategic interaction by combining information about empirical facts that are neither neural nor psychological in character with thin RCT applications that show how thin RCT's implications vary when one modifies or removes specific axiomatic requirements. To give one example, Alexander (2010) examines various influential models of rational deliberation - including perfectly rational agents who maximize their expected utility given the available information, Pettit's, 1995, virtually rational agents, Kahneman et al.'s, 1982, boundedly rational agents, Skyrms', 1990, dynamically rational agents - and demonstrates that distinct axiomatizable rationality assumptions (e.g. expected utility maximization, various sets of heuristics) lead to systematically dissimilar deliberative outcomes when agents' local interactions figure in the dynamics of rational deliberation. The availability of these thin RCT applications does not *per se* exclude that some thick RCT applications grounded in neuro-psychological detail may also count as explanatory under the counterfactual account (e.g. Northcott and Alexandrova, 2015, on thick RCT applications grounded in psychological detail; Craver and Alexandrova, 2008, on thick RCT applications grounded in neural detail). Still, the point remains that those thin RCT applications enable RCTs to accurately answer a wide range of questions about real-world agents' choices without making any assumption about neuro-psychological substrates, thereby pointing to counterfactual explanations grounded in structural dependences between thin RCT's axiomatic requirements and real-world agents' choices.[12]

[12] Modifying and removing some axiomatic requirements may prompt significant changes in thin RCT's domain of applicability (e.g. Fishburn, 1970). These changes constrain the explanatory import of various thin RCT applications, but do not prevent RCTs from identifying interrelated variations in thin RCT's axiomatic requirements and thin RCT's implications. For several

In Sections 4-7 below, I defend my thesis that thin RCT can explain real-world agents' choices in the two senses I explicated against the most influential objections put forward by the critics of RCT. More specifically, I address in turn: the objection from *mere representations* (e.g. Hodgson, 2012, Saatsi, 2016); the objection from *causal/ mechanistic explanations* (e.g. Alexandrova and Northcott, 2013, Craver, 2006); the objection from *axioms' untenability* (e.g. Sen, 1985, Sugden, 1991); and the objection from *interdisciplinary consilience* (e.g. Craver and Alexandrova, 2008, Quartz, 2008). I shall argue that these objections cast doubt on the explanatory import of several thin RCT applications, but do not undermine my thesis that thin RCT can explain real-world agents' choices.

4. Objection from Mere Representations

The objection from *mere representations* holds that thin RCT cannot explain real-world agents' choices on the alleged ground that explaining these choices requires one to supplement thin RCT with accurate (or at least approximate) empirical information about the neuro-psychological substrates of such choices (e.g. Hodgson, 2012, Saatsi, 2016). The objection proceeds as follows. The implications derivable from thin RCT's axioms enable RCTs to *represent* a wide variety of phenomena as solutions to constrained optimization problems (e.g. Becker, 1976, on racial discrimination and family relations, Kable and Glimcher, 2007, on the activation patterns of anatomically delimited neural areas). However, the mere fact that a theory enables one to represent a wide variety of

axiomatic requirements can be modified without prompting significant changes in thin RCT's domain of applicability (e.g. Bossert et al., 2006, on various modifications of transitivity, Starmer, 2000, on some modifications of independence).

phenomena by no means implies that such theory *explains* those phenomena (e.g. Kaplan and Craver, 2011, Saatsi, 2011). Moreover, the objection goes, explaining real-world agents' choices requires one to demonstrate that the implications derivable from thin RCT's axioms actually hold for these agents' choices (e.g. Morgan, 2001, Sugden, 2011). Doing so, in turn, requires one to supplement thin RCT with accurate (or at least approximate) empirical information about the neuro-psychological substrates of such choices (e.g. Hodgson, 2012, on information about psychological substrates). Hence, thin RCT cannot *per se* explain real-world agents' choices. As Hodgson puts it, thin RCT "cannot provide a real explanation" because to explain individual choices RCTs "have to consider the real […] psychological determinants of human behaviour" (2012, 94 and 100; also Lehtinen, 2013, 185, for the contention that "as-if claims are frequently made in order to […] correctly describe [behaviour, but] as-if claims in themselves […] never explain why the entity of interest behaves in the way it does").

This objection correctly notes that the mere fact that thin RCT enables RCTs to represent a wide variety of choices does not imply that thin RCT explains these choices. However, there are at least two reasons to doubt that the objection undermines the claim that thin RCT can explain real-world agents' choices. First, demonstrating that the implications derivable from thin RCT's axioms actually hold for real-world agents' choices does not *generally* require RCTs to supplement thin RCT with accurate (or even approximate) empirical information about the neuro-psychological substrates of these choices. For RCTs can (and often do) provide accurate representations of real-world agents' choices without having to provide accurate (or even approximate) representations of the neuro-psychological substrates of these choices (e.g. Fumagalli, 2014, Mäki, 2011). And second, thin RCT applications can have *some* explanatory import even in cases where RCTs are unable to demonstrate that the implications derivable from thin RCT's axioms

actually hold for real-world agents' choices. And in some cases, this explanatory import is sufficient to enable RCTs to explain the choices of real-world agents.

To illustrate this, consider the often-made distinction (e.g. Forber, 2010, Reiner, 1993, Resnik, 1991) between *how-actually* explanations, which identify what events or factors in fact account for the occurrence or specific properties of the investigated phenomena (e.g. what computational limitations in fact cause violations of RCT's axioms in a given choice setting), and *how-possibly* explanations, which only identify what events or factors may account for those phenomena's occurrence or properties (e.g. what computational limitations may cause observed violations of RCT's axioms). How-possibly explanations do not constitute well-confirmed how-actually explanations, and cannot be turned into such explanations in the absence of empirical information about real-world targets (e.g. Fumagalli, 2016). Moreover, it may be difficult for RCTs to demarcate between how-possibly explanations that respectively do and fail to provide informative explanatory insights about their targets (e.g. Verreault-Julien, 2019). Still, it would be implausible to maintain that any how-possibly explanation that does not meet the requirements for well-confirmed how-actually explanation has no explanatory import (e.g. Bokulich, 2014, Mäki, 2013, Rohwer and Rice, 2013). In fact, several authors point to the existence of a continuum in terms of explanatory import between well-confirmed how-actually explanations and distinct sets of how-possibly explanations (e.g. Forber, 2012, Marchionni, 2018, Weisberg, 2006).[13]

[13] Different authors provide distinct criteria to demarcate between how-actually and how-possibly explanations (e.g. Forber, 2010, regards how-actually and how-possibly explanations as different species of explanation, Resnik, 1991, distinguishes how-actually and how-possibly explanations by their degree of empirical support). Still, most authors concur that one can demarcate between how-actually and how-possibly explanations in terms of the sort of information provided by such explanations. The idea is that how-actually explanations provide information concerning actual states of affairs, whereas how-possibly explanations provide information concerning logically

These considerations have significant implications for the explanatory potential of thin RCT applications. To see this, consider again the counterfactual account of explanation (analogous considerations hold mutatis mutandis for the unificationist account). As noted in Section 2, RCTs frequently examine how thin RCT's implications vary when one modifies or removes some of this theory's axiomatic requirements. Modifying and removing thin RCT's axiomatic requirements does not *per se* provide well-confirmed how-actually explanations of real-world agents' choices in the absence of empirical information about the examined choice settings (e.g. Section 3 on information about agents' payoffs). Moreover, several thin RCT applications fail to count as explanatory under the counterfactual account because they fail to accurately fit real-world agents' choices. Still, several thin RCT applications accurately represent what choices are made by different sets of agents, enabling RCTs to determine in what respects and to what degree the choices of real-world agents deviate from the choices of the agents posited by thin RCT across a wide range of actual and counterfactual conditions (e.g. Hindriks, 2013, Ylikoski and Aydinonat, 2014). These thin RCT applications provide informative how-possibly explanations of real-world agents' choices because they enable RCTs to accurately answer many what-if-things-had-been-different questions about these choices by pointing to interrelated variations in thin RCT's axiomatic requirements and thin RCT's implications (Section 3).

A critic of RCT may grant that thin RCT could in principle provide some how-possibly explanations of real-world agents' choices. However, she may object that thin RCT can actually explain real-world agents' choices only if this theory is used as a *baseline* for

possible, nomically possible, physically possible, etc. states of affairs (e.g. Verreault-Julien, 2019, 22-26). I take this demarcation criterion to be sufficiently detailed for my evaluation.

constructing thick (as opposed to thin) RCT applications. The thought would be that the implications derivable from thin RCT's axiomatic requirements can ground explanations of real-world agents' choices only if these requirements are subsequently modified or eliminated through some process of decreasing abstraction (e.g. Lindenberg, 1992), thereby turning thin RCT applications into thick RCT applications. Now, RCTs often acquire valuable explanatory insights by complementing thin RCT applications with thick RCT applications that modify or eliminate the axiomatic requirements underlying thin RCT applications (e.g. footnote no.20 for an illustration). However, explaining real-world agents' choices does not *generally* require one to turn thin RCT applications into thick RCT applications. In fact, as I argue in Section 5, turning thin RCT applications into thick RCT applications may *hamper* (rather than improve) RCTs' ability to explain choices. For the axiomatic requirements underlying some thin RCT applications play a crucial explanatory role, in the sense that thin RCT's implications cannot be derived if one modifies or removes those requirements.[14]

A critic of RCT may grant that the axiomatic requirements underlying some thin RCT applications are needed to derive thin RCT's implications. However, she may object that - for all RCTs are typically able to show - these axiomatic requirements may be playing a *merely representational* (as opposed to explanatory) role. The idea would be that such axiomatic requirements may be needed to represent the physical (e.g. neuro-psychological) facts that ultimately explain real-world agents' choices without doing any explanatory work on their own. As Saatsi remarks in the debate about mathematical

[14] The mere fact that an axiomatic requirement is needed for deriving some implications does not *per se* imply that this requirement plays a crucial explanatory role (e.g. think of cases where the derived implications fail to be explanatory). Still, the axiomatic requirements underlying some thin RCT applications are crucial for gaining the explanatory insights yielded by such applications (Section 5), and this suffices to vindicate my claim that removing these requirements may hamper (rather than improve) RCT's explanatory potential.

explanations of empirical facts, "mathematics can be […] indispensable to expressing or representing non-mathematical facts which themselves do all the explanatory [work]" (2016, 579; also Saatsi, 2011).

This objection points to an insightful distinction concerning two different roles that mathematics may play in the explanation of empirical facts. Still, there are at least two reasons to doubt that the objection undermines the claim that thin RCT can explain real-world agents' choices. First, such objection specifically targets the explanatory role of mathematics in criticizing indispensability arguments for mathematical realism, which "infer the reality of [specific mathematical posits from these posits'] explanatory indispensability" (Saatsi, 2016, 1062; also Lyon, 2012). Yet, my defence of thin RCT's explanatory potential neither rests on particular realist commitments nor infers the reality of specific mathematical posits from these posits' explanatory indispensability. And second, one may consistently endorse the aforementioned distinction concerning the two different roles that mathematics may play in the explanation of empirical facts and argue that thin RCT's mathematical formalism can play an explanatory role in RCT explanations of real-world agents' choices. To illustrate this, consider Saatsi's distinction between the ontologically committing explanatory role "played by a fact that bears an ontic relation of explanatory relevance to the explanandum" and the ontologically non-committing explanatory role "played by something that allows us to grasp, or (re)present, whatever plays [an ontologically committing] explanatory role" (Saatsi, 2016, 1056). One may plausibly take thin RCT's mathematical formalism to play an ontologically non-committing explanatory role without having to ascribe to such formalism an ontologically committing explanatory role (e.g. Saatsi, 2016, 1046-1051). And such ontologically non-committing explanatory role may be sufficient to license the claim that thin RCT can explain real-world agents' choices (e.g. Section 3 on various cases where the physical

facts that are invoked to explain real-world agents' choices are neither neural nor psychological in character).

5. Objection from Causal/Mechanistic Explanations

The objection from *causal/mechanistic explanations* holds that thin RCT cannot explain real-world agents' choices on the alleged ground that explaining these choices requires one to make accurate assumptions about the causal and/or mechanistic underpinnings of such choices (e.g. Alexandrova and Northcott, 2013, Craver, 2006). The objection goes as follows. Thin RCT applications abstract away from assumptions about the causal and/or mechanistic underpinnings of choices. By doing so, they fail to include several factors that are known to causally influence choices (e.g. neural activations of areas involved in reward valuation). This, in turn, crucially hampers the explanatory potential of these applications. To be sure, one may occasionally improve one's explanations of specific choices by abstracting away from assumptions about the causal and/or mechanistic underpinnings of these choices (e.g. Strevens, 2008, ch.3). Yet, the objection goes, explaining real-world agents' choices requires one to make at least *some* accurate assumptions about the causal and/or mechanistic underpinnings of such choices. For to explain a phenomenon is to "provide some information about its causal history" (Lewis, 1986, 217; also Salmon, 1984) and to identify the "components, activities, and organizational features of the mechanism that in fact produces the phenomenon" (Craver, 2006, 361; also Kaplan and Craver, 2011). As Alexandrova and Northcott put it, thin RCT models "do not identify any actual causes", and "it is only […] subsequent research, often featuring close empirical study, that achieves explanations" (2013, 262 and 265; also

Hodgson, 2012, 100, for the claim that RCT "does not explain anything unless it points to an underlying causal mechanism").[15]

This objection correctly notes that making accurate assumptions about the causal and/or mechanistic underpinnings of real-world agents' choices often helps RCTs to explain such choices (e.g. Sections 2-4 above). Still, there are at least three reasons to doubt that the objection undermines the claim that thin RCT can explain real-world agents' choices. First, several entrenched accounts of explanation deny that making accurate assumptions about causal and/or mechanistic underpinnings constitutes a *necessary* condition for explanation (e.g. Batterman and Rice, 2014, Pincock, 2015, Sober, 1983, Ylikoski, 2013, on non-causal accounts of explanation). The availability of these accounts does not *per se* undermine the claim that making accurate assumptions about the causal and/or mechanistic underpinnings of phenomena may be occasionally required for explaining such phenomena. However, it challenges the critics of RCT to explicate why exactly in the specific case of RCT, explaining real-world agents' choices would require RCTs to make accurate assumptions about the causal and/or mechanistic underpinnings of these choices. And the critics of RCT have hitherto failed to address this justificatory challenge.[16]

[15] Philosophers put forward different characterizations of the notions of cause (e.g. Lewis, 1986, Salmon, 1984) and mechanism (e.g. Glennan, 2005, Weiskopf, 2011). I gloss over these definitional concerns since the cogency of my evaluation does not rest on which of these characterizations one endorses. In particular, I do not aim to establish whether there are precise and plausible characterizations of cause and/or mechanism which license the claim that thin RCT applications provide causal and/or mechanistic explanations of choices. For a recent discussion of the role abstractions can play in causal explanations, e.g. Okasha, 2016. For a similar discussion concerning mechanistic explanations, e.g. Boone and Piccinini, 2016.

[16] The claim that one may obtain explanatory insights about real-world phenomena from a theory's mathematical formalism without making assumptions about the causal and/or mechanistic underpinnings of these phenomena is often put forward in the literature on mathematical explanations (e.g. Baker, 2009, Baron, 2016, Batterman, 2010, Pincock, 2007). I do not expand on this literature since my defence of thin RCT's explanatory potential does not directly rest on how one conceptualizes the relationship between thin RCT explanations and mathematical explanations.

Second, several accounts that emphasize the explanatory significance of information about causal and/or mechanistic underpinnings allow that one may explain real-world agents' choices *irrespective* of whether one makes accurate assumptions about the causal and/or mechanistic underpinnings of such choices. To give one example, the proponents of the counterfactual account of explanation frequently emphasize the explanatory significance of causal patterns of counterfactual dependence (e.g. Woodward and Hitchcock, 2003). Yet, one may retain these authors' emphasis on patterns of counterfactual dependence without requiring that these patterns be causal (e.g. Reutlinger, 2016, Rice, 2015, Saatsi, 2016). In fact, leading proponents of the counterfactual account allow that one may explain several phenomena without making any empirical assumption concerning the causal and/or mechanistic underpinnings of these phenomena. By way of illustration, Woodward (2003) prevalently discusses causal (rather than non-causal) patterns of counterfactual dependence. Still, he allows that in some cases counterfactual information may be explanatory even if it does not track causal patterns of counterfactual dependence (ibid., ch.5; also Hitchcock and Woodward, 2003, 191, for the claim that "when a theory […] answers a what-if-things-had-been-different question, but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a non-causal explanation"). Moreover, and irrespective of the interpretative issue whether Woodward is plausibly taken to advocate applying his account to non-causal explanations, several authors emphasize the suitability of the counterfactual account to target non-causal patterns of counterfactual dependence (e.g. Baron et al., 2017, Bokulich, 2012, Povich, 2018, Saatsi and Pexton, 2013).

And third, making accurate assumptions about the causal and/or mechanistic underpinnings of choices may *hamper* (rather than enhance) RCTs' ability to explain

choices. My claim is not merely that RCTs may explain some agents' choices *in spite of* considerable ignorance of these choices' causal and mechanistic underpinnings. Rather, my main point is that some thin RCT applications succeed in explaining choices precisely *because* they abstract away from causal and mechanistic underpinnings, and that these applications provide some explanatory insights besides those provided by thick RCT applications that make accurate assumptions about such underpinnings. To see this, let us briefly compare thin RCT with Simon's (1957) theory of bounded rationality (henceforth, BRT), which focuses on the identification of satisfactory (rather than optimal) solutions to agents' choice problems. BRT makes some accurate assumptions about the causal and mechanistic underpinnings of choices (e.g. limitations in agents' cognitive and computational abilities). As a result, BRT applications provide RCTs with causal and mechanistic explanatory insights that are not provided by thin RCT applications (e.g. footnote no.20). Yet, thin RCT applications also provide some explanatory insights besides those provided by BRT applications. For thin RCT's representation theorems cannot be derived within BRT (e.g. Blaug, 1992, part III), and BRT characterizes choice situations "as fundamentally *incomplete* [...] with regard to both the information that economic agents have *and* the agents' cognitive and computational capacities" (Bueno and Colyvan, 2011, 363). These considerations imply neither that thin RCT is in general more explanatory than BRT nor that making accurate assumptions about causal and/or mechanistic underpinnings is irrelevant for explaining choices. Still, the point remains that thin RCT can explain choices, and that thin RCT applications provide some explanatory insights besides those provided by BRT applications.[17]

[17] A critic of thin RCT may grant that thin RCT applications provide some explanatory insights besides those provided by BRT applications, yet object that thick RCT applications typically provide more explanatory insights than thin RCT applications. I am not concerned here with comparing thin RCT applications' and thick RCT applications' overall explanatory contributions. For my evaluation, it suffices to note that if thick RCT applications were shown to provide more explanatory insights than thin RCT applications, this would bear on the justifiability of developing

6. Objection from Axioms' Untenability

The objection from *axioms' untenability* holds that thin RCT cannot explain real-world agents' choices on the alleged ground that the axiomatic requirements underlying thin RCT applications are inconsistent with the available behavioural findings (e.g. Sen, 1985, Sugden, 1991). The objection can be explicated as follows. Various entrenched accounts of explanation require that the explanans provides accurate (or at least approximate) characterizations of the examined targets as a minimum condition for adequate explanation (e.g. Hempel, 1965, Strevens, 2008). In particular, several authors hold that whether thin RCT can explain real-world agents' choices crucially depends on whether these agents' preferences satisfy thin RCT's axiomatic requirements (e.g. Sen, 1973, Sugden, 1991). Unfortunately, the objection goes, these axiomatic requirements are typically motivated on a priori grounds and are rarely tested against the evidence (e.g. Bruni and Sugden, 2007, Sen, 1985). Moreover, the available behavioural findings document that real-world agents' preferences frequently violate such axiomatic requirements (e.g. Loomes et al., 1991, on violations of transitivity, Allais, 1953, on violations of independence). In fact, such findings indicate that real-world agents' choices reflect contextual elements of the choice situation (e.g. framing of choice options) rather than stable and consistent preference orderings. This, in turn, undermines the claim that thin RCT can explain real-world agents' choices. For many choice patterns "are so sensitive to [context] that it is difficult to associate these [choice patterns to any] context-free preferences" (Rabin, 2002, 662; also Tversky, 1996, 189).[18]

thin (rather than thick) RCT applications, but would not directly bear against my claim that thin RCT applications can explain real-world agents' choices (e.g. footnote no.21).

[18] The objection from axioms' untenability targets the putative lack of descriptive (rather than normative) validity of thin RCT's axioms. Some critics of thin RCT doubt not just the descriptive

This objection nicely elucidates often-made complaints concerning the putative untenability of thin RCT's axiomatic requirements. Even so, there are at least two reasons to doubt that the objection undermines the claim that thin RCT can explain real-world agents' choices. First, thin RCT's axiomatic requirements have received more *empirical support* than the objection from axioms' untenability states. To be fair, thin RCT's axiomatic requirements have been shown to be inconsistent with the behavioural findings collected in various choice settings (e.g. Cubitt et al., 2001, Glimcher, 2011, ch.5). Yet, thin RCT's axiomatic requirements fit vast sets of findings across several choice settings (e.g. Gilboa, 2009, Smith, 1991), and the reported experimental violations frequently tend to decrease when agents are given time and incentives to learn during experiments (e.g. Loomes et al., 2003, Ross, 2014). And second, RCTs have modified and even relinquished some of thin RCT's axiomatic requirements so as to *increase* these requirements' fit with the available behavioural findings (e.g. Bhattacharyya et al., 2011, Machina, 2008). In this context, pointing to some empirical findings contrary to specific axiomatic requirements does not *per se* exclude that thin RCT applications which build on these axiomatic requirements can explain real-world agents' choices. For RCTs may be able to provide other empirical findings showing that those axiomatic requirements, while failing to hold in various choice settings, hold in the specific choice settings they target. To illustrate this, consider RCTs' responses to observed violations of independence. For some applications, these violations prompted RCTs to modify or even relinquish independence (e.g. Starmer, 2000). For other applications, RCTs have succeeded in accommodating the reported violations of independence without having to

validity, but also the normative validity of specific axioms (e.g. Sugden, 1991). I do not expand on normative criticisms of thin RCT's axioms because such criticisms do not directly bear on the issue whether thin RCT can explain the choices that are in fact made by real-world agents.

modify or relinquish such axiomatic requirement (e.g. Gilboa, 2009, ch.12, on thin RCT applications that accommodate various reported violations of independence by including reference to social payoffs into the description of choice options; also Guala, 2000, for similar remarks concerning RCTs' ability to accommodate various reported violations of transitivity).[19]

A critic of RCT may grant that thin RCT could in principle explain real-world agents' choices in cases where thin RCT's axiomatic requirements are shown to hold in the specific choice settings targeted by RCTs. However, she may object that thin RCT cannot explain real-world agents' choices because this theory cannot explain *why* its own axiomatic requirements hold (or fail to hold) in any given choice setting (e.g. Okasha, 2016). The idea would be that the explanantia figuring in thin RCT applications are left entirely unexplained by thin RCT, and that this precludes thin RCT from providing adequate explanations of choices. Suppose, for the sake of argument, that thin RCT applications leave it entirely unexplained why thin RCT's axiomatic requirements hold (or fail to hold) in any given choice setting. This arguably shows that thin RCT provides *partial* (as opposed to *complete*) explanations of real-world agents' choices. Even so, the partial character of thin RCT's explanations does not *per se* bear against the claim that thin RCT can explain real-world agents' choices. For most explanations are partial in the sense that they presuppose some unexplained explanans on pain of explanatory regress

[19] A critic of thin RCT may grant that thin RCT's axiomatic requirements hold across several choice settings, yet object that RCTs cannot determine whether specific axioms hold in the choice settings they target unless they make psychological assumptions (e.g. Sen, 1973, 243, on RCTs' alleged need to take a "peep into the head of the [agent]"; also Bhattacharyya et al., 2011, 145, for the claim that RCTs cannot "conclude whether or not [an] agent satisfies [specific axiomatic requirements] simply on the basis of [their] observations of the agent's choice behaviour without referring to the [agent's preferences]"). I mention these claims in passing because such claims are premised on psychological interpretations of preferences and so do not provide independent reasons to doubt the explanatory potential of RCT applications grounded on non-psychological (e.g. thin) interpretations of preferences (e.g. Hands, 2013, Thoma, 2019).

(e.g. Povich, 2018). And the mere fact that there are explanatory tasks that some theory does not address falls short of indicating that this theory lacks explanatory power altogether (e.g. Strevens, 2013). In particular, thin RCT applications may enable RCTs to explicate why agents whose preferences have specific structural properties make some choices (rather than others) even if these applications do not answer the question why those agents' preferences have such structural properties (rather than others). To put it differently, a proposed thin RCT explanation of choices is not rendered explanatorily empty just because it does not answer the question why its own axiomatic requirements hold (or fail to hold) in the examined choice settings.[20]

7. Objection from Interdisciplinary Consilience

The objection from *interdisciplinary consilience* holds that thin RCT cannot explain real-world agents' choices on the alleged ground that thin RCT's axioms entail false neuro-psychological hypotheses (e.g. Craver and Alexandrova, 2008, Quartz, 2008). This objection targets not so much the alleged inconsistency of thin RCT's *axioms* with the available *behavioural* findings, but rather the alleged inconsistency of these axioms' *implications* with the available *neuro-psychological* findings. For this reason, pointing to the alleged fact that thin RCT's axioms are consistent with the available behavioural findings (Section 6) would not address the objection from interdisciplinary consilience. Similarly, it would be of little import to observe that thin RCT's axioms are not meant to accurately fit neuro-psychological findings (Section 5). For although thin RCT's axioms

[20] This obviously does not exclude that neuro-psychological findings may help RCTs to explain why specific thin RCT's axiomatic requirements hold (or fail to hold) in specific settings (e.g. Guala, 2019, Padoa-Schioppa, 2011, on transitivity). In this respect, thick RCT applications can nicely complement thin RCT applications, helping RCTs to ascertain whether the explanatory insights they obtain in specific choice settings can be plausibly expected to hold in other choice settings.

are not meant to accurately fit neuro-psychological findings, these axioms' implications can in principle contradict such findings. Paraphrasing Quartz, while thin RCT does not make any "explicit assumptions about the physiology of the brain", the implications of this theory's axioms "could either be confirmed or disconfirmed by neuroscience" (2008, 463; also Hausman, 2008, for similar claims about psychological findings).

This objection nicely elucidates how considerations of interdisciplinary consilience may bear on the plausibility of specific thin RCT's axioms. Still, there are at least two reasons to doubt that the objection undermines the claim that thin RCT can explain real-world agents' choices. First, it is dubious that thin RCT entails *any* specific hypotheses concerning the neuro-psychological substrates of choice. Hence, it is even more dubious that thin RCT entails *false* hypotheses concerning such substrates. To be sure, the critics of thin RCT often complain that this theory is inconsistent with generic hypotheses about what cognitive and computational abilities agents would have to possess to make choices that fit thin RCT's axioms (e.g. Craver and Alexandrova, 2008, on human individuals' inability to perform the computations allegedly required to satisfy thin RCT's axioms). Yet, the critics have hitherto failed to specify *exactly what* false neuro-psychological hypotheses would be entailed by thin RCT. Indeed, it is not even clear what computations (e.g. one-stage versus multi-stage integrations of reward values) agents would have to execute to satisfy a literal interpretation of thin RCT, and whether such computations would involve psychological processes, neural processes or both (e.g. Fumagalli, 2019). And second, one may consistently ascribe great relevance to considerations of interdisciplinary consilience and yet deny that compatibility with *all* neuro-psychological findings is a *necessary* requirement for explaining choices. In fact, it would be overly restrictive to impose this requirement on RCTs' attempts to explain choices. For RCTs can frequently explain choices by building theories whose implications neglect or fail to

fit specific subsets of these findings (e.g. Fumagalli, 2014, on various neural findings). That is to say, even if thin RCT's axioms entailed some false neuro-psychological hypotheses, this would not *per se* imply that thin RCT cannot explain real-world agents' choices.

A critic of RCT may grant that thin RCT could in principle explain real-world agents' choices even if thin RCT's axioms entailed some false neuro-psychological hypotheses. However, she may object that thin RCT cannot explain real-world agents' choices on the alleged ground that the best available neuro-psychological findings entail that thin RCT is untenable or false (e.g. Camerer, 2008). The thought would be that the best available neuro-psychological findings do not contradict thin RCT *directly*, but undermine thin RCT by contradicting hypotheses that, if true, would support such theory (e.g. Fehr and Rangel, 2011). Let us assume, for the sake of argument, that the best available neuro-psychological findings contradict hypotheses that, if true, would support thin RCT. This result is shown to undermine thin RCT only if one demonstrates that thin RCT holds *just* in case the hypotheses putatively contradicted by neuro-psychological findings hold. However, it is highly dubious that the critics of thin RCT are able to demonstrate that thin RCT holds just in case such hypotheses hold. For many different sets of neuro-psychological processes can generate choice patterns that satisfy thin RCT's axiomatic requirements (e.g. Vromen, 2010), and whether a given neuro-psychological process generates choice patterns that satisfy these requirements may crucially depend on which other neuro-psychological processes interact with such process (e.g. Muldoon and Bassett, 2016). In this respect, it is telling that leading critics of thin RCT acknowledge that this theory "cannot be falsified by the observation that the algorithmic structure of the human brain is incompatible with [the] computations [putatively posited by such theory]" (Glimcher et al., 2007, 145).

By way of illustration, consider the much-discussed hypothesis that expected utility is *literally* computed in the brain by specific neural areas (e.g. Glimcher, 2011, ch.12-15, on dopaminergic activations in the ventral striatum and the medial prefrontal cortex). In recent years, several studies have suggested that activations in anatomically delimited neural areas track rewards' subjective values during decision-making across multiple reward types (e.g. Padoa-Schioppa, 2011). *Prima facie*, these findings might seem to vindicate a literal interpretation of thin RCT. Indeed, such findings have been claimed to "provide unambiguous evidence that […] preference functions are part of the neural mechanism for choice, as opposed to being a purely descriptive ('as if') theoretical construct" (Kable and Glimcher, 2007, 1625). Nonetheless, the critics of thin RCT have hitherto failed to demonstrate that thin RCT holds just in case expected utility is literally computed in the brain by specific neural areas (e.g. Fumagalli, 2019). Moreover, there are several reasons to doubt that thin RCT holds just in case expected utility is literally computed in the brain by specific neural areas.

To give one example, an individual may in principle make choices that fit thin RCT's axiomatic requirements even if no anatomically delimited set of neural areas in her brain undergo activations that reliably fit thin RCT's axiomatic requirements. In fact, various choice procedures can generate choice patterns that fit thin RCT's axiomatic requirements even if no anatomically delimited set of neural areas undergo activations that reliably fit such requirements (e.g. Bernheim, 2009, on sequential choice procedures). Conversely, the alleged fact that the activation patterns of particular neural areas reliably fit thin RCT's axiomatic requirements by no means guarantees that the involved individual makes choices that fit such requirements. For areas whose activations fit thin RCT's axiomatic requirements frequently interact with areas undergoing dissimilar activations

(e.g. Berridge and O'Doherty, 2014), with violations of thin RCT's axiomatic requirements possibly resulting at the level of choices (e.g. Fumagalli, 2013). And the neural substrates of reward valuation appear to be highly distributed, with different kinds of value-related signals being computed in the human brain and several areas being involved in integrating such signals (e.g. Rushworth et al., 2012).[21]

8. Conclusion

Over the last few decades, a number of authors have built on the contrast between thick and thin applications of RCT to argue that thin RCT lacks the potential to explain the choices of real-world agents. In this paper, I have drawn on several often-cited RCT applications to demonstrate that despite this prominent critique there are at least two different senses in which thin RCT can explain such choices. I then defended this thesis against the most influential objections put forward by the critics of RCT. My thesis that thin RCT can explain real-world agents' choices in the two senses explicated in this paper implies neither that thin RCT is in general more explanatory than thick RCT nor that empirical findings from neuro-psychology are irrelevant for explaining choices. Still, if my thesis is correct, the often-made claim that thin RCT cannot explain real-world agents'

[21] A critic of thin RCT may grant that some thin RCT applications can explain real-world agents' choices, yet object that devoting intellectual resources to develop thin (rather than thick) RCT applications is not an efficient use of intellectual resources because the explanatory contribution of thin RCT applications is too limited to justify RCTs' focus on such applications (e.g. Northcott and Alexandrova, 2015). This objection bears on the justifiability of developing thin (rather than thick) RCT applications, but does not directly bear against my claim that thin RCT can explain real-world agents' choices. Moreover, showing that the explanatory contribution of thin RCT applications is too limited to justify RCTs' focus on such applications would require the critics of thin RCT to explicate what conception of explanation they presuppose, what measures of modelling costs and benefits they adopt, and how such modelling costs and benefits are to be traded-off. In fact, even showing that the explanatory contribution of thin RCT applications is too limited to justify RCTs' focus on such applications would not exclude that RCTs may justifiably devote intellectual resources to develop thin RCT applications. For thick and thin RCT applications do not always compete, and RCTs can acquire valuable explanatory insights by combining thick and thin RCT applications (e.g. footnote no.20).

choices is mistaken, and thin RCT has a greater explanatory potential than its critics maintain.

REFERENCES

Alexander, J. 2010. Local Interactions and the Dynamics of Rational Deliberation. *Philosophical Studies*, 147, 102-121.

Alexandrova, A. and Northcott, R. 2013. It's just a feeling: why economic models do not explain. *Journal of Economic Methodology*, 20, 262-267.

Allais, M. 1953. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, 21, 503-546.

Baker, A. 2009. Mathematical Explanation in Science. *British Journal for the Philosophy of Science*, 60, 611-633.

Baron, S. 2016. Explaining Mathematical Explanation. *The Philosophical Quarterly*, 66, 458-480.

Baron, S., Colyvan, M. and Ripley, D. 2017. How mathematics can make a difference. *Philosophers' Imprint*, 17, 1-19.

Batterman, R. 2010. On the Explanatory Role of Mathematics in Empirical Science. *British Journal for the Philosophy of Science*, 61, 1-25.

Batterman, R. and Rice, C. 2014. Minimal Model Explanations. *Philosophy of Science*, 81, 349-376.

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

Becker, G. 1996. *Accounting for Tastes*. Cambridge: Harvard University Press.

Bernheim, B. 2009. On the potential of neuroeconomics: a critical (but hopeful) appraisal. *American Economic Journal: Microeconomics*, 1, 1-41.

Berridge, K.C. and O'Doherty, J.P. 2014. From Expected Utility to Decision Utility. In *Neuroeconomics: Decision Making and the Brain*. 2nd Ed., Edited by Glimcher, P. and Fehr, E. 335-354. Elsevier: Academic Press.

Bhattacharyya, A., Pattanaik, K. and Xu, Y. 2011. Choice, internal consistency and rationality. *Economics and Philosophy*, 27, 123-149.

Blaug, M. 1992. *The methodology of economics - or how economists explain*. 2nd Ed. Cambridge University Press.

Bokulich, A. 2012. Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science*, 79, 725-737.

Bokulich, A. 2014. How the tiger bush got its stripes: 'How possibly' vs 'how actually' model explanations. *The Monist*, 97, 321-338.

Boone, W. and Piccinini, G. 2016. Mechanistic Abstraction. *Philosophy of Science*, 83, 686-697.

Bossert, W., Sprumont, Y. and Suzumura, K. 2006. Rationalizability of choice functions on general domains without full transitivity. *Social Choice and Welfare*, 27, 435-458.

Boudon, R. 2003. Beyond rational choice theory. *Annual Review of Sociology*, 29, 1-21.

Bradley, R. 2017. *Decision Theory with a Human Face*. Cambridge University Press.

Bruni, L. and Sugden, R. 2007. The road not taken: how psychology was removed from economics and how it might be brought back. *The Economic Journal*, 117, 146-173.

Bueno, O. and Colyvan, M. 2011. An Inferential Conception of the Application of Mathematics. *Nous*, 45, 345-374.

Camerer, C.F. 2008. The Potential of Neuroeconomics. *Economics and Philosophy*, 24, 369-379.

Caplin, A., Dean, M., Glimcher, P. and Rutledge, R.B. 2010. Measuring beliefs and rewards: A neuroeconomic approach. *Quarterly Journal of Economics*, 125, 923-960.

Craver, C.F. 2006. When mechanistic models explain. *Synthese*, 153, 355-376.

Craver, C.F. and Alexandrova, A. 2008. No revolution necessary: neural mechanisms for economics. *Economics and Philosophy*, 24, 381-406.

Cubitt, R., Starmer, C. and Sugden, R. 2001. Discovered Preferences and the Experimental Evidence of Violations of Expected Utility Theory. *Journal of Economic Methodology*, 8, 385-414.

Dietrich, F. and List, C. 2016. Mentalism Versus Behaviourism in Economics: a Philosophy-of-Science Perspective. *Economics and Philosophy*, 32, 249-281.

Fehr, E. and A. Rangel. 2011. Neuroeconomic Foundations of Economic Choice - Recent Advances. *Journal of Economic Perspectives*, 25, 3-30.

Ferejohn, J. 2002. Symposium on explanations and social ontology 1: Rational choice theory and social explanation. *Economics and Philosophy*, 18, 211-234.

Ferejohn, J. and Satz, D. 1995. Unification, universalism, and rational choice theory. *Critical Review*, 9, 1-2, 71-84.

Fishburn, P. 1970. Intransitive Indifference in Preference Theory: A Survey. *Operations Research*, 18, 207-228.

Forber, P. 2010. Confirmation and explaining how possible. *Studies in History and Philosophy of Science*, 41, 32–40.

Forber, P. 2012. Conjecture and explanation: a reply to Reydon. *Studies in History and Philosophy of Science*, 43, 298-301.

Friedman, M. 1974. Explanation and Scientific Understanding. *The Journal of Philosophy*, 71, 5-19.

Fumagalli, R. 2013. The Futile Search for True Utility. *Economics and Philosophy*, 29, 325-347.

Fumagalli, R. 2014. Neural Findings and Economic Models: Why Brains have Limited Relevance for Economics. *Philosophy of the Social Sciences*, 44, 606-629.

Fumagalli, R. 2016. Why we cannot Learn from Minimal Models. *Erkenntnis*, 81, 433-455.

Fumagalli, R. 2019. (F)utility Exposed. *Philosophy of Science*. In Press.

Gaertner, W. and Xu, Y. 1999. On the structure of choice under different external references. *Economic Theory*, 14, 609–620.

Gibbard, A. and Varian, H.R. 1978. Economic Models. *The Journal of Philosophy*, 75, 664-677.

Gilboa, I. 2009. *Theory of Decision Under Uncertainty*. Cambridge University Press.

Glennan, S. 2005. Modeling Mechanisms. *Studies in History and Philosophy of Science*, 36, 443-464.

Glimcher, P. 2011. *Foundations of Neuroeconomic Analysis*. Oxford University Press.

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

Green, D.P. and Shapiro, I. 1994. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven: Yale University Press.

Guala, F. 2000. The logic of normative falsification: Rationality and experiments in decision theory. *Journal of Economic Methodology*, 7, 59-93.

Guala, F. 2006. Has game theory been refuted? *The Journal of Philosophy*, 103, 239-263.

Guala, F. 2012. Are preferences for real? Choice theory, folk psychology, and the hard case for commonsensible realism. In A. Lehtinen, et al. (Eds.), *Economics for real: Uskali Mäki and the place of truth in economics*, 137-155. Routledge.

Guala, F. 2019. Preferences: Neither Behavioural Nor Mental. *Economics and Philosophy*, 35, 383-401.

Hands, W. 2006. Individual psychology, rational choice, and demand. *Revue de Philosophie Economique*, 13, 3-48.

Hands, W. 2013. Foundations of Contemporary Revealed Preference Theory. *Erkenntnis*, 78, 1081-1108.

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

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

Hausman, D.M. 2011. Mistakes about preferences in the social sciences. *Philosophy of the Social Sciences*, 41, 3-25.

Hausman, D.M. 2012. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Hempel, C. 1965. *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hindriks, F. 2013. Explanation, understanding, and unrealistic models. *Studies in History and Philosophy of Science*, 44, 523-531.

Hitchcock, C. and Woodward, J. 2003. Explanatory generalizations, Part II: Plumbing explanatory depth. *Nous*, 37, 181-199.

Hodgson, G. 2012. On the Limits of Rational Choice Theory. *Economic Thought*, 1, 94-108.

Kable, J. and Glimcher, P. 2007. The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10, 1625-1633.

Kahneman, D., Slovic, P. and Tversky, A. 1982. *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kaplan, D.M. and Craver, C.F. 2011. The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, 78, 601-627.

Kitcher, P. 1981. Explanatory Unification. *Philosophy of Science*, 48, 507-531.

Kitcher, P. 1989. Explanatory Unification and the Causal Structure of the World. *In Kitcher, P. and Salmon, W. Eds. Minnesota Studies in the Philosophy of Science*, 13, 410-505. University of Minnesota Press.

Kuorikoski, J. and Lehtinen, A. 2010. Economics Imperialism and Solution Concepts in Political Science. *Philosophy of the Social Sciences*, 40, 347-374.

Kuorikoski, J. and Ylikoski, P. 2010. Explanatory relevance across disciplinary boundaries: the case of neuroeconomics. *Journal of Economic Methodology*, 17, 219-228.

Kuorikoski, J. and Ylikoski, P. 2015. External Representations and Scientific Understanding. *Synthese*, 192, 3817-3837.

Lazear, E. 2000. Economic Imperialism. *Quarterly Journal of Economics*, 115, 99-146.

Lehtinen, A. 2013. Three kinds of 'as-if' claims. *Journal of Economic Methodology*, 20, 184-205.

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

Lewis, D. 1986. Causal Explanation. In Lewis, D. *Philosophical Papers, Vol.II*, 214-240. Oxford University Press.

Lindenberg, S. 1992. The Method of Decreasing Abstraction. In *Rational Choice Theory. Advocacy and Critique*. Coleman, J. and Fararo, T. (Eds.), ch.1. London: Sage Publications.

Loomes, G., Starmer, C. and Sugden, R. 1991. Observing violations of transitivity by experimental methods. *Econometrica*, 59, 425-439.

Loomes, G., Starmer, C. and Sugden, R. 2003. Do Anomalies Disappear in Repeated Markets? *The Economic Journal*, 113, C153-C166.

Loomes, G. and Sugden, R. 1982. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92, 805-824.

Lyon, A. 2012. Mathematical explanations of empirical facts, and mathematical realism. *Australasian Journal of Philosophy*, 90, 559-578.

Machina, M. 1982. Expected Utility Analysis without the Independence Axiom. *Econometrica*, 50, 277-323.

Machina, M. 2008. Non-expected Utility Theory. In *The New Palgrave Dictionary of Economics*. 2nd Ed., Durlauf, S. and Blume, L., 74-84. New York: Palgrave Macmillan.

Mäki, U. 1992. On the method of isolation in economics. *Poznan Studies in the Philosophy of Science and the Humanities*, 26, 316-351.

Mäki, U. 2001. Explanatory unification: double and doubtful. *Philosophy of the Social Sciences*, 31, 488-506.

Mäki, U. 2011. Models and the locus of their truth. *Synthese*, 180, 47–63.

Mäki, U. 2013. Scientific Imperialism: Difficulties in Definition, Identification, and Assessment. *International Studies in the Philosophy of Science*, 27, 325-339.

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

Marchionni, C. 2018. What is the problem with model-based explanation in economics? *Disputatio*, 9, 603-630.

McMullin, E. 1985. Galileian Idealization. *Studies in History and Philosophy of Science*, 16, 247-273.

Morgan, M. 2001. Models, stories and the economic world. *Journal of Economic Methodology*, 8, 361-384.

Morgan, M. 2006. Economic man as model man: ideal types, idealization and caricatures. *Journal of the History of Economic Thought,* 28, 1-27.

Muldoon, S. and Bassett, D. 2016. Network and multilayer network approaches to understanding human brain dynamics. *Philosophy of Science*, 83, 710-720.

Northcott, R. and Alexandrova, A. 2015. Prisoner's dilemma doesn't explain much. In Martin Peterson (Ed.). *The Prisoner's Dilemma*, 64-84. Cambridge University Press.

Nowak, M. 2006. Five rules for the evolution of cooperation. *Science*, 314, 1560-1563.

Okasha, S. 2016. On the interpretation of decision theory. *Economics and Philosophy*, 32, 409-433.

Padoa-Schioppa, C. 2011. Neurobiology of Economic Choice: A Good-Based Model. *Annual Review of Neuroscience*, 34, 333-359.

Pettit, P. 1995. The Virtual Reality of *homo economicus*. *Monist*, 78, 308-329.

Pincock, C. 2007. A role for mathematics in the physical sciences. *Nous*, 41, 253-275.

Pincock, C. 2015. Abstract Explanations in Science. *British Journal for the Philosophy of Science,* 66, 857-882.

Pollak, R. 2003. Gary Becker's Contributions to Family and Household Economics. *Review of Economics of the Household*, 1, 111-141.

Povich, M. 2018. Minimal Models and the Generalized Ontic Conception of Scientific Explanation. *British Journal for the Philosophy of Science*, 69, 117-137.

Quartz, S.R. 2008. From cognitive science to cognitive neuroscience to neuroeconomics. *Economics and Philosophy*, 24, 459-471.

Rabin, M. 2002. A perspective on psychology and economics. *European Economic Review*, 46, 657-685.

Reiner, R. 1993. Necessary Conditions and Explaining How-Possibly. *Philosophical Quarterly*, 44, 58-69.

Reiss, J. 2012. The explanation paradox. *Journal of Economic Methodology*, 19, 43-62.

Resnik, D. 1991. How-Possibly Explanations in Biology. *Acta Biotheoretica*, 39, 141-149.

Reutlinger, A. 2016. Is There A Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science*, 83, 733-745.

Rice, C. 2012. Optimality Explanations: A Plea for an Alternative Approach. *Biology and Philosophy*, 27, 685-703.

Rice, C. 2015. Moving Beyond Causes: Optimality Models and Scientific Explanation. *Nous*, 49, 589-615.

Roche, W. and Sober, E. 2017. Explanation = Unification? A New Criticism of Friedman's Theory and a Reply to an Old One. *Philosophy of Science*, 84, 391-413.

Rohwer, Y. and Rice, C. 2013. Hypothetical Pattern Idealization and Explanatory Models. *Philosophy of Science*, 80, 334-355.

Rohwer, Y. and Rice, C. 2016. How are Models and Explanations Related? *Erkenntnis*, 81, 1127-1148.

Ross, D. 2014. Psychological versus Economic Models of Bounded Rationality. *Journal of Economic Methodology,* 21, 411-427.

Ross, L.N. 2015. Dynamical Models and Explanation in Neuroscience. *Philosophy of Science*, 82, 32-54.

Rushworth, M., Kolling, N., Sallet, J. and Mars, R. 2012. Valuation and decision-making in frontal cortex: one or many serial or parallel systems? *Current Opinion in Neurobiology*, 22, 946-955.

Saatsi, J. 2011. The Enhanced Indispensability Argument: Representational versus Explanatory Role of Mathematics in Science. *British Journal for the Philosophy of Science*, 62, 143-154.

Saatsi, J. 2016. On the 'Indispensable Explanatory Role' of Mathematics. *Mind*, 125, 1045-1070.

Saatsi, J. and Pexton, M. 2013. Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations. *Philosophy of Science*, 80, 613-624.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Samuelson, P. 1938. A Note on the Pure Theory of Consumer's Behavior. *Economica*, 5, 61-71.

Satz, D. and Ferejohn, J. 1994. Rational Choice and Social Theory. *The Journal of Philosophy*, 91, 71-87.

Sen, A. 1973. Behaviour and the concept of preference. *Economica*, 40, 241-259.

Sen, A. 1985. Rationality and uncertainty. *Theory and Decision*, 18, 109-127.

Sen, A. 1987. Rational behavior. In *The New Palgrave: A Dictionary of Economics*, Eatwell J., Milgate M., Newman P. (eds.), 68-76. London, Macmillan.

Simon, H. 1957. *Models of Man: Social and Rational*. New York: John Wiley & Sons.

Skyrms, B. 1990. *The Dynamics of Rational Deliberation*. Harvard University Press.

Smith, V. 1991. Rational choice: the contrast between economics and psychology. *Journal of Political Economy*, 99, 877-897.

Sober, E. 1983. Equilibrium Explanation. *Philosophical Studies*, 43, 201-210.

Starmer, C. 2000. Developments in nonexpected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 3, 38, 332-382.

Strevens, M. 2008. *Depth: an account of scientific explanation*. Cambridge: Harvard University Press.

Strevens, M. 2013. No understanding without explanation. *Studies in History and Philosophy of Science*, 44, 510-515.

Sugden, R. 1991. Rational Choice: A Survey of Contributions from Economics and Philosophy. *The Economic Journal*, 101, 751-778.

Sugden, R. 1993. An axiomatic foundation for regret theory. *Journal of Economic Literature*, 60, 159-180.

Sugden, R. 2011. Explanations in search of observations. *Biology and Philosophy*, 26, 717-736.

Sugden, R. 2013. How fictional accounts can explain. *Journal of Economic Methodology*, 20, 237-243.

Thoma, J. 2019. Folk Psychology and the Interpretation of Decision Theory. Manuscript Under Review.

Tversky, A. 1996. Rational theory and constructive choice. In Arrow, K, Colombatto, E., Perlman, M. and Schmidt, C. Eds. *The rational foundations of economic behaviour*, 185-197. Basingstoke: Macmillan Press.

Verreault-Julien, P. 2019. How could models possibly provide how-possibly explanations? *Studies in History and Philosophy of Science*, 73, 22-33.

Von Neumann, J. and Morgenstern, O. 1947. *Theory of Games and Economic Behavior*, 2nd Ed. Princeton University Press.

Vredenburgh, K. 2019. A unificationist defence of revealed preferences. *Economics and Philosophy*. In Press.

Vromen, J. 2010. On the surprising finding that expected utility is literally computed in the brain. *Journal of Economic Methodology*, 17, 17-36.

Weisberg, M. 2006. Robustness Analysis. *Philosophy of Science*, 73, 730-742.

Weisberg, M. 2007. Three Kinds of Idealization. *The Journal of Philosophy*, 104, 639-659.

Weiskopf, D. 2011. Models and mechanisms in psychological explanation. *Synthese*, 183, 313-338.

Woodward, J. 2003. *Making Things Happen. A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. and Hitchcock, C. 2003. Explanatory generalizations, Part I: A Counterfactual Account. *Nous*, 37, 1-24.

Ylikoski, P. 2013. Causal and Constitutive Explanation Compared. *Erkenntnis*, 78, 277-297.

Ylikoski, P. and Aydinonat, E. 2014. Understanding with Theoretical Models. *Journal of Economic Methodology*, 21, 19-36.

Ylikoski, P. and Kuorikoski, J. 2010. Dissecting explanatory power. *Philosophical Studies*, 148, 201-219.