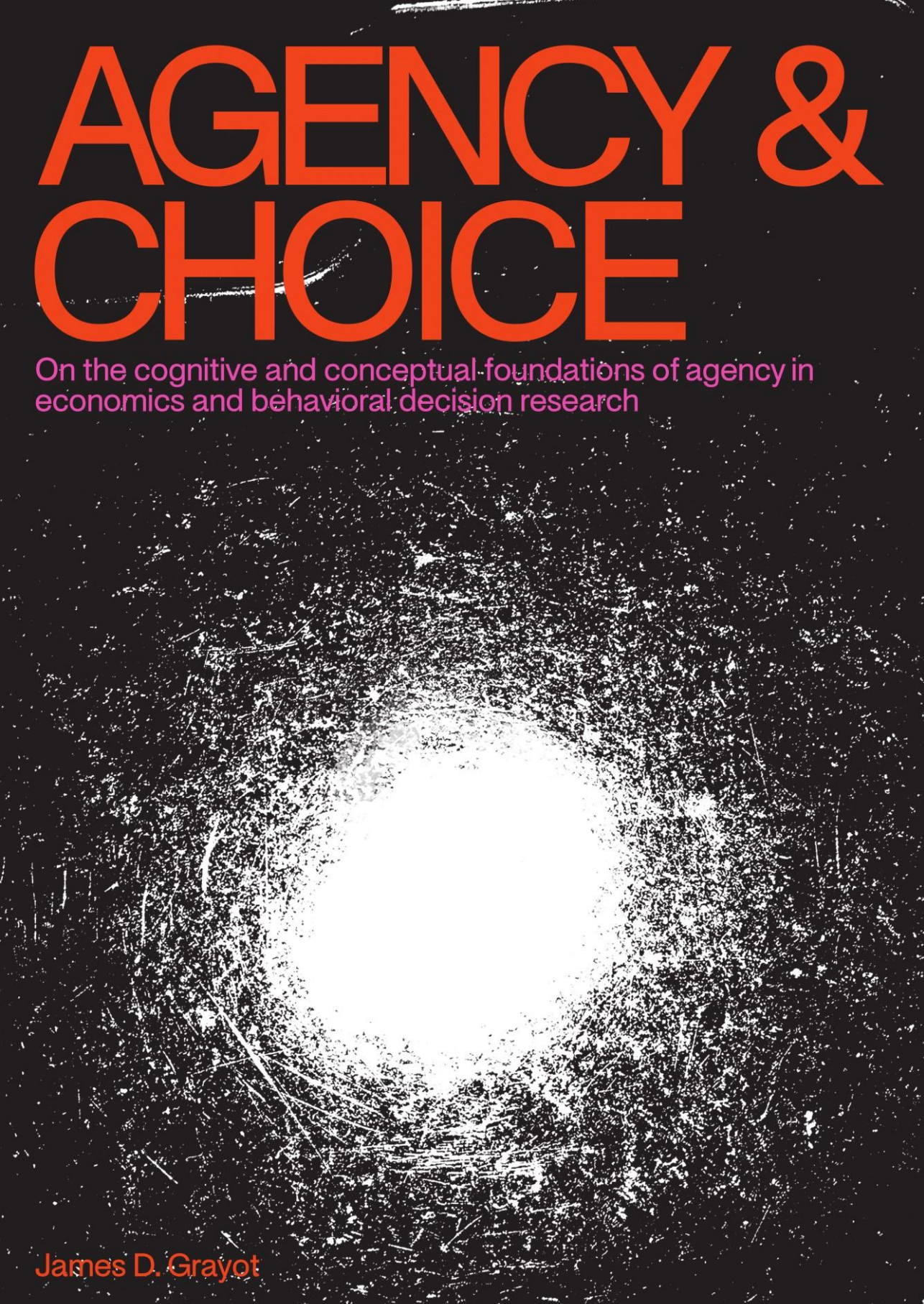
AGENCY & CHOICE

On the cognitive and conceptual foundations of agency in economics and behavioral decision research



James D. Grayot

James D. Grayot

Agency & Choice

On the cognitive and conceptual foundations of agency in economics and behavioral decision research

PhD thesis—Erasmus University Rotterdam

SUPERVISORS
Prof.dr. Jack Vromen
Dr. Conrad Heilmann

Agency and Choice:
On the cognitive and conceptual foundations of agency
in economics and behavioral decision research

Actorschap en keuze:
De cognitieve en conceptuele fundamenten van actorschap in
economisch en beslissingsonderzoek

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
Rector Magnificus

Prof.dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on
Friday the 8th of February at 13:30 hrs.

by

James Daniel Grayot
born in California, United States of America

Erasmus University Rotterdam

Doctoral Committee:

Promotors:

Prof.dr. J.J. Vromen
Dr. H.C.K. Heilmann

Other members:

Prof.dr. K.I.M. Rohde
Prof.dr. D. Ross
Prof.dr. F. Guala

James D. Grayot
Agency & Choice

ISBN:
Cover design by Moritz Oberberger.

CONTACT:
james.grayot@gmail.com

Table of Contents

Preface

How I ended up in the Netherlands and how this thesis came to exist are difficult to explain. My choice to pursue a PhD wasn't based on a single decision or definitive life-event: it was something that took shape slowly, that emerged from moves in seemingly different directions. Even with the benefit of hindsight, it feels less like a narrative of my own design and more like a series of moments that happen to include me. Verily, I owe my disordered success to two qualities that shouldn't but nevertheless coexist in me: pathological curiosity and persistent skepticism. This inspiring and frustrating duality is the reason I chose to be an academic and it's likely the reason I've committed myself to studying how agency and choice relate to the science of decision making.

My introduction to academic philosophy was through the study of language, not economics. As a bachelor student at Humboldt State University, I was exposed to the writings of Wittgenstein, Ryle, Austin, and Ebersole, and to the subtle but caustic methods of Ordinary Language Philosophy. I am grateful to John W. Powell who mentored me as a young student, and who taught me to be wary of conventional philosophical inquiry. For better or worse, my own skepticism is modeled after John's.

At San Jose State University I met Anand Vaidya, a tenacious logician who (re)introduced me to the wonders of logic and to the varieties of modal knowledge. As a master student, my skepticism subsided (somewhat) and my appreciation for traditional analytic philosophy improved. This was, to be sure, a strange inversion of intellectual progress! I did research on the metaphysics of necessity, on the plurality of logical systems, and on the cognitive foundations of mathematical and logical knowledge. I expected that these subjects would be the topic of my graduate thesis and the focus of my academic career.

This abruptly changed when I joined a reading group in the final semester of my master's degree. The subject was theories of justice and ethical aspects of economics. This was a revelatory moment as I'd never seriously considered how human reasoning relates to economics. (Admittedly, I'd always loathed economics and avoided it during my bachelor studies). It was here that I began to critically explore decision-theoretic concepts like utility, preference, and rational choice, and became acquainted with the political and, more importantly, the psychological implications of preference measurement. For reasons I still don't grasp, I shifted my emphasis from logic and language to decision making, and began fervently studying economics.

It was Anand who encouraged me to apply to the Erasmus Institute for Philosophy of Economics (EIPE). For that recommendation—and so much more—I am thankful to him.

Having lived in Rotterdam for six years, and having become a member of the EIPE community, I see now that my own story is not all that unique, which is to say, that EIPE is a wonderfully diverse place—both intellectually and culturally. I've had the good fortune to not only work with some of the best and brightest minds, but I've forged professional relationships that I trust to carry me into the future. Among those best and brightest minds are my PhD supervisors, Jack Vromen and Conrad Heilmann (and Constanze Binder for her supervision during my research master), and my inner committee members, Don Ross, Francesco Guala, and Kirsten Rohde. I am indebted to each of you.

To Jack and Conrad: Your respective supervising styles complimented one another, and in this way, were ideal for someone like me. Jack, you regularly stoked my curiosity and encouraged me to pursue research on topics that others have found too difficult or too fringe. This support allowed me to trek deeper into interdisciplinary territory and satisfy my own, often ambling, curiosity. Yet, your own shrewd approach toward philosophical analysis kept me grounded. It is my hope that this thesis meets your very high standards of conceptual clarity. Conrad, time and again you exceeded the call of supervisory duty. Not only did you make yourself available around the clock (to cater to my obscene hours), but your comments were often more detailed than the original manuscript I sent. You've raised project management to an art form, and I've reaped all the benefits of it. I cannot thank you enough for your commitment to my thesis, for the many hours you've spent sifting through my scattered notes, and for your council during many fraught moments. I could not have completed this without your help.

To my fellow PhDs: Osman Çağlar Dede, Akshath Jitendranath, Willem van der Deijl, Philippe Verreault-Julien, Vaios Koliofotis, Huub Brouwer, Daphne Truijens, Melissa Vergara-Fernandez, Attilia Ruzzene, Jasper van der Herik. As a group, you are a sterling example of why EIPE and the Erasmus School of Philosophy are first-class research institutes. You are some of the finest people I know, and it is because of your support and friendships that I have thrived in Rotterdam. In particular, I'm thankful to Çağlar (Dede), for standing by my side throughout the research master and PhD programs (and for agreeing to stand by my side one more time during the PhD defense). I've spent more time with you than anyone over the last six years. You're my brother and probably one of the silliest, most genuine people I know.

Finally, my life Rotterdam would not be the same if it weren't for a select group of people, and thus, this preface would be incomplete if I didn't acknowledge them: Cristina Silva, Ovidiu Stanciu, Yiannis Tsoskonouglou—it is because of you three that I can call Rotterdam my second home. Few people outside our circle can really understand or appreciate what life on 1e Middellandstraat was like. It was many things, but most of all it was (and still is) an adventure. I know that whatever the future

brings, you three will be there, supporting me. Cristina, you especially have been a rock for me. Your infectious optimism is what has keep me afloat all these years.

To Mom, Dad, Katie, Curtis, (and Bryan): I am who I am because of each of you: Thank you for teaching me patience, integrity, work ethic, wit, (and spontaneity).

To Ashley: Thank you for putting up with me these last few years. You are my best friend and partner in crime.

Το Zoi: Σε ευχαριστώ που με εμπνέεις καθημερινά. Δεν ξέρεις πόσο ευτυχισμένο με κάνεις.

Chapter 1
Introduction

1. Agency and choice

Much of economics is devoted to the study of choice and the consequences of choices; in fact, rational choice theory forms one of the building blocks of microeconomic theory. Choice can be understood in a variety of forms and analyzed from a variety of perspectives (individual, temporal, interactive, collective, aggregated, and so on). In studying choices, economists have formulated—sometimes implicitly and sometimes explicitly—different views about the concept of *agency* and its relationship to rational choice. Naturally, concepts such as agency and rationality are also systematically studied by other disciplines, most notably philosophy, psychology, and cognitive science. These other disciplines have often highlighted different aspects of these operative concepts.

Orthodox approaches to economics often portray agents as being fully rational, which means that, (i) people have well-defined preferences and make decisions so as to maximize those preferences, (ii) preferences accurately reflect a person's information about their options, and (iii) people have the ability to update their beliefs about their options in light of changing information. Economists may disagree about the specific requirements underpinning (i) – (iii), but most, if not all, will submit to these criteria as the defining characteristics of economic science.

Of course, people are not fully rational. Decades of experimental research and interdisciplinary collaborations between economists, psychologists, and neuroscientists have produced an unending list of anomalies which serve to challenge orthodox interpretations of economic theory. This research reveals that not only is rationality unreliably demonstrated in human choice and inference, but also that the vast majority of human behavior is driven by automatic rather than controlled, and emotional rather than reflective processes. That individuals are cognitively constrained and prone to systematic errors in thinking and reasoning is now well known as *bounded rationality*. Research in the behavioural decision sciences, and notably in behavioural economics and neuroeconomics, has been developing in sometimes quite close interaction with these interdisciplinary efforts.

In this thesis, I offer a philosophical perspective on the different conceptions of agency and choice as they are understood and employed in economics and behavioral decision research—this perspective is two-fold: on the one hand, philosophical analysis can clarify ambiguities in definitions and concepts that can and do arise within interdisciplinary research. This is of particular importance given how philosophical concepts such as *mind*, *cognition*, and *intentionality* feature in economic studies of

rational choice. Hence, one project of this thesis is to subject contemporary research on questions about agency and choice to such philosophical scrutiny.

On the other hand, the questions and topics discussed in this thesis can be understood as an exercise in philosophy of science: they deal explicitly with questions and topics that pertain to the theoretical and empirical practices of scientists. This includes traditional microeconomic disciplines, such as decision and game theory, as well as interdisciplinary collaborations in behavioral economics, neuroeconomics, and experimental psychology.

2. Economics meets psychology and cognitive science

In recognizing that ordinary humans are boundedly rational, economists and decision researchers who utilize rational choice theory are faced with a difficult choice: one can stick to the standard concepts and tools of orthodox economics and bracket-out decision anomalies which challenge orthodoxy; or one can confront the evidence head-on and modify economic concepts and tools accordingly. Of course, how one reacts to this dilemma will depend on what they interpret the target and underlying units of economic analysis to be. Not surprisingly, opinions have been and remain divided on the (increasing) role of psychology and cognitive science in economics. Consider the following passages:

> Because psychology systematically explores human judgment, behavior, and well-being, it can teach us important facts about how humans differ from the way they are traditionally described by economists. (Rabin, 1998, p. 11)

> Because economics is the science of how resources are allocated by individuals and by collective institutions like firms and markets, the psychology of individual behavior should underlie and inform economics, much as physics informs chemistry; archaeology informs anthropology; or neuroscience informs cognitive psychology. (Camerer, 1999, p. 10575)

It is implied by the first two passages that economics needs psychology, or that it has much to learn from it, because individual persons are centers of decision-making—which is to say, that choices are the outcome of their subjective beliefs and conscious and unconscious desires. It can be inferred from these points that some economists take the concepts of utility and preference to be psychologically real, and they hold out hope that cognitive psychology or neuroscience can illuminate where and/or how these concepts are realized. Hence, even if persons are not ideally or systematically rational, perhaps some part of them—or their brains—is.

Now consider the following:

> Neuroscience evidence cannot refute economic models because the latter make no assumptions or draw no conclusions about physiology of the brain. Conversely, brain science cannot revolutionize economics because it has no vehicle for addressing the concerns of the latter. Economics and psychology differ in the question they ask. Therefore, abstractions that are useful for one discipline will typically be not very useful for the other. (Gul & Pesendorfer, 2008, p. 4)

> That economic agents and people have different properties should strike no one as surprising. Whereas people are pre-theoretical entities found in the world, economic agency is a theoretical construction elaborated as part of the development of a family of models. (Ross, 2012, p. 691)

By contrast, the latter two passages imply that economics doesn't need psychology because economic agents are not human. Concepts like utility and preference are theoretical constructions—they are a necessary part of the economist's toolkit; but no poking around inside of the head of individuals will reveal what utility is or where preferences come from. Any entity can, in theory, be modeled as an economic agent and this means that individual person isn't special. But this suggests that persons may not centers of decision-making because choice, as it is traditionally conceived by economists, is the outcome of both internal processes and external forces.

The passages above reveal an interesting but crucial tension in contemporary economics concerning agency and choice: given the bounded rationality of ordinary humans, and, given the tools and concepts of orthodox economics, researchers are faced with the joint dilemma of re-evaluating their conception of economic agency *and* with defining more suitable candidates for the ascription of utilities and/or preferences. As will become evident in this thesis, this tension pulls in different directions and gives rise to conflicting ideologies about the future of economics and decision research.

3. Four questions about agency and choice

The considerations above give rise to a number of philosophical and methodological questions for scientific disciplines in the employ of rational choice theory. The chapters in this thesis are centered around four sets of questions:

Chapter 2: What does it mean to describe choice evidence as "mental" or "behavioral"? How useful are such labels for interpreting decision phenomena, and what are the implications of their use in contemporary economic research?

Chapter 3: To what extent are persons like economic agents, and under what conditions do persons approximate economic agency? What does social cognition have to do with economic agency?

Chapter 4: How has interdisciplinary research on internal conflict and self-control impacted the concept of economic agency? What are the conceptual and ontological challenges of integrating economic formalism with psychological insights?

Chapter 5: What are the advantages and disadvantages of interpreting choice as the outcome of dual processes? How has the dualistic narrative shaped the discipline of behavioral economics?

A well-informed analysis of these questions must inevitably address themes from the broader canon of analytic philosophy, including philosophy of science and philosophy of mind. Below I provide an overview of themes and debates which pertain to each set of questions above. This overview will provide context for some of the more philosophically nuanced issues regarding the cognitive and conceptual foundations of agency and choice.

3.1 *On the curious role of mental states in economics*

Few debates in the history and philosophy of science are as unrelenting as those which concern the scientific status of mental states. It is said that economic theory formalizes microeconomic explanations by representing agents' desires in terms of a utility function over various outcomes and their beliefs in terms of a subjective probability function over various states of the world (Reiss, 2013; Rosenberg, 2018). These together entail a preference ordering. For most rational choice theorists, the logic underlying economic explanations is similar to the logic underlying ordinary folk-psychological reasoning, viz. both rely on the ascription of mental states to explain choice-behavior. Rosenberg (2018) describes this as "folk psychology formalized". Yet, it may surprise some to learn that the ontology of mental states is important to the study of economic methodology: not only is it relevant to the selection and interpretation of evidence, but, for some, the identity of economics as a scientific discipline depends entirely on whether it permits or denies non-choice data—this includes, among other things, mental states (Davis, 2006; Bruni & Sugden, 2007; Hands, 2009, 2013; Hausman, 1998; Ross, 2014; Edwards, 2012). There are two main views that are helpful to introduce at this point.

Behaviorism, broadly construed, is the position that humans are stimulus-response machines, and that behavior can be described and explained without making reference to mental events or to internal psychological processes (Graham, 2017). Behaviorists tend to regard individual actions as patterned—or conditioned—responses to external forces. These patterned responses may evolve into ever more sophisticated dispositions as new experiences feed into a person's behavioral repertoire. This is what allows individuals to learn from their environment. Yet, the history of

behaviorism as both a theoretical doctrine and a series of scientific programs in the history and philosophy of science is quite complex: its role in economics is tied up in its role in psychology. To paraphrase Graham (2017), behaviorism can be interpreted in (at least) three ways, i.e. methodologically, psychologically, and/or analytically (I will not review their differences here).

Mentalism, by contrast, is the position that humans are more than stimulus-response machines, and that in order to understand individuals' decisions and choice-behaviors, economists may need to investigate the goings-on of the mind and/or brain. But like behaviorism, there are different variants of mentalism. One approach, dubbed "mindful economics" (Camerer, 2008; Hausman, 2008) has gained traction as a catch-all phrase for models that either include psychological information or make claims (i.e. predictions, explanations) about psychological phenomena in relation to economic behavior. The conventional wisdom here is that because mental states serve to predict and rationalize agents' behavior, mental states should be included in economists' everyday ontology of scientific objects. Mindful economics is generally not restrictive about what counts as psychological information. However, there are those within the mentalist camp who wish to distinguish mental states from purely physiological and neural states (Dietrich & List, 2013, 2016; Okasha, 2016). This move is based on the idea that folk-psychological concepts are a class of scientific objects all their own and this special status allows them to play a unique role in economic models. But what is folk-psychology, exactly?

Folk psychology refers to a patchwork of linguistic practices and sense-making norms according to which people predict and interpret each other's actions. For many philosophers, folk psychology is synonymous with commonsense, wherein everyday psychological idioms—belief, desire, and intention being the most cited examples—are used to ascribe mental states (McGeer, 2007; Hutto & Ratcliffe, 2007; Hutto, 2007—see Ratcliffe, 2006, for compelling counterarguments). The relevance of folk psychology for economic methodology rests in the *functional* role that mental-state terms play: beliefs and desires don't just represent internal, psychological processes—their function as a sense-making technology is tied-up in the behavioral patterns that these terms support and describe (Davidson, 1974; Dennett, 1971, 1978, 1989; Ross, 2005; cf. Fodor, 1987). In this way, what permits rational choice theorists to formalize folk psychology is the belief that mental states explain by virtue of their commonsense functions (Elster, 1983; Pettit, 1991, 2000; Reiss, 2013; Rosenberg, 2018).[1]

While there is indeed a major positive shift in attitude regarding the permissibility of mental states for explaining decision phenomena (this is likely due to the growing

[1] There is, of course, no denying that folk-psychological practices are underwritten by various neurobiological processes. But, what differentiates folk psychology from cognitive psychology or neuroscience is the recognition that mental state ascriptions are linguistic practices, and that words like "belief" and "desire" refer not to brain states but to behaviors.

popularity of behavioral economics), there are interesting, if contentious, assumptions built into recent defenses of mentalism which seem to ignore decades of careful toiling over how mental state ascriptions relate to actual folk-psychological practices.[2] This has implications for current debates in economics about whether decision theoretic concepts like utility, belief, and preferences should be interpreted as mental states or, by contrast, as dispositions to act and behave in certain ways.

In **Chapter 2**, I evaluate the relevance of the mentalism-behaviorism (MB) dichotomy in economics in light of recent debates and subsequent arguments in favor of mentalism. The MB dichotomy in economics has historical ties to debates in the history and philosophy of science concerning the foundations of psychological explanation. In this chapter, I argue that there are two problems with current conceptions of the MB dichotomy as it pertains to how economists and decision researchers interpret and gather evidence. First, it is unclear what the MB dichotomy pertains to or is about exactly—which is to say, economists and decision researchers may have different motivations for endorsing mentalism and/or for opposing behaviorism. Second, and more importantly, it is unclear how the MB dichotomy is supposed to improve or advance empirical research in economics and decision research—in particular, supporters of mentalism have the difficult task of clarifying what mentalism entails or consists in (beyond vapid appeals to folk psychology). In response to the first problem, I consider two common motivations for endorsing mentalism: one motivation appeals to the *choice-theoretic foundations* of economics; the other appeals to *scientific practice* in economics. In response to the second problem, I argue that the MB dichotomy likely won't advance or improve scientific practice in contemporary economic settings because neither mentalism (nor behaviorism) are equipped to analyze and resolve explanatory problems that are unique to non-choice data, i.e. psychological and neuroscientific data. I conclude by discussing the limitations of functionalism, the mainstay of the mentalism defense book, and suggest alternative schemas to the MB

[2] Some philosophers of mind and cognitive scientists are skeptical of the propositional attitude interpretation of folk psychology because it presumes an internalist (neocartesian) picture of the individual. In fact, there are a number of reasons why philosophers reject this view; but three will suffice to make the case. Firstly, the propositional attitude interpretation of folk psychology presupposes that individuals have first-person epistemic authority (self-knowledge) about their mental states. But introspection is not always reliable as people are prone to confabulation and other forms of error or self-deception about their beliefs, desires, etc. (McGeer, 1996; Nisbett & Ross, 1980). Secondly, Introspection, understood as the process of accessing self-knowledge, is not psychologically realistic if it excludes external sources of information—namely, other people and norms of reinforcement. Thus, self-knowledge is constructed with the help of others through processes of enculturation. This means that the terms used to pick out mental states have a commissive and regulative element (McGeer, 2007, 2015; Hutto, 2007). Thirdly, propositional attitudes are crude semantic approximations of cognitive and affective states that aren't well understood by cognitive neuroscience. It may be, and likely is, the case that there is nothing structurally analogous to beliefs or desires in the brain (Dennett, 1991; Hutto, 2007; Hutto & Myin, 2012).

dichotomy, some of which are employed in neighboring areas of the cognitive and behavioral sciences.

3.2 *Individualism versus anti-individualism: an ontological debate*

Many debates in the social and behavioral sciences revolve around the idea that collective action can be explained in terms of individual behavior. Such views emphasize the importance of persons *as* intentional agents and assume that collective actions can be investigated by appealing to the internal psychological states of individuals. Moreover, such views hold that social phenomena—such as markets and business cycles, voting trends, surges in innovation, language conventions, and other artifacts of social interaction—can be decomposed into the actions of individuals despite their apparent complexity. This popular albeit controversial view is known as individualism and is often, though perhaps misleadingly, called methodological individualism (Hodgson, 2007; Ross, 2005).

In principle, individualism supposes that if some social phenomenon is decomposable into the actions of individual persons, then knowledge of the causes of their behaviors—what could be called "micro-foundations"—should be sufficient to understand how the social phenomenon occurs and produces further social phenomena. However, what constitutes a micro-foundation is a contingent matter rather than a principled one. For instance, individualism could be read as an ontological thesis, meaning that individual persons have a special, theoretical status among other objects in the world; what we then perceive as collective actions and events are merely epiphenomena, i.e. events that supervene on the actions of individuals. Or, individualism could be read as a metaphysical thesis, meaning that collective actions are *bona fide* phenomena, but that individual persons are causally necessary to produce such phenomena. Or, individualism could be read as an explanatory thesis, meaning that collective actions are descriptively redundant to the extent that knowledge of the mechanics of individual choice are more parsimonious or more informative than explanations which reside at the social level.

That there is discrepancy over which is the correct interpretation of individualism raises a critical issue for proponents of it—namely, that it is uncertain what is the right criterion for decomposing and thereby understanding social phenomena. What serves the function of a micro-foundation in one context may be entirely inappropriate in another. This issue is further complicated by the fact that individualism is not a theory per se (similar to folk psychology), but a family of theses that loosely correspond to researchers' concerns about socially-embedded individuals.

Yet, in market contexts (which is nearly all contexts), people are bounded—both rationally and individually (Ross, 2005; Davis, 2014); and the institutional and informational structures through which people are bounded are external to individuals.

Hence, it seems unlike that the same structural dynamics which produce social action—those which simultaneously constrain and support how individuals choose and act with others—could be interpreted as or read off internal decision processes.

In **Chapter 3** I argue that individualism is problematic as a basis for investigating social interaction. In so doing, I examine Don Ross's (2005, 2006) account of "multiple-selves" as a way of reconciling individuals' bounded rationality with their bounded individuality. Ross argues that individual persons are complex aggregations of selves, which arise in response to external pressures to regulate individual behaviors and enable the tracking of public norms and conventions. I thus investigate the different roles that selves play in Ross's broader philosophy of economics and I identify separate projects that arise therein. To this end, I distinguish three different roles for selves, which are *evolutionary*, *narrative*, and *economic*, and I argue that these roles contribute to two distinct, but overlapping, projects. My aim is to show that there is a tension underlying these projects, but that it's difficult to say where this tension arises because of how selves are multiply understood and used to defend these projects. I will argue that, while it is not problematic to conceive of selves according to their different roles, we should not presume that the functions or properties of selves in one role can serve the same purposes for different projects.

3.3 *Mixing metaphors: dual-selves or dual-processes*

Philosophers often speak of carving nature "at its joints". Since Plato (see *Phaedrus*), this figure of speech has been used to describe the analytical exercise of partitioning the world into manageable parts and properties—what some philosophers call "natural kinds" (cf. Campbell, O'Rourke, and Slater, 2011). The aim of such an exercise is not simply to determine which parts and properties of the world are fundamental, as this is a job for physicists; it is, rather, to understand which categories are instrumental and conducive to understanding the natural world. The reason philosophers speak of carving nature at its joints is because knowledge of fundamental parts and properties isn't sufficient to provide understanding of more complex objects and processes. (If it were, then all of natural science would devolve into fundamental physics.) However, some phenomena, namely social phenomena like choice formation, do not lend themselves to easy carving, as it were, in which case researchers rely on metaphors to take some of the explanatory burden. Consider the feeling of being "of two minds" about a situation, or of feeling loath to accomplish a task. What does it mean to be of two minds about a decision? There are different ways of cashing out this idiom, and the analogy of carving nature at its joints is particularly instructive here:

Multiple-self models of intrapersonal and intertemporal choice emerged in decision theory and game theory to help economists better understand the dynamics of internal conflict and to predict—and hopefully explain—choice anomalies and

inconsistencies that arise over time. This modeling technique is achieved by depicting the individual decision-maker as a coalition of temporally distinct "selves" who must cooperate (or compete) to satisfy their respective ends. Although early intertemporal choice models were not intended to identify the psychological determinants of choice (each temporal self was taken to be an independent utility-maximizing agent, cf. Samuelson, 1937) subsequent time-preference models by Strotz (1956) and Phelps & Pollak (1968) proposed to partition individuals into selves (or generations) with distinctive motives. It was thus demonstrated that myopic and weak-willed behaviors could be the result of a tradeoff between short and long-term interests. Thaler & Shefrin (1981; cf. Shefrin & Thaler, 1988) were among the first to conceive of this multiple-self approach in an explicitly dualistic framework between a long-run "planner" self and short-run "doer" self.

On the other hand, the concept of bounded rationality invoked by many behavioral economists and decision researchers relies on the notion of information processing. Clearly humans are not "von Neumann computers"; yet, the idea that the brain can be interpreted as a computer has roots in the cognitive revolution of the 1960's and 1970s wherein the majority of human mental activities began to be interpreted as information processing (Baars, 1996; Garner, 1987; Daugman, 2001; Mirowski, 2002). Despite philosophical debates about the nature of computational theories of cognition,[3] the computer metaphor has been widely and repeatedly reinforced in the behavioral sciences under the assumption that humans—or rather, their brains—actually perform computations when reasoning and problem-solving. Part of what has made this brain-as-a-computer metaphor gain so much traction is that it builds upon a secondary, though perhaps more confusing notion in the cognitive and behavioral science—that notion of cognitive processing. Hence, *dual-process theories* of reasoning and judgment are another means of capturing internal conflict. While there are dual-process theories for nearly every aspect of cognition, the primary assumption behind dual-process theories is that both conscious and unconscious thinking depends on the interplay of separate cognitive modes: one mode is said to involve processes that are fast, reactive, and automatic, while the other mode is said to involve processes that are slow, controlled, and deliberative (Schneider & Shiffrin, 1977; Epstein, 1994; Stanovich & West, 2000; Lieberman, 2003). This distinction allows researchers to discern "higher" cognitive processing, which are associated with deliberative judgments and the ability to reason logically, from "lower", more primitive information processing, which is usually associated with affective states and visceral responses.

[3] See van Gelder (1995), Hutto et al (2018); cf. Piccinini (2013), Piccinini & Bahar (2015).

Over the last two decades, interesting collaborations between economists and psychologists have given rise to integrative models which weave together the metaphor of the multiple-self with the metaphor of the dual-information processor.

In **Chapter 4**, I critically examine how multiple-self models of intrapersonal and intertemporal choice have been integrated with dual-process and dual-system theories from social psychology and cognitive science. I adopt the term "multi-agent model" to denote models which conceive of multiple agents with multiple psychological abilities within the individual. Such models seem to be growing in popularity given their purported ability to predict and explain reasoning errors and decision anomalies due to internal conflict or lack of self-control. In particular, I analyze how multi-agent models conceive of and employ "selves" and "systems" for the purposes of representing intrapersonal and intraneural conflict. The chapter is structured according to three claims. The first and second claims establish that multi-agent models are conceptually as well as ontologically ambiguous. The third claim argues that such ambiguities can lead to problems in scientific understanding. The examination of multi-agent models is not only critical to understanding how economists and psychologists jointly interpret and model self-control problems, but it further presents an important opportunity to study the effects of cross-disciplinary pollination of concepts and theories.

3.4 *Why is two the magic number? Further challenges for dual process theories*

The explanatory heuristic of parsing individuals into manageable parts has historically taken *two* to be the magic number, often using a dualistic framework to contrast competing aspects of the human will. As described by Evans & Frankish (2009), the legacy of framing human thought as dualistic has roots in Plato, Augustine, Freud, James, and so forth; and as I suggested above, both cognitive and behavioral scientists have latched on (hard) to this framework. Behavioral economists' preference for partitioning human activity (critical thinking, decision-making) into dual process and dual systems seems to be more than merely a passing fad.

However, the faith in dual process theory indicates more than an interest in improved modeling. In fact, it was recently argued that behavioral economics, construed as an independent field of research, is closer in kind to cognitive science than it is to orthodox economics. Angner & Loewenstein (2012) observe a number of links between behavioral economics and cognitive science, which they attribute to the success of behavioral economics as an independent discipline. These links range from shared theoretical commitments, e.g., both disavow positivist methodological doctrines in the behavioral sciences, to historical affiliations, e.g., behavioral economics emerged from the field of behavioral decision research. The claim that behavioral economics has a kinship with cognitive science represents a bold new step in a series of reflections on the relationship between economics and psychology. However, Angner &

Loewenstein's appraisal of the links between behavioral economics and the cognitive sciences is uncritical in ways that reinforce the problems above. It takes for granted (and even seems to celebrate) the freedom with which behavioral economists have explored the bounds of human rationality. It doesn't consider whether the insights and resources accumulated from the cognitive sciences are credible or well-founded, which is to say, it does not actively engage with debates in psychology or cognitive science. This is representative of a broader trend in the literature on economics and psychology, in which greater emphasis is placed on the history of interdisciplinary exchanges than on issues which may be pertinent to the philosophy of science (see Lewin, 1996; Rabin, 1998; Sent, 2004; Camerer, Loewenstein, & Rabin, 2011; and Heukelom, 2014).[4]

In fact, it could be argued that this lack of emphasis on the philosophy of science has something to do with how behavioral economics is generally conceived. Angner & Loewenstein write that, "These days, as it is typically employed 'behavioral economics' refers to the attempt to increase the explanatory and predictive power of economic theory by providing it with more psychologically plausible foundations" where psychological plausibility means "consistent with the best available psychology" (2012, p. 642).

In **Chapter 5**, I confront the success story of behavioral economics by investigating the broader role that dual process theory has played as a psychological framework: Cognitive scientists and philosophical psychologists alike have criticized the theoretical foundations of the standard view of dual process theory and have argued against the validity and relevance of evidence used to support it. Moreover, recent modifications of dual process theory in light of these criticisms have generated additional concerns regarding its applicability and irrefutability. I argue that this should raise concerns for behavioral economists who see dual process theory as providing psychologically realistic foundations for their models. In particular, it raises the possibility that dualistic models are not as descriptively accurate or reliable as behavioral economists presume them to be. In fact, the case can be made that the popularity of dual process theory in behavioral economics has less to do with the empirical success of dualistic models, and more to do with the convenience that the dualism narrative provides economists looking to sort out decision anomalies. I will argue that the growing number of criticisms against DPT leaves behavioral economists with something of a dilemma: either they stick to their purported ambitions to give a realistic

[4] Investigations into the interdisciplinary exchanges between economics and psychology tend to focus on the historical episodes that led the disciplines to come together, with the emphasis on how economics has changed as a result of importing psychological concepts and theory. Such investigations tend to presume the credibility or factivity of psychological concepts and theories rather than engage them directly.

description of human decision-making and modify their use of DPT, or they stick to DPT and modify their ambitions.

4. Outlook

To conclude, two notes are in order. First, the chapters of this thesis are conceived of as independent research articles and are intended to be read that way. For this reason, there is no signaling to former or latter chapters—each is a stand-alone essay. But this also means there is occasional repetition in the listing of references and explication of concepts. But this is minimal. Second, in most instances, the term "economics" denotes microeconomics or some area of microeconomics, e.g., decision theory, game theory, behavioral economics, and so on.

The goal of this thesis is to provide a philosophical analysis at two levels: one is to understand and analyze the operative concepts *agency* and *choice*, and to track their various forms and distillations across economics and behavioral decision research. Two is to understand how theories and models germane to these operative concepts travel between scientific disciplines, and to assess how this promotes and limits interdisciplinary collaboration.

In **Chapter 6** I offer concluding remarks and consider where one goes from here. First, Chapters 2 – 5 project two main approaches to reconciling the tension between agency and choice. One approach views individual persons as the primary objects of study for economics, and as such, psychology and neuroscience can help locate a more appropriate locus for the study of choice. The second approach views individual persons not as the primary object of study, (economic agents are the primary study, and they are ontologically distinct from persons). As such, choice should be construed as the outcome of external (market) pressures, which include important socio-cognitive supports. Hence, for each of these approaches, there are new pursuits and new philosophical questions to be considered.

5. Bibliography

Angner, E., & Loewenstein, G. (2012). Behavioral economics. In Mäki, U. (Ed.), *Philosophy of Economics* (pp. 641-689). Amsterdam: Elsevier.

Baars, B. J. (1986). *The Cognitive Revolution in Psychology* (Vol. 157). New York: Guilford Press.

Bruni, L., & Sugden, R. (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. *The Economic Journal*, *117*(516), 146-173.

Campbell, J. K., O'Rourke, M., & Slater, M. H. (Eds.). (2011). *Carving Nature at its Ioints: Natural Kinds in Metaphysics and Science*. MIT Press.

Camerer, C. (1999). Behavioral economics: Reunifying psychology and economics. *Proceedings of the National Academy of Sciences*, 96(19), 10575-10577.

Camerer, C. (2008). The case for mindful economics. In Caplin A., & Schotter A. (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 43-69). New York: Oxford University Press,

Camerer, C. F., Loewenstein, G., & Rabin, M. (Eds.). (2011). *Advances in Behavioral Economics*. Princeton university press.

Davidson, D. (1974). Psychology as philosophy. In *Philosophy of Psychology* (pp. 41-52). Palgrave Macmillan, London.

Davis, J. B. (2006). The turn in economics: neoclassical dominance to mainstream pluralism?. *Journal of Institutional Economics*, 2(1), 1-20.

Davis, J. B. (2015). Bounded rationality and bounded individuality. In *A Research Annual* (pp. 75-93). Emerald Group Publishing Limited.

Daugman, J. G. (1993). Brain metaphor and brain theory. In *Computational neuroscience* (pp. 9-18). MIT Press.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87-106.

Dennett, D. C. (1978). Three kinds of intentional psychology. *Perspectives in the Philosophy of Language: A Concise Anthology*, 163-186.

Dennett, D. C. (1989). *The Intentional Stance*. MIT press.

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

Dietrich, F., & List, C. (2016). Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics & Philosophy*, 32(2), 249-281.

Evans, J. S. B., & Frankish, K. E. (Eds.). (2009). *In Two Minds: Dual Processes and Beyond*. Oxford University Press.

Edwards, J. (2016). Behaviorism and control in the history of economics and psychology. *History of Political Economy*, 48(suppl_1), 170-197.

Fodor, J. A. (1987). *Psychosemantics: The problem of Meaning in the Philosophy of Mind* (Vol. 2). MIT press.

Elster, J. (1983). *Explaining Technical Change: A Case Study in the Philosophy of Science*. CUP Archive.

Gardner, H. (1987). *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic books.

Graham, G. (2017). Behaviorism, In Zalta E. (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University, Spring 2017 Edition.

Guala, F. (2012). Are preferences for real? Choice theory, folk psychology, and the hard case for commonsensible realism. In In Lehtinen A., & Ylikoski P. (Eds.), *Economics for Real* (pp. 151-169). Routledge.

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. In Caplin A., & Schotter A. (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 3-42). New York: Oxford University Press.

Hands, D. W. (2009). Economics, psychology and the history of consumer choice theory. *Cambridge Journal of Economics*, 34(4), 633-648.

Hands, D. W. (2013). Foundations of contemporary revealed preference theory. *Erkenntnis*, 78(5), 1081-1108.

Hausman, D. M. (1998). Problems with realism in economics. *Economics & Philosophy*, 14(2), 185-213.

Hausman, D. M. (2008). Mindless or mindful economics: a methodological evaluation. In Caplin A., & Schotter A. (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 125-151). New York: Oxford University Press.

Hodgson, G. M. (2007). Meanings of methodological individualism. *Journal of Economic Methodology*, 14(2), 211-226.

Hutto, D.D. (2007). The narrative practice hypothesis: origins and applications of folk psychology. *Royal Institute of Philosophy Supplements*, 60, 43-68.

Hutto, D.D., & Ratcliffe, M. (Eds.). (2007). *Folk Psychology Re-assessed*. Dordrecht: Springer.

Hutto, D. D., & Myin, E. (2012). *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press.

Hutto, D. D., Myin, E., Peeters, A., & Zahnoun, F. (2018). Putting computation in its place. In Sprevak M., & Colombo M. (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 272-282). London: Routledge.

Kitcher, P. (1984). 1953 and all that. A tale of two sciences. *The Philosophical Review*, 93(3), 335-373.

Lewin, S. B. (1996). Economics and psychology: Lessons for our own day from the early twentieth century. *Journal of Economic Literature*, 34(3), 1293-1323.

Mäki, U. (1998). Aspects of realism about economics. *Theoria: An International Journal for Theory, History and Foundations of Science*, 1, 301-319.

McGeer, V. (1996). Is" self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. *The Journal of Philosophy*, 93(10), 483-515.

McGeer, V. (2007). The regulative dimension of folk psychology. In Hutto, D.D., & Ratcliffe, M. (Eds.), *Folk Psychology Re-assessed* (pp. 137-156). Springer, Dordrecht.

McGeer, V. (2015). Mind-making practices: the social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18(2), 259-281.

Mirowski, P. (2002). *Machine Dreams*. Cambridge: Cambridge University Press

Nisbett, R.E. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall.

Okasha, S. (2016). On the interpretation of decision theory. *Economics & Philosophy*, 32(3), 409-433.

Pettit, P. (1991). Decision theory and folk psychology. In Bacharach, M. & Hurley, S. (Eds.), *Foundations of Decision Theory: Issues and Advances*, (pp.147-75). Blackwells, Oxford.

Pettit, P. (2000). Rational choice, functional selection and empty black boxes. *Journal of Economic Methodology*, 7(1), pp.33-57.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, 37(3), 453–488.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, 36(1), 11-46.

Ratcliffe, M. (2006). 'Folk psychology' is not folk psychology. *Phenomenology and the Cognitive Sciences*, 5(1), 31-52.

Reiss, J. (2013). *Philosophy of Economics: A Contemporary Introduction*. Routledge.

Rosenberg, A. (2018). *Philosophy of Social Science*. Routledge.

Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Ross, D. (2006). The economic and evolutionary basis of selves. *Cognitive Systems Research*, 7(2-3), 246-258.

Ross, D. (2012). The Economic Agent: Not Human, But Important. In Mäki, U. (Ed.), *Philosophy of Economics* (pp. 691-735). Amsterdam: Elsevier.

Ross, D. (2014). Psychological versus economic models of bounded rationality. *Journal of Economic Methodology*, 21(4), 411-427.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences*, 23(5), 645-665.

Van Gelder, T. (1995). What might cognition be, if not computation?. *The Journal of Philosophy*, 92(7), 345-381.

Chapter 2

Two problems for the mentalism-behaviorism dichotomy in economics

1. Introduction

Few oppositions in the history and philosophy of science are as convoluted and thereby as polarizing as the mentalism-behaviorism dichotomy (henceforth "M-B dichotomy"). Historical analyses suggest that the M-B dichotomy is critical to the disciplinary identity of economics, and that disputes about the role of psychological explanations in economics are predicated on how one conceives economics as a science (Hausman, 1998; Davis, 2006; Bruni & Sugden, 2007; Backhouse & Medema, 2009; Hands, 2009, 2014; Mäki, 2010; Ross, 2011, 2014).[1] Moreover, recent debates about the interpretation of decision-theoretic concepts have complicated how economics relates to psychology. The literature indicates that *utility* and *preference* can take on very different meanings, which has a significant effect on how one interprets the M-B dichotomy in economics (Dowding, 2002; Camerer, 2008; Hausman, 2008, 2012; Guala, 2012, *unpublished*; Hands, 2013; Clarke, 2016; Okasha, 2016; Dietrich & List, 2016).

According to Dietrich & List (2016), behaviorists and mentalists are divided on two questions, namely on "whether or not the evidence base of economics should be restricted to choice behavior" and, on "whether the relations or functions playing a preference-or-belief role in economic theories should be treated as mere theoretical constructs or as corresponding to real phenomena" (2016, p. 267). This characterization of the dichotomy follows a controversial series of exchanges as recorded in Caplin & Schotter (2008). For instance, Gul & Pesendorfer (2008) are now infamous for arguing that nonchoice data—essentially, any data that is not strictly behavioral in nature—cannot be used to support *or* reject economic theories and methodology. This is, they argue, because economics is the science of aggregated choice-behavior, nothing more (2008, p. 3). Gul & Pesendorfer believe that economics and psychology are fundamentally different disciplines insofar as they "address different questions, utilize different abstractions, and address different types of empirical evidence" (2008, p. 4). Understandably, Gul & Pesendorfer's position has generated much backlash among historians and methodologists of economics. Among responses to Gul & Pesendorfer, Dietrich & List's (2016) is perhaps the most definitive: they argue that not only should

[1] For an historical overview of the role of psychological explanation in economics, see Loewenstein (1992), Lewin (1996), Rabin (1998), Sent (2004), and recently Edwards (2016). For commentary on the importance of behavioral economics to the economics discipline, see Camerer, Loewenstein, & Rabin (2011), Angner & Loewenstein (2012), and Heukelom (2014).

psychological evidence be admissible for explanatory purposes in economics, but that it is *indispensable* for those purposes.[2]

In response to these exchanges, and in light of newly unfolding debates about its role in economics and behavioral decision research, I argue that there are two problems with current interpretations of the M-B dichotomy. First and foremost, it is uncertain what, exactly, the dichotomy pertains to or what is implicated by it: different token debates reveal that researchers have different motivations for endorsing mentalism and/or for opposing behaviorism—there is not one dichotomy. I refer to this as the primary problem. Second, it is uncertain how the dichotomy informs, or is informed by, empirical research in economics: while researchers may have different motivations for endorsing mentalism and/or for opposing behaviorism, it seems that those who take the dichotomy most seriously have the least to say about cutting-edge contemporary research in other domains of economics that draw on psychological data. I refer to this as the secondary problem.

In response to the primary problem, I qualify and compare arguments that appeal to the *choice-theoretic foundations* of economics and distinguish them from arguments that appeal to *scientific practice* in economics. However, given the complicated history of economics (especially alongside psychology and other behavioral sciences), it is an open question whether arguments that appeal to choice-theoretic foundations can resolve the M-B dichotomy—it seems to me that this is a job better suited to historians. In response to the secondary problem, I argue that the M-B dichotomy does not advance or improve scientific practice in economics in the domains of decision research because neither mentalism nor behaviorism are equipped to analyze nonchoice data or to resolve explanatory problems. In explicating these two problems, I bracket out philosophical considerations of the role of functional explanations; this is a topic which needs to be dealt with separately as it requires more sophisticated diagnosis from the perspective of philosophy of mind. After addressing the primary and secondary problems I argue that functionalism—as it is understood and defended by philosophers of economics—invites more problems than solutions, and may not be epistemically reliable for explanatory purposes which contemporary economists claim to pursue.

The paper has the following structure: In section 2, I characterize the M-B dichotomy via exchanges between Gul & Pesendorfer (2008) and Dietrich & List (2016). I then sketch the primary problem and secondary problem in response to Dietrich & List's conception of the M-B dichotomy, which includes a critical analysis of their version of mentalism. Section 3 considers alternative interpretations of the M-B dichotomy and motivates my response to the primary problem. Section 4 then introduces two examples from behavioral economics and neuroeconomics and shows why the

[2] Going forward, all instances of "economics" refer to microeconomics.

M-B dichotomy to is not suited to advance empirical decision research. This leads me to further explore the limitations of functionalism for explanatory purposes. Section 5 anticipates rebuttals and concludes.

2. Mentalism and behaviorism from a "philosophy-of-science" perspective

To get a better idea of how the M-B dichotomy has played out in recent debates, I examine Dietrich & List's (2016) critique of Gul & Pesendorfer (2008). This exegesis will establish two things: first, it shows why radical behaviorism is problematic from a philosophy-of-science perspective; second, it summarizes the advantages of mentalism as a counter-position to behaviorism. This sets up the more critical discussion, which investigates what exactly mentalism entails.

2.1 *On the supposed threat of radical behaviorism*

Gul & Pesendorfer's The Case for Mindless Economics (2008) aims is to establish the "proper" domain of economic research in light of attempts to refine its descriptive and normative foundations (cf. Kahneman, 1994; Rabin, 1998, 2002; Camerer, Prelec & Loewenstein, 2004, 2005). In particular, Gul & Pesendorfer condemn the use of nonchoice data, viz. psychological and neuroscientific evidence, to explain choice-behavioral phenomena and to generate more realistic economic models. Both the paper's message and overall tone has garnered harsh criticism from philosophers and methodologists alike (Camerer, 2008; Hausman, 2008; cf. Ross, 2014). In this respect, Dietrich & List (2016) are no different.

Dietrich & List identify three separate and contentious claims in *Mindless Economics*—these are: (i) the only evidence that should be used to test economic theories is evidence about people's choice behavior; (ii) the content of any economic theory consists solely in its choice-behavioral implications; two theories that are choice-behaviorally equivalent should be seen as equivalent simpliciter; (iii) any economic theory should take the form of a representation of choice behavior, and that representation should ideally take the form of attributing to the agents the maximization of some objective function (2016, pp. 253-254).[3] They argue that such a radical interpretation of economics is both scientifically naïve and misinformed about the benefits of an expanded evidence base, which they demonstrate by explicating several misconceptions which underlie the behaviorist methodology (2016, pp. 254-259). Let's consider two of those misconceptions:

[3] Dietrich & List clarify that each of these claims corresponds to a different behaviorist thesis—the first to "psychological behaviorism, the second to "analytical or logical behaviorism" and the third to "methodological behaviorism" (2016, pp. 253-254). Although each are problematic, it is the first two that portend a truly radical threat.

> Gul & Pesendorfer argue that the only evidence that should be used to test economic theories is evidence about people's choice behavior. But there is no systematic reason why the evidence base of economics should be restricted in this way… [The] idea that the evidence base of a particular scientific discipline should be fixed once and for all lacks any justification, given the history of science and the experience of other scientific disciplines. Rather, the evidence base of any science is changeable and dynamic, and there is no reason why economics should be an exception. (Dietrich & List, 2016, pp. 255-256)

This is the "misconception of a fixed evidence base"; it posits that economics, or any natural science for that matter, need not restrict what it admits as evidence for the development of theory. Indeed, Gul & Pesendorfer offer no compelling reasons for limiting the role of psychology and excluding nonchoice data from standard economics. Their justification, which declares that economics *just is* the science of observable choice, begs the question. I return to this point below. Let's consider another:

> While [maximization] may be a useful starting point for the explanation of behavior to search for some objective function that a given agent maximizes, there is no principled reason why our best theories of economic behavior should *necessarily* be based on the notion of maximization… Which *form* of a theory best explains human behavior is a contingent, empirical question, which can be settled only by actual scientific research, not by methodological stipulation. (Dietrich & List, 2016, pp. 258)

This is the "maximization dogma". It states that utility maximization is a central theoretical component of revealed preference theory, which means that it is a central theoretical component of most behaviorist methodologies. But why presume that individuals maximize anything at all? Dietrich & List argue that there is no *a priori* reason to think that an individual's behavior should maximize some objective function *and* that the evidence base of economics should be fixed once and for all.[4] As such, the misconceptions identified by Dietrich & List indicate that Mindless Economics is based merely on a definition of what Gul & Pesendorfer think economics is rather than on a well-founded argument against mentalism (cf. Camerer, 2008, for initial reactions to their argument from a definition).

If Gul & Pesendorfer's arguments for behaviorism issue from a stipulative attitude about what economics is or should be, then Dietrich & List's counterarguments for mentalism issue from a "naturalistic attitude", which is analogous to scientific

[4] Though, Dietrich & List do not distinguish between *maximization* of individual utility via choice behavior and *optimization* at the level of the model. Indeed, it may be the case that flesh-and-blood individuals do not actually maximize expected utility—not consciously anyway; but this does not preclude economic models from utilizing the mathematics of constrained optimization (cf. Ross, 2014 for useful discussion). For this reason, I do not endorse Dietrich & List's claim that the maximization dogma is necessarily a misconception.

practice in the natural sciences (2016, p. 268). This natural attitude entails the following: that *if* an entity or property is among the ontological commitments of a well-established scientific theory, *then* it ought to be taken "at face value", which is to say, admitted for the purposes of explanation (this is, of course, presuming that one does not have independent reasons for disqualifying those entities or properties). By implication, even if one views economics as just the science of observable choice-behavior, it is nevertheless committed to entities and properties other than choice-behavior by virtue of the theories it posits to model and rationalize behavior: "…when our best theories of economic decision-making are committed to certain mental-state constructs in the technical sense (i.e. relations or functions playing a preference-or-belief role), we should treat these as corresponding to real features of the world, unless we wish to reject those theories themselves (Dietrich & List, 2016, p. 268). But this natural attitude needs to be fleshed out a bit further.

In order to show that mentalism is explanatorily superior to behaviorism, and not merely an alternative theoretical framework, it must be demonstrated that there is *no* independent reason for rejecting those aspects of economic theory which would commit one to the existence of mental states. Dietrich & List do this by showing that revealed preference theory (RPT) is untenable unless mental-state constructs do play a functional role beyond the merely operational sense in which they represent agents' preference orderings. Dietrich & List take is as given that RPT is not an ontological thesis (2016, p. 268), which allows them to surmise that any epistemological interpretation of RPT is compatible with mentalism *because* the functional roles that subjective probabilities and utility functions play is mediated by economists' commonsense understanding of intentional states.[5] By disjunction, behaviorists who invoke RPT are, at least implicitly, committed to folk psychology (though, what it means to be "committed" to folk psychology isn't clear). Once they've shown that RPT is compatible with mentalism, Dietrich & List proceed to explicate what a mentalistic explanation entails—which turns out to be quite unlike other "mindful" conceptions of economics that permit different forms of nonchoice data to play explanatory roles (cf. Camerer 2008, Hausman 2008 in response to Gul & Pesendorfer, 2008).

Dietrich & List defend a unique version of mentalism which can be expressed by way of two core convictions:

[5] In particular, they argue that, whereas an epistemological interpretation of RPT merely restricts the evidence base of a theory to an agent's choice behavior, an ontological interpretation RPT would restrict a theory's ontological commitments to choice behavior. They argue that epistemological RPT is plausible but ontological RPT is not because standard economic theories appeal to mental-state constructs like subjective probabilities and utility functions. They infer that subjective probabilities and utility functions "rationalize" and "systematize" choice behavior by playing a functional role with regard to unobservable entities. For the argument, see (2016, pp. 265-268).

The mind-brain distinction principle purports that decision-theoretic explanations should not conflate the mind, which is constituted by unobservable mental states, with the brain, which is constituted by physiological and neural states. The mind-brain distinction principle is based on the belief that folk-psychological concepts, viz. beliefs, desires, and intentions, are "macro-level" phenomena and therefore distinct from micro-foundational events.[6]

The non-reduction principle purports that decision-theoretic explanations should not be reduced to neuroeconomic explanations because the appropriate level of explanation in economics *is* the level at which mental-state constructs operate functionally, i.e. the level at which folk-psychological concepts apply. The non-reduction principle is based the belief that decision theory represents—or is intended to represent—the *reasons* behind agents' actions (2016, pp. 272-273).

These two convictions are integral to understanding how Dietrich & List's conception of mentalism is different from others. Principally, they take decision theory to codify agent's reasons for choice—reasons are different from whatever transpires 'in the head' of the agent—and reasons depict commonsense ways of thinking about acts and outcomes. The folk-psychological nature of reason-based decision theory entails that mentalism is not compatible with or reducible to neuroeconomic explanation. Thus, their natural attitude toward the M-B dichotomy entails that economics is not only committed to the existence of unobservable mental states, but that its mode of explanation should preserve their ontological status, which is (presumably) non-physical. For this reason, their conception of mentalism is also intrinsically *functionalist*. But how do they understand functionalism? And, how do they determine when some mental state plays the correct functional role? They state that:

> Mental states are, at least in part, states that play a certain role for an agent. Beliefs, for example, play the role of representing certain features of the world from the agent's perspective, and preferences play the role of motivating the agent's actions. Functionalism is the view that what makes something a mental state is simply that it plays the relevant role. (Dietrich & List, 2016, pp. 268-69)

They go on to clarify that functionalism only requires that a preference-or-belief role be indicative of an underlying mental state—it need not be *constitutive* of it.

[6] The conviction I've called the "mind-brain distinction principle" played a more prominent role in an early working draft of their paper (cf. 2012, p. 3). Presumably, their belief that the mind is constituted by "macro-level phenomena" stems from the assumption that unobservable mental states are qualified via public discourse, namely, folk-psychological practices, and then further regulated by sense-making norms. Although Dietrich & List (2016) never defend this claim (they assume it to be true), the assumption that mental states are macro-level phenomena is echoed throughout the published version.

Functionalism, they argue, works by preserving "certain structural features [of the world] but which still abstracts away from many substantive details" (2016, p. 269).

Dietrich & List see this (thin) conception of functionalism to be advantageous to their explanatory goals, and further, to be agreeable with good scientific practices in both the natural and social sciences. At this point, I will not comment on functionalism, except to say that their argument for mentalism depends entirely and precariously upon it. I will return to the topic of functionalism in section 4.

2.2 *Two problems with the "philosophy-of-science" perspective*

Having discussed some of the limitations of radical behaviorism and having sketched an argument for mentalism, I now present two problems with Dietrich & List's conception of the M-B dichotomy. As we'll see, these two problems are representative of much of the debate surrounding the M-B dichotomy.

The *primary problem* with Dietrich & List's conception of M-B dichotomy is that it's not well-defined, which is to say, the domain of their investigation it is simultaneously too narrow and too wide. It is too narrow in the sense that it focuses solely on the shortcomings of Gul & Pesendorfer's (2008) case for mindless economic, which is arguably not representative of less-radical behavioristic accounts. Though, it is too wide in the sense that it makes bold generalizations about the epistemic aims of economics without considering whether their own goals are congruous with other domains of economic research, especially those which already recognize non-choice data to be relevant to explaining decision phenomena.

About the narrowness of their investigation. Although their account may seem like an easy target for criticism, it is important to appreciate that Gul & Pesendorfer discussion and refutation of the "neuroeconomics critique" (2008, p. 4) is not based on what behavioral economists and neuroeconomists actually do. Mindless Economics thus does not offer an argument against mentalism (it never claims to provide one), and for this reason, it gives no impetus for thinking that economic models which utilize nonchoice data would generate bad inferences about human behavior. (To their credit, Gul & Pesendorfer are forth-coming about this—the clearly state that they are not interested in evaluating the contributions of neuroeconomics.) This fact should give one pause to consider the strength and relevance of Dietrich & List' preferred version of mentalism: It is neither established how Mindless Economics fit into the broader spectrum of behavioristic economics, nor whether there are advantages to strictly choice-based models, such as "flexibility" or "sophistication" of revealed preference models (Gul & Pesendorfer, 2008). In fact, Dietrich & List seem to take it for granted that once radical behaviorism is shown to be false—or scientifically naïve—then more moderate forms of behaviorism will acquiesce to mentalism in virtue of their underlying ontological commitments. But this does not "clarify the distinction

between behaviourism and mentalism", which they claim to do (2016, pp. 259-265). In fact, it does the opposite. By locking onto a single, radical account of behaviorism, their conception of the M-B dichotomy and defense of mentalism seem to be rather myopic. It loses the breadth and nuance of what debates that stem from the M-B dichotomy are about.

What this narrow construal of the M-B dichotomy signifies is that Dietrich & List very likely have their own definition of economics in mind. Consider the following passage:

> Setting aside technicalities, the logic underlying this explanation is very similar to the logic underlying folk-psychological reasoning with its ascription of beliefs and desires to explain behaviour. Economic decision theory can thus be seen as a more sophisticated and scientific reconstruction of folk psychology. (2016, p. 250)

This passage not only sheds light on why they hold the convictions they do, but it further indicates that they interpret economics to be an extension of decision theory.[7] This assumption is far from uncontroversial.[8] However, for the present purposes, we need only focus on those problems which arise due to the narrowness of this assumption.

The *secondary problem* is that Dietrich & List do not engage with scientific practice in detail. This is illustrated by the fact that, once they present the main premises of their version of mentalism, there is little engagement with contemporary empirical research in economics (cf. 2016, pp. 271-277). A consequence of this is that one does not see how that their preferred version of mentalism can be used to advance empirical decision research. This is surprising given that a strong motivation for their endorsing mentalism is that it is "in line with best scientific practice" (2016, p. 252).

The secondary problem is easier to understand if we reconsider their non-reduction principle. Recall that this conviction purports that decision-theoretic explanations should not be reduced to neuroeconomic explanations because the appropriate level of explanation in economics is the level at which at which mental-state constructs operate functionally. *Prima facie*, one could read this conviction as merely a word of

[7] See also the following quote: "Decision theory thereby exemplifies a familiar feature of science more generally, which Quine described as commonsense gone self-conscious (Quine, 1960). As Lewis (1983, p. 114) puts it: '[Decision theory] is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference, and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematized.'" (2016, p. 250).

[8] One of the more vivid demonstrations of this point comes from Ross (2014), who argues that economics ought to be regarded as a sociological science, not a science of individual decision-making. His justification for this point is quite complex, as it leverages support from the history of economics as well as from the philosophy of science.

caution to economists not to blindly pursue micro-foundations for explanatory purposes. But what one needs to keep in mind is that their mind-brain distinction principle (which is a precondition for the non-reduction principle) supposes that folk-psychological concepts cannot be reduced to physiological or neural states in the brain. While there are many plausible arguments in support of the non-reducibility of mind and mental states in the philosophical literature, most are controversial, and it's never established which one Dietrich & List have in mind.

For this reason, if mentalism fits with the best scientific practice in economics, it's not altogether clear what this means. It's a truism that behavioral economists and most neuroeconomists are committed to mentalism insofar as they rely on mental-state constructs and other forms of psychological evidence in their research—but this is beside the point. Much empirical research also includes physiological and neural data. But this does not entail that empirical decision research is *inherently* reductionistic. On the contrary, many behavioral economic (and neuroeconomic) models include physiological and neural data in order to *qualify* choice-behavioral explanations that continue to employ mental-state constructs (cf. Harrison, 2016). Of course, the inclusion of physiological and neural data alongside choice-behavioral data gives rise to other complex questions and explanatory problems (Grayot, *forthcoming*). Thus, it's quite probable that researchers in other empirical domains of economics would not find Dietrich & List's conception of mentalism useful precisely because it doesn't specify how types of nonchoice data would function as explanantia.

In sum, the above analysis shows that Dietrich & List's failure to consider whether their definition of economics is congruous with others' means they haven't established a solid foundation for a philosophy-of-science analysis. For this reason, the primary problem can be read as a meta-theoretical criticism of their conception of the M-B dichotomy. While their argument for expanding the evidence base of economics is well taken, Dietrich & List's aversion to micro-foundations limits their ability to appraise different forms of psychological evidence beyond mental-state constructs. Implicitly, they seem to worry that taking an anti-behavioristic approach to economics entails explanatory reduction.[9] However, as I discuss further in section 3, the choice-theoretic foundations of economics are not uncontroversial, and Dietrich & List make the mistake of thinking that decision theory's commitment to the existence of mental states by way of functional relations entails the same for economics, broadly construed. But this is assumed and not defended carefully. For this reason, even if Dietrich & List's arguments for mentalism *were* compelling for the purposes of decision theory (which remains a matter of debate), there's no guarantee this would

[9] Oddly, they seem to agree with Gul & Pesendorfer about this: "We agree with one methodological concern voiced by Gul and Pesendorfer: the concern about the appropriate *level of explanation* in economics. Here, we think, Gul and Pesendorfer are right in criticizing the attempts of the most radical neuroeconomists to reduce decision theory to neuroscience" (2016, p. 252).

be useful for other purposes. In fact, there is no discussion of how mentalism could be put to use for empirical purposes.[10] To recap, this could be seen as a consequence of Dietrich & List's core convictions, viz. the non-reduction principle. Their worries about reductionism overshadow what I take to be a more pressing scientific concern, that is, whether mentalism is equipped to advanced empirical research. This further emphasizes why their conception of the M-B dichotomy is ill-conceived and why their defense of mentalism does not do justice to the broader debate.

3. Clarifying the domains of the M-B dichotomy in economics

In light of the above problems one may ask, what is the M-B dichotomy actually about? Dietrich & List seem to be talking past Gul & Pesendorfer rather than engaging them on a common ground. To see what's gone wrong here, the section below considers two alternative interpretations of the M-B dichotomy. I then argue that a proper philosophy-of-science evaluation must be able to distinguish between different domains of economics research—this means distinguishing theoretical from empirical goals.

3.1 *Alternative interpretations of the M-B dichotomy*

Okasha (2016) argues that decision theory is, in fact, committed to both mentalism *and* behaviorism—in particular, he argues that "the mentalistic interpretation is preferable if our aim is to use decision theory for descriptive purposes, but if our aim is normative then the behavioristic interpretation cannot be dispensed with" (2016). Okasha's account shares much in common with Dietrich & List, so we can start there.

It posits the same anti-behavioristic arguments as Dietrich & List (2016) and supposes that mental states are very likely real. By disjunction, he concludes that the mentalistic interpretation is thus preferable because "there seems to be every reason to regard an agent's credence and utility functions, as defined by Savage's theorem, as psychologically real, and as capable of explaining her preferences and choices" (2016, p. 421). However, unlike Dietrich & List (2016), Okasha's treatment involves a distinction: he identifies the relation between *utility and preferences* as separate from the relation between *preferences and choice* (Okasha, 2016, p. 410). While philosophical decision theorists are more sensitive to the possibility that utility represents something internal to the agent, economists (typically) regard utility and preference as one and the same (Okasha, 2016, p. 410). For this reason, he argues that if one wants to make in-roads toward resolving the M-B dichotomy, one must first recognize

[10] Instead of showing (or speculating) how their version of mentalism could be used for empirical purposes, they merely list-off publications which are compatible with mentalism (2016, p. 273). But this does not suffice as a demonstration of the explanatory power of mentalism.

these relations as distinct. This is because "mentalism and behaviorism… denote alternative positions about the relation between utility and preference, which are compatible with different views about how preference relates with choice" (2016, p. 416).

The above distinction is useful to the present discussion not because it strengthens the case for mentalism *per se*, but because it shows how the utility-preference relation *presupposes* mentalism when viewed from a decision-theoretic perspective (though, it still doesn't tell one what mental-state constructs functionally correspond to). Okasha's discussion indicates that prewar notions of cardinal utility, which sought to mathematically represent agent's internal states, were not altogether displaced by sophisticated expected utility (EU) models (Savage, 1954; Jeffrey, 1965). It may have been the case that early EU decision models were agnostic about the real psychological determinants of preferences; but this agnosticism did not preclude mentalism—rather, it reserved a role for utility that was seen to be prior to and independent of an agent's preference relation (a fact that Okasha links to modern neuroeconomists' attempts to identify EU optimization procedures directly in the brain, cf. Glimcher, 2004; Glimcher et al., 2005). In this way, Okasha does not see a big difference between early EU models and later non-EU models (such as cumulative prospective theory—cf. Wakker, 2010). Both EU and non-EU models suppose that the utility-preference relation is indicative of some internal optimization process; it is thus an empirical matter what processes realize this relation as captured by the mathematical models. Okasha's defense of mentalism is not actually based on an appeal to best scientific explanation; rather, it is an appeal to the historical concept of utility and to those theories of measurement which presuppose the reality of the mental.[11]

In contrast to Okasha, Guala (*unpublished*) challenges the M-B dichotomy, arguing that neither interpretation is consistent with scientific practice in choice theory and behavioral economics. To this end, Guala argues that "preferences are dispositions with a multiply realizable casual basis, which explains why economists are reluctant to make a commitment about their interpretation" (*unpublished*, p. 1). Before I get into the nuts and bolts of Guala's interpretation of preferences-as-dispositions, let's look at how he conceives the M-B dichotomy more generally.

Guala takes it as given that behaviorism is simply false from an explanatory perspective and rehearses several familiar arguments against the stronger forms of behaviorism. I will not review these arguments here except to say that Guala is less sensitive to the different variants of behaviorism, which makes his dismissal of it a little trite (*unpublished*, pp. 5-6). However, Guala does not presume that mentalism is

[11] Okasha is careful not to overstate the descriptive accuracy of mentalism: "Faced with a very specific pattern in our data—preferences that satisfy the Savage axioms—we hypothesize the existence of entities—credence and utility functions that combine in a particular way—to explain the data. This explanation is elegant, and the rule of combination—mathematical expectation—is highly intuitive. Of course, there is no guarantee that the explanation is correct, but there is strong inductive evidence in favour of it" (Okasha, 2016, p. 422).

vindicated by disjunction the way that both Dietrich & List and Okasha do. He argues that, one can grant that psychological information is indispensable for economic explanation without assenting to or endorsing mentalism: "…although behavioral economics has undoubtedly introduced psychological concepts and mechanisms in economics science, it has given no reason to interpret utility or preferences as mental" (*unpublished*, p. 5). His justification for this stems from the apparent heterogeneity of choice phenomena. In particular, he argues that choice behaviors are idiosyncratic and are realized through unique causal arrangements of external stimuli and internal psychological mechanisms in responses to that stimuli (*unpublished*, pp. 6-8). Thus, even if one *knew* the psychological mechanisms underlying an agent's choices, it's likely that this information could not be generalized for economic purposes.

This argument is based on the notion that economics studies aggregate choice (not individual decisions), and that economic explanations of choice should be based on aggregative properties somehow (cf. Dowding, 2002; Ross, 2005, 2014). This establishes why preferences are not *merely* mental. Let's now consider his reconstruction of the concept of preference to see why mentalism does not fit with "best scientific practice" in economics.

Guala adopts the view that utilities and preferences are essentially synonymous—this lends support to Okasha's claim that economists aren't interested in utility separate from preferences. Although Guala admits that the economic concept of preference is tied up in certain mental-state constructs, viz. belief (cf. Hausman, 2012), he argues that this does not warrant an exploration of utility (or related mental-state constructs) as separate from preferences. To understand why, we need to reflect on two things: the role of preferences in economics and the multiple realizability of preferences. Guala describes the role of preferences as follows:

> In general, preferences are *explanatorily relevant* and help formulating counterfactual claims about future or hypothetical scenarios, which may inform the decisions of scientists and policy-makers… [But,] even when they are genuinely explanatory…preferences do not provide information about many interesting questions. They do not tell us, for example, *how* – through which causal mechanism – a given price variation may affect [one's] behavior. They tell us that A (an agent) does B (engages in a certain behavior) in C (a set of circumstances), without saying how B and C are causally related. (*unpublished*, p. 6)

The first passage describes how preferences relate to one epistemic goal of economics, viz. prediction. Preferences are thought to enable predictions because they provide information about the "relative attractiveness" of states of affairs by identifying, at a certain level of generality, the patterns which motivate individuals to act. To say that 'A does B in circumstance C' is to describe an agent's *disposition* to act in a predictable way in C, *not* to explain how their mental states causally produce actions relevant to the achievement of B. The second passage thus indicates how preferences are

multiply realized given that, as stated above, different arrangements of external stimuli (in C) and psychological mechanisms (in A) may lead to B. Economists, argues Guala, are only interested in this causal arrangement insofar as they can identify the conditions (in both A and C) which may be generalized to predict choice-behaviors (B).

In sum, Guala's preference-as-disposition view does not deny the explanatory importance of psychological information (whether construed as unobservable mental states or as physiological states); though it strongly denies that individual psychological information is *sufficient* to enable broader choice-behavioral predictions and explanations. The preference-as-disposition view "spare[s] us the trouble of giving extremely complicated and heterogenous descriptions of the causes of behavior" (*unpublished*, p. 7).

3.2 *Different questions imply different arguments*

Clearly Okasha and Guala are up to different things, which is evinced by their divergent approaches to the M-B dichotomy. Okasha's interpretation presupposes mentalism because mental-state constructs, like utilities and credences, are built into the foundation of decision theory. But this does not constitute an argument for the existence of mental states or for the empirical adequacy of mental-state constructs; it is simply a definition, one which has experienced many overhauls over the years (cf. Starmer, 2000). For this reason, the question of whether or not mental-state constructs are descriptively accurate is an empirical issue which this account of mentalism cannot resolve (Okasha, 2016, p. 423). Contrast this with Guala's argument that preferences are dispositions and that neither behaviorism nor mentalism is tenable. This interpretation likewise stems from a definition, viz. that economics is a social science, one that does not have the same epistemic goals as psychology (this echoes Gul & Pesendorfer's line of reasoning; cf. Ross, 2014). The problem with Guala's interpretation is not that he presupposes utility and preferences to be synonymous (since this just follows from his definition of what a preference is according to rational choice theory), but that he does not take the next step, which is to consider how the preference-as-disposition fits with empirical research.

What are we to make of all of this? My claim is that there is not one but several points of contention which stem from the M-B dichotomy. I will characterize these points of contention as distinct questions. The first and most profound question is *what is economics about*? This question is clearly implicated by the dissimilarities between by Dietrich & List and Gul & Pesendorfer. There's no need to rehearse their differences again. Then, a second and more subtle question follows from the first, which is *what is the appropriate level of explanation* in economics? This, for instance, explains why Guala's account, which is amenable to psychological information for explanatory

purposes, rejects mentalism. In order to understand agents' preferences (-as-dispositions) one needs more than information beyond individual mental states; genuine explanation requires information about the external environment as well. In the above discussions, these two questions are deeply entangled, which makes getting a grip on the whole M-B dichotomy difficult.

It is for this reason I propose a third question, namely *what is the relevance of the M-B dichotomy*? This question inverts the perspective introduced by Dietrich & List by recognizing that there are fundamentally different viewpoints about the epistemic aims of economics and decision research. This suggests that before one can determine whether economics is mentalistic or behavioristic, one must first locate such debates in a scientific domain where the target phenomena and epistemic goals are well-defined. As such, the third question is not intended to resolve disagreements about fundamental questions; rather, it shows how different viewpoints may be evaluated from within a specific domain of application. In line with this third question, it's clear that what all the above accounts disagree about are the *choice-theoretic foundations* of economics. But this leaves a considerable void in one's understanding of the M-B dichotomy for none of the above accounts discuss how the dichotomy applies to empirical decision research.

4. Empirical challenges for the M-B dichotomy

Now that the primary problem has been addressed and the more fundamental questions disentangled, we're in a better position to consider the relevance of the M-B dichotomy. One way to do this is to look at what economists are actually up to in the lab (and in the field) and to consider the problems they actually face. This is doubly important for one may have the impression, following the discussions above, that behavioral economists and neuroeconomists are concerned *only* with fine-grained, micro-foundational analyses of choice behavior. Even if this were true (it isn't), this assumption oversimplifies what empirical decision research consists of. Behavioral economists and neuroeconomists alike must contend with a wide array of data to derive psychologically realistic models, and this raises many interesting and difficult philosophical questions about the epistemic aims of scientific explanation. Thus, in lieu of previously shallow appeals to scientific practice, I now consider whether, and if so how, the M-B dichotomy is useful to the purposes of empirical decision research.

4.1 *Two examples from empirical decision research*

"Dual-self" models of intrapersonal and intertemporal conflict. Research into the causes of intrapersonal and intertemporal conflict has lead economists to generate various multi-agent frameworks. Some frameworks posit temporal "selves" which

correspond to an individual's changing attitudes and interests over time, while other models link selves to underlying cognitive "systems" which regulate decision processes. The overlap of these types of frameworks in different behavioral economic contexts has given rise to highly complex models. These are most commonly known as "dual-self" models (cf. Benabou & Tirole, 2002; Bernheim & Rangel, 2003, 2004; Benhabib & Bisin, 2005; Loewenstein & O'Donoghue, 2005; Fudenberg & Levine, 2006, 2011). Dual-self models are thought to be *descriptively* superior to standard decision-theoretic models for they capture the intrapersonal dynamics of bargaining and conflict that lead persons to make suboptimal choices, such as preference reversals and other self-control problems like weakness of will and addiction. Dual-self models are innovative for they incorporate personal-level information (like pro-attitudes and first-person intentions) and subpersonal information (like emotions, low-level cognitive processes) to generate a psychologically realistic depiction of the decision maker. This enables prediction of choice behavior based on how personal-level and subpersonal-level processes are affected by changes in both the time and location of decision situations.

However—and this is crucial for the present discussion—dual-self models *need not* replicate exactly the casual-physiological processes that underlie individuals' decisions to be psychologically realistic. This is demonstrated by the fact that the terms "selves" and "systems" are often loosely defined in relation to the decision agent (i.e. in relation to underlying processes), and they are often used in concert with other mental-state constructs (i.e. utility, belief, willpower, etc.). The upshot of not rigidly fixing these terms is that it provides researchers both theoretical space to revise formal models and empirical space to explore the situational conditions under which decisions turn out to be irrational. But dual-self models give rise to a number of explanatory challenges. First, the possibility that selves and systems may, and often do, pick out different reasoning processes (some conscious, some unconscious) means that the same model may implicitly confuse different kinds of internal conflict, leading to false (or underdetermined) predictions of decision phenomena. Second, and more problematically, the comparison and evaluation of dual-self models is nearly inconceivable unless one knows how mental-state constructs relate both functionally *and* physiologically to selves and/or systems.

With these explanatory problems in mind, what could the M-B dichotomy provide researchers by way of a solution or guiding principles? Perhaps the most obvious move here is to see whether either of the dichotomy's disjuncts can offer guidance; and seeing as how behaviorism is at cross purposes with the aims of dual-self modeling, any solution is likely to be mentalistic in nature.[12]

[12] Behaviorism is at cross purposes with dual-self modeling because the act of partitioning individuals into sub-personal agents presupposes that choice data alone is insufficient to explain decision anomalies.

Thus far, the closest we've come to uncovering guiding principles for an account of mentalism has come from Dietrich & List's core convictions, what I referred to as the "mind-brain distinction principle" and the "non-reduction principle". Recall that the mind-brain distinction principle simply posits that folk-psychological concepts (which inform mental-state constructs) should *not* be conflated with physiological events in the brain, the reason being that the former are constituted by macro-phenomena and the latter by micro-phenomena. Further, the non-reduction principle qualifies the appropriate level of explanation in economics by defining decision-theoretic explanations in terms of reasons. How can these principles help us? Presumably, one could argue that the mind-brain distinction principle seeks only to avoid ontological ambiguity given the uncertain relationship between folk-psychology and individual neuro-physiology. Fair enough. But what about the non-reduction principle? Dietrich & List would have us believe that it aids researchers by identifying the *right kind* of evidence for decision-theoretic purposes:

> What makes 'higher-level" mental states, such as beliefs and desires, more explanatorily useful than 'lower-level' patterns of neural activity is precisely that they abstract away from a large number of physical details that are irrelevant, even detrimental, to the explanatory purposes at hand. (2016, p. 275)

This "abstracting away" from physical details is precisely what has lead anti-behaviorists (Okasha and Guala included) to endorse functionalism, and it seems to be a common refrain for functionalists across the sciences in general.[13] But is *more* functionalism what economists really need? Dual-self models are controversial in large part because they attempt to model the intrapersonal dynamics of decision-making as a strategic interaction between different internal structures (which the functionalist is not required to define). But, because it is not agreed upon what distinguishes different 'internal structures', the modeling of intrapersonal dynamics involves a very high-degree of functional ascription (not just to standard mental-state constructs, but also to the low- and mid-level physiological processes which subvene mental-state constructs). Functionalists tend to view this as an asset because it alleviates the burden of addressing tough ontological questions. But if the goal of economics, construed as a positive science, is to better understand how intrapersonal conflict is related to choice behavior, then some tough questions cannot be avoided. For this reason, the case can be made that functionalism is part of the explanatory problem rather than part of the

[13] While Dietrich & List's endorsement of functionalism has been clearly stated, Okasha's and Guala's respective endorsements are subtly different. Okasha never discusses functionalism explicitly, though his defense of mentalism for descriptive purposes makes no sense *unless* he indeed endorses some account of functionalism. By contrast, Guala is more careful in his elucidation of functionalism (*unpublished*, pp. 5-7, 15-18).

solution (Grayot *forthcoming*). In the next sub-section, I return to address why cases like these are particularly damning for functionalism.

The take away message here is not that mentalism is inherently flawed or that functionalism is a dead-end (one may simply change their convictions and adopt better criteria for what counts as a functional explanation). The problem is that one cannot resolve these sorts of empirical challenges from the perspective of the M-B dichotomy: the dichotomy operates at the level of meta-theory and thus cannot tell you what kind of mentalistic evidence is the right kind, let alone tell you how to assimilate different types of evidence for the purposes of scientific explanation.

Information processing models of economic behavior. Some neuroeconomic research into how humans process information applies optimization models directly to neurons and neural modules in the brain. Within these types of models, the individual, i.e. their choice behaviors, play little to no role in the explanation of how brains satisfy rewards, except as part of a larger feedback mechanism (which is often left underdetermined, cf. Grayot *forthcoming*). Dietrich & List are right that radical neuroeconomic accounts like these don't fully explain how individuals make decisions because such models do not account for "higher-level" factors like reasons and environmental phenomena that motivate persons to act. But, from the perspective of the M-B dichotomy, it's also important not to confuse models that are strictly interested in human neurobiology (dubbed "neuro-cellular economics") with neuroeconomic models which are designed to show how individual choice behaviors relate to more generalized brain processes (dubbed "behavioral economics in the scanner"). While these alternative neuroeconomic approaches are certainly interested in the micro-foundations of choice behavior, this doesn't mean that they exclude other forms of psychological data which are part of the mentalist repertoire.[14] With this in mind, the case can be made that not all neuroeconomic models are so radical that they are irrelevant to the current discussion.

Closely related to the above example, neuroeconomists have investigated whether the intrapersonal conflict modeled by dual-self models is an indication underlying dual systems in the brain. Unlike behavioral economists who portray this conflict by way of functional relations between mental-state constructs, some neuroeconomists believe that there are two "systems" at work, whose information networks generate a kind of physiological conflict between neural modules (cf. McClure et al., 2004; McClure et al., 2007; Brocas & Carrillo, 2008, 2014). However, other neuroeconomists think this idea is wrong-headed. Alternative research indicates that reward and information networks aren't discretely dualistic, and that the processes involved in decision-making are distributed throughout the entire brain (cf. Glimcher et al., 2007;

[14] Moreover, a convincing case has been made that these different styles of neuroeconomics seem to be converging, which means that the interpretation of neuroeconomics as radically reductive may not be representative of where research is eventually headed (Vromen, 2011).

Kable & Glimcher, 2007). For this reason, the latter argue, it's better to think of the brain as a unitary decision system, or they don't speak of systems at all (cf. Rustichini, 2008).

One may ask, how does all this relate to M-B dichotomy? –isn't this a physiological dispute, a disagreement about how the brain works? Perhaps it is. But this doesn't mean that resolving the dispute depends solely upon gathering more physiological and neural data. Quite the opposite. Many such debates among neuroeconomists arise in the first place because of preexisting discrepancies about the meaning and interpretation of economic concepts—whether the data collection methods of neuroeconomists is invalid is another issue. By way of example, consider the concept of *utility*.

The idea (or hope) that utility is literally computed in the brain has been the holy grail of nearly all major neuroeconomic research. Yet, there is no consensus what the best interpretation for the economic concept of utility is (a measure of satisfaction? a representation of changing hedonic states? an analog of monetary value?); for this reason, there is no consensus about where to *look* for utility in the brain, let alone consensus about whether utility is "computed" in any meaningful, economic sense of the word (cf. Vromen, 2010). These conceptual discrepancies about the meaning of the concept of utility have a direct impact on how experiments are developed, how models are constructed, and how data are interpreted (Fumagalli, 2013).

It may seem that we are beating a dead-horse at this point: if one grants that the M-B dichotomy is ill-equipped to answer difficult empirical questions? The reason is that many of the same researchers who defend mentalism invoke non-reduction as a justification for not considering neuroeconomic practices among the body of empirical work which is relevant to the M-B dichotomy. Albeit, that researchers do not find the pursuit of micro-foundations relevant to their own ends does not constitute an argument against the inclusion of neuroeconomics within the broader domain of empirical decision research. Moreover, what both examples above demonstrate is that those engaged in debates which stem from the M-B dichotomy are stuck at the meta-theoretical level.

4.2 *Further limitations for functional explanations*

Thus far I've attempted to bracket my discussion of functionalism by considering only those contexts in which is pertains to economics. To recap, Dietrich & List argue that economics is committed to mentalism in virtue of the functional roles the mental states play: "our best economic theories of individual decision making are certainly committed to certain relations or functions that formally play preference-or-belief roles… And since playing those roles is indicative of mental states, we should, at least provisionally, treat the 'mental-state relations or functions' posited by our theories as corresponding to real mental states of the agents in question" (2016, p. 269). However,

I suggested in section 2.2 that this understanding of functionalism is rather thin, that it may not be as advantageous as Dietrich & List presume it to be. I'd like now to reconsider some further objections to functionalism, first from a more traditional philosophical perspective, and second from the perspective of the study of mind and human cognition.

An intrinsic challenge for any functionalist account of behavior that utilizes mental-state ascriptions is how to determine what counts as a "mental state". As I argued in section 2.1 and demonstrated in 4.1, functional interpretations of mentalism in economics and decision theory tend to lack well-defined criteria for demarcating mental states, and for this reason, it's not always obvious what does the explanatory work in a decision-theoretic model. However, to portray this as merely a criterial problem would oversimplify the complexity of questions that underlie functional explanations in general. The specification of the *right* criteria depends on far more than identifying the "functionally relevant properties" of some phenomena (Dietrich & List, 2016, p. 275). It depends, *in primis*, on how one construes functional theory, and thereby, how one conceives the logical relationship between functional statements—whatever those happen to be—and scientific theory.[15] To make this point salient, let's consider two problems which are associated with the "holism" critique of functionalism in philosophy of mind (Levin, 2017). Simply put, the holism critique posits that mental states are often *interdefined* with other mental states, each of which are defined by functional relations. Here is a brief description of how this works:

> Functionalists hold that mental states are to be characterized in terms of their roles in a psychological theory—be it common sense, scientific, or something in between—but all such theories incorporate information about a large number and variety of mental states… In addition, differences in the ways people reason, the ways their beliefs are fixed, or the ways their desires affect their beliefs—due either to cultural or individual idiosyncrasies—might make it impossible for them to share the same mental states. These are regarded as serious worries for all versions of functionalism. (Levin, 2017, §5.1)

The holism critique gives rise to potentially many problems for functionalism (cf. Block, 1980; Rey, 1997); however, for the purposes of this paper, I concentrate on just two challenges.

The first challenge concerns how researchers recognize and intelligibly track the ascription of folk-psychological labels to mental states. Because folk-psychological practices are believed to be inherently intentional and commissive (Dennett, 1991;

[15] There are multiple varieties of functionalism in the history of philosophy of science, and each can be defined by a unique logical relationship between its attendant properties (construed as predicates), statements or propositions, and theory (cf. Quine, 1976; Lewis, 1966). For instance, Levin (2017) distinguishes four types of functional theory: *machine state functionalism*, *psycho-functionalism*, *analytic functionalism*, and *role-functionalism*.

McGeer, 2007; 2015), the problem of tracking what mental states pick out turns out to be a normative issue, rather than a descriptive one (this seems to be part of Dietrich & List's motivation for pursuing mentalism in the first place). While many see this as a good reason for adopting functionalism, what is often missing from functional explanations is an account of how folk-psychological labels actually serve their role—which is to say, how they identify mental-state *types*, and further, how one can trust that they pick out the same types in different circumstances (White, 1986, 2007; Block, 2002). Thus, without further empirical support, e.g. socio-linguistic or anthropological evidence of, any claim in support of folk-psychological practice to justify the ascription of mental states runs the risk of being recursive or *ad hoc*.

Of course, this first horn of the critique is only a problem if one has a demanding notion of explanation in mind. Some functionalists may find the holism critique to be a non-issue because they take it for granted that the commonsense nature of folk-psychology is sufficient to determine what counts as a mental state. To wit, one may argue that the motivation for adopting functionalism in the first place is that it does not require stringent criteria to provide approximate explanations, or better yet, to make useful predictions. Again, this may be an adequate reason for adopting functionalism for *some* decision-theoretic purposes. But that is not what *this* paper is concerned with. My argument that M-B dichotomy does not advance or improve empirical research in economics because neither mentalism (nor behaviorism) are suited to these purposes assumes that behavioral economics and neuroeconomics aspire to be psychologically realistic (Angner and Loewenstein, 2012).[16] This leads me to the second horn of the holism critique.

A more problematic issue for functional explanations, beyond the matter of determining what counts as a mental state, can be seen as a down-stream implication of the mind-body problem. It is this: if one endorses the existence of mental states as entities separate from physiological states, then one must also endorse the interaction of mental states. This gives rise to an altogether new set of difficult questions. Some have referred to this as the "mind-mind" problem to convey that functional descriptions of non-physiological processes compounds the challenges of psychological explanation because it renders uncertain how cognitive processes produce human behavior (Jackendoff, 1988). However, even if one is not concerned with mental causation *per se* or with parsing supervenience relations, the mind-mind problem raises important epistemic issues for functional explanations because it raises the possibility that researchers of the mind cannot (easily) distinguish between action-producing mental states within the same individual. Various examples of this point have been

[16] This is not an uncontroversial assumption. I'm thus willing to concede that behavioral decision research can aspire to be psychologically realistic while appealing to different standards of explanatoriness. However, the issue of what psychological realism entails for economic explanations is beyond the scope of this paper.

discussed in the context of representational and computational theories of mind and have thus given rise to alternative conceptions of cognition (Rowlands, 2010; cf. Verala et al., 2017). If one looks to the current state of research in cognitive science and philosophical psychology, there is a growing confidence in the view that the brain, the body, and the environment constitute extensive cognitive systems (cf. Menary, 2010).[17] The complexity of these cognitive systems of course gives rise to new problems and hard(er) questions about the nature and properties of the mental (Adams & Aizawa, 2008). But these movements away from the traditional metaphysics of the mind have permitted researchers to more carefully explicate the functional relations between cognitive processes and systems (whatever those may be) and human action (Wheeler, 2010; Hurley, 2010). It is perhaps for this reason that labels like "mentalism" and "behaviorism" are falling rapidly out of fashion with philosophers and cognitive scientists.

The holism critique thus demonstrates the *epistemic* limitations of functionalism. The first instance of the critique inquires how functional explanations can discern mental states without invoking recursive definitions or criteria, viz. other mental states or functional relations. This raises challenges both for the identification of mental-state *types*, and for the comparison of mental states between individuals. The second instance of the critique inquires how functional explanations identify and track the interaction of distinct mental states within the individual. This raises challenges above and beyond the first set for it puts pressure on how researchers achieve scientific understanding of cognitive processes. Thus, if either horn of the holism critique holds, then any psychological explanation predicated on functional property ascriptions is going to be epistemically suspect unless it is based on well-defined criteria. The importance of these challenges is that neither hinges on the causal efficacy of mental states or their ontological status.

How do these philosophical objections to functionalism relate to this paper's analysis of the relevance of the M-B dichotomy? There is a strangely sanguine attitude among economists and especially decision theorists toward functionalism, and this suggests that either they aren't informed about the limitations of functional explanations or that they don't see these limitations as legitimate problems for economic research. I've now provided two reasons for thinking that functionalism may not be epistemically reliable for the purposes of psychological explanation in general. However, it's important that one not construe the explanatory limitations of functionalism as a reason to abandon the use of mental-state ascriptions or to embrace eliminative materialism. This is not my intention. Indeed, these philosophical objections serve to remind researchers that *if* they are going to invoke functionalism to defend the

[17] I purposely use the term "extensive" here so as not to favor any particular theory of embodied, extended, or enactive cognition, which are the popular candidates for non-Cartesian theories of mind and human cognition.

explanatory role of mental states for economic purposes, they ought to know what the stakes are. This, recall, is Guala's reasoning for not endorsing mentalism: if what economists care about are preferences, and if the *real* causal basis of preferences depends upon a confluence of individual psychological and physiological states, which may be multiply realized by different external conditions, then only a functionalist account of preferences will capture the dispositional nature of individual choice-behaviors. Interestingly, this view is not altogether different from Dietrich & List's understanding of functionalism. The major difference is that Dietrich & List (and to a lesser extent, Okasha) presuppose that this functional relationship between preferences and choice-behavior is further mediated by another kind of functional entity, viz. mental states, and it's this kind of functional explanation that is problematic. Of course, Dietrich & List can get off the hook by arguing that, for their purposes, it doesn't matter what mental states *actually* are. Fair enough. But, if one believes that economics is a positive science and should aim to be descriptively realistic, then they must appreciate the limitations of functionalism to discern and qualify psychological evidence.

5. Concluding remarks

I've argued that there are two main problems with how philosophers and methodologists understand the M-B dichotomy in economics and decision research. The primary problem is that it's not certain what the M-B dichotomy is really about because most debates are ill-conceived: researchers may have different motivations for endorsing mentalism and/or for opposing behaviorism, but recent debates reveal these motivations to be entangled. To make this point vivid, I investigated how Dietrich & List (2016) analyze and respond to the behaviorism of Gul & Pesendorfer (2008). Though Dietrich & List provide compelling arguments for expanding the evidence-base of economic models beyond choice-behavioral data, they presuppose that economics and decision theory have the same scientific goals and the same ontological commitments. In effect, they talk past Gul & Pesendorfer. Dietrich & List infer that mentalistic models "fit" with scientific practice in economics. I challenged this inference on the grounds that they provide no criteria for discerning mental states; further, they do not explain how mentalism advances or improves empirical research in economics.

The upshot of the primary problem is that it demonstrates how to identify and remedy confusion about the M-B dichotomy by distinguishing arguments that appeal to choice-theory from arguments that appeal to scientific practice. It was for this reason that I explored alternative interpretations of the dichotomy by Okasha (2016) and Guala (*unpublished*). Okasha's account better articulates reasons for endorsing mentalism given the history of decision-theory, whereas Guala's account better articulates reasons for rejecting behaviorism given the scientific goals of economics. Yet, I

argued that these two interpretations are incomplete if one wishes to avoid a theoretical stalemate.

The secondary problem is understanding how the M-B dichotomy advances and improves empirical research in economics. Specifically, I investigated how mentalism functions in the context of decision research by examining two controversial modeling techniques, one from behavioral economics and one from neuroeconomics. The goal here was to find out whether adopting mentalism would yield methodological benefits. I surmised that its benefits are redundant. Adopting mentalism does little to improve research in these areas because it does not have the tools to make sophisticated judgments about psychological and physiological data.

The upshot of the secondary problem is two-fold. First, it demonstrates the limitations of the M-B dichotomy to improve or advance empirical research. If it is relevant to scientific practice, I argue, it is so at a very general level only. Second, my diagnosis of the secondary problem reveals the degree to which functionalism is an empirical concern, especially if one invokes functionalism as a justification for mentalism. Many philosophers and methodologists of economics seem to overlook the fact that functional explanations can lead to misunderstandings about how the mind-brain works. This can be detrimental to the study of individual decision-making.

The M-B dichotomy is no doubt important. It's important because it tells us something about how philosophers and methodologists conceive economics as a science. It's also important because it indicates the status of philosophy-of-science research. I conclude with a word of caution: if researchers are going to continue to leverage the M-B dichotomy, they must recognize its limitations. The dichotomy is indeed potentially useful for making coarse approximations and for locating one's intuitions in debates. Beyond these purposes, researchers should be wary of arguments which invoke the terminology of "mentalism" and "behaviorism".

6. Bibliography

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. John Wiley & Sons.

Angner, E., & Loewenstein, G. (2012). Behavioral economics. In Mäki, U. (Ed.), Philosophy of Economics (pp. 641-689). Amsterdam: Elsevier

Backhouse, R.E., & Medema, S.G. (2009). Defining economics: the long road to acceptance of the Robbins definition. *Economica* 76, 805-820.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3), 871-915.

Benhabib, J. & Bisin, A. (2005). Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions. *Games and Economic Behavior*, 52(2), 460-492.

Bernheim, B.D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review*, 94(5), 1558-1590.

Block, N. (1980). Troubles with functionalism. *Readings in Philosophy of Psychology*, *1*, 268-305.

Block, N. (2002). The harder problem of consciousness. *The Journal of Philosophy*, *99*(8), 391-425.

Brocas, I., & Carrillo, J.D. (2008). The brain as a hierarchical organization. *The American Economic Review*, 98(4), 1312-1346.

Brocas, I., & Carrillo, J.D. (2014). Dual-process theories of decision-making: A selective survey. *Journal of Economic Psychology*, 41, 45-54.

Bruni, L., & Sugden, R. (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. *Economic Journal* 117(516), 146-173.

Caplin, A., & Schotter, A. (Eds.). (2008). *The Foundations of Positive and Normative Economics: A Handbook*. New York: Oxford University Press.

Camerer, C. (2008). The case for mindful economics. In Caplin A., & Schotter A. (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 43-69). New York: Oxford University Press,

Camerer, C., Loewenstein, G., & Prelec, D. (2004). Neuroeconomics: Why economics needs brains. *The Scandinavian Journal of Economics*, 106(3), 555-579.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1), 9-64.

Camerer, C., Loewenstein, G., & Rabin, M. (Eds.). (2011). *Advances in Behavioral Economics*. Princeton University Press.

Clarke, C. (2016). Preferences and positivist methodology in economics. *Philosophy of Science* 83, 192-212.

Clarke, C. *Unpublished*. Interpreting economic models: four pledges for mentalists and behaviourists.

Davis, J. B. (2006). The turn in economics: neoclassical dominance to mainstream pluralism?. *Journal of Institutional Economics*, 2(1), 1-20.

Dennett, D.C. (1991). Real patterns. *The journal of Philosophy*, *88*(1), 27-51.

Dowding, K. (2002). Revealed preference and external reference. *Rationality and Society* 14, 259-284.

Dietrich, F., & List, C. (2016). Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics & Philosophy*, 32(2), 249-281.

Edwards, J. (2016). Behaviorism and control in the history of economics and psychology. *History of Political Economy*, 48(suppl_1), 170-197.

Fudenberg, D., & Levine, D.K. (2006). A dual-self model of impulse control. *The American Economic Review*, 96(5), 1449-1476.

Fudenberg, D., & Levine, D.K. (2011). Risk, delay, and convex self-control costs. *American Economic Journal: Microeconomics*, *3*(3), 34-68.

Fumagalli, R. (2013). The futile search for true utility. *Economics & Philosophy* 29, 325-347.

Glimcher, P.W. (2004). *Decisions, Uncertainty and the Brain*. Cambridge MA: MIT Press.

Glimcher, P.W., Dorris, M.C., & Bayer, H.M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games and Economic Behaviour* 52, 213-256.

Glimcher, P.W., Kable, J., & Louie. K. (2007). Neuroeconomic studies of impulsivity: now or just as soon as possible? *American Economic Review* 97(2), 142-147

Grayot (*forthcoming*). From selves to systems: on the intrapersonal and intraneural dynamics of decision making. *Journal of Economic Methodology.*

Guala, F. (2012). Are preferences for real? Choice theory, folk psychology, and the hard case for commonsensible realism. In Economics for Real (pp. 151-169). Routledge.

Guala, F. (*Unpublished*). Preferences: neither behavioral nor mental.

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. In Caplin A., & Schotter A. (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 3-42). New York: Oxford University Press.

Hands, D.W. (2009). Economics, psychology and the history of consumer choice theory. *Cambridge Journal of Economics*, 34(4), 633-648.

Hands, D.W. (2013). Foundations of contemporary revealed preference theory. *Erkenntnis* 78, 1081-1108.

Hands, D.W. (2014). Paul Samuelson and revealed preference theory. *History of Political Economy*, 46, 85-116.

Hausman, D. M. (2008). Mindless or mindful economics: a methodological evaluation. In Caplin A., & Schotter A. (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 125-151). New York: Oxford University Press.

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

Heukelom, F. (2014). *Behavioral economics: A history*. Cambridge University Press.

Hurley, S.L. (2010). Varieties of externalism. In Menary, R. (Ed.), *The Extended Mind* (pp. 101-153). MIT Press.

Jackendoff, R. (1986). Conceptual Semantics in Meaning and Mental Representations. *Versus*, (44-45), 81-97.

Jeffrey, R. (1965). *The Logic of Decision*. Chicago: Chicago University Press.

Kable, J.W., & Glimcher. P.W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12), 1625-1633.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 18-36.

Levin, J, (2017). Functionalism., In Zalta E. (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University, Winter 2017 Edition.

Lewin, S. B. (1996). Economics and psychology: Lessons for our own day from the early twentieth century. *Journal of Economic Literature*, 34(3), 1293-1323.

Lewis, D. (1966). An Argument for the Identity Theory. *Journal of Philosophy*, 63, 17–25.

Lewis, D. (1983). *Philosophical Papers* (Vol 1). Oxford: Oxford University Press.

Lowenstein, G. (1992). The fall and rise of psychological explanations in the economics of intertemporal choice. *Choice Over Time* (pp. 3-34). Russell Sage Foundation, New York.

Loewenstein, G., & O'Donoghue T. (2005). Animal Spirits: Affective and Deliberative Processes in Economic Behavior. *CMU Working Paper*.

McClure, S.M., Laibson, D.I., Loewenstein, G., & Cohen J.D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695), 503-507.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience*, 27(21), 5796-5804.

McGeer, V. (2007). The regulative dimension of folk psychology. In Hutto, D.D., & Ratcliffe, M. (Eds.), *Folk Psychology Re-assessed* (pp. 137-156). Springer, Dordrecht.

McGeer, V. (2015). Mind-making practices: the social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, *18*(2), 259-281.

Menary, R. (Ed.). (2010). *The extended mind*. MIT Press.

Okasha, S. (2016). On the interpretation of decision theory. *Economics & Philosophy*, 32(3), 409-433.

Quine, W.V. (1960). *Word and Object.* Cambridge. MIT Press.

Quine, W.V (1976). Two dogmas of empiricism. In *Can Theories be Refuted?* (pp. 41-64). Springer, Dordrecht.

Rabin, M. (1998). Psychology and Economics. *Journal of Economic Literature*, 36(1), 11-46.

Rabin, M., (2002). A perspective on psychology and economics. *European Economic review*, *46*(4-5), 657-685.

Rey, G. (1997). *Contemporary philosophy of mind*: Blackwell

Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press.

Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Ross, D. (2011). Estranged parents and a schizophrenic child: choice in economics, psychology and neuroeconomics. *Journal of Economic Methodology* 18, 217-231

Ross, D. (2014). Economic versus psychological models of bounded rationality. *Journal of Economic Methodology* 21, 41-27.

Rustichini, A. (2008). Dual or unitary system? Two alternative models of decision making. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4), 355-362.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: Dover.

Sent, E.M. (2004). Behavioral economics: how psychology made its (limited) way back into economics. *History of Political Economy*, *36*(4), 735-760.

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

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences*, 23(5), 645-665.

Varela, F.J., Thompson, E., & Rosch, E. (2017). *The Embodied Mind: Cognitive Science and Human Experience*. MIT press.

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

Vromen, J. (2011). Neuroeconomics: two camps gradually converging: what can economics gain from it?. *International Review of Economics*, 58(3), 267-285.

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

White, S. (1986). Curse of the Qualia. *Synthese*, 68, 333-368.

White, S. (2007). Property Dualism, Phenomenal Concepts, and the Semantic Premise. In Alter, T. & Walter. S. (Ed.), *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism* (pp. 210–248). Oxford University Press on Demand.

Wheeler, M. (2010). In defense of extended functionalism. In Menary, R. (Ed.), *The Extended Mind* (pp. 245-270). MIT Press.

Chapter 3

The quasi-economic agency of human selves[1]

1. Introduction

With all the recent advances in behavioral economics (including advances in experimental psychology and neuroeconomics) perspectives about economic agency have shifted away from the traditional, neoclassical conception of the rational agent. It is now recognized that humans are boundedly rational, which means that persons typically do not think and behave like *homo economicus* agents. Among the methodologies for modeling boundedly rational individuals, *multiple-self* models have gained considerable popularity as tools for representing the dynamics of intrapersonal choice under various conditions and constraints. Multiple-self models typically work by isolating features endogenous to individuals that motivate them to act in different ways. Generally, these features are taken to correspond to autonomous structures within the individual and, as such, are modeled *as if* they were independent agents (that is, independent agents who can reason together). Some multiple-self models conceive of selves as temporal agents (Thaler & Shefrin, 1981; Laibson, 1997; O'Donoghue & Rabin, 1999, 2001), whereas other models conceive of selves as cognitive processes in, or mapped onto, the brain (Benhabib & Bisin, 2004; Jamison & Wegner, 2009; cf. Brocas & Carrillo, 2008, 2012). Yet, there is another sense in which individuals are thought to contain selves which is not well-represented in the economics literature:

According to Don Ross (2005, 2006, 2010) individual persons are complex aggregations of selves. These selves arise in response to external pressures to regulate individual behaviors, and they enable the tracking of public norms and conventions. In contrast with the many approaches to multiple-self modeling in behavioral decision research that focus explicitly on the cognitive-psychological basis of intrapersonal conflict, Ross argues that selves are not reducible to brain functions since they are conjunctions of neural and social activity, spanning the brain, body, and environment. In this way, selves are not the type of object that can be studied in isolation of the systems of which they are part. Rather, they are the virtual embodiment of individual and cultural *narratives* that are cultivated over the course of a person's biography. It is because individuals have selves that they can engage in and maintain interpersonal relationships in the first place. It is thus believed that enculturated selves enable

[1] For the publication, see Grayot, J. (2017). The Quasi-Economic Agency of Human Selves. *Œconomia. History, Methodology, Philosophy*, (7-4), 481-511. It can be found at: https://journals.openedition.org/oeconomia/2790

persons to navigate complex social networks free of the computational burden of continuously problem-solving coordination dilemmas.

Ross's conception of enculturated selves marks an important contribution to the study of economic agency for it challenges the idea that individuals are, or should be regarded as, centers of decision-making. His anti-individualistic perspective, which he has described as "non-anthropocentric" neoclassicism, presents a view of the individual person that is bound by social and institutional constraints (2005). In this regard, his understanding of multiple selves touches upon familiar projects of bounded *individuality* and the economics of identity which have also been discussed at length by Davis (2003, 2011). Yet, Ross's account of selves is interesting because it forges novel links with the cognitive and behavioral sciences in ways that other accounts have not. It proposes that selves played (and continue to play) a critical role in the evolution of human social intelligence, namely through the sending and receiving of linguistic signals. Thus, for Ross, the economic function of selves as a behavior stabilization technology is tied up in their ability to recognize and respond to linguistic and other signaling conventions. This makes for a much richer, albeit more convoluted, account of selves than others discussed in the economics literature.

In this paper, I investigate the roles that selves play within Ross's anti-individualistic framework, and I identify separate projects that may be (and in some instances, have been) attributed to him based on different interpretations of what selves are believed to be. To this end, I distinguish three different roles for selves—these are *evolutionary*, *narrative*, and *economic*—and I argue that these roles contribute to two distinct, but overlapping, projects in Ross's broader philosophy of economics.[2] One project is to give an account of the emergence of human socio-cognitive abilities, and to show how those abilities are necessary for market behavior. Another project is to give an account of economic agency that is both amenable to neoclassical economic methodology while remaining sensitive to the fact that humans are not ideal economic agents. With these three roles and two projects in mind, my aim is to show that there is tension underlying these projects, but that it's not clear where these tensions arise precisely because of how selves are multiply understood and used to defend these projects. I will argue that, while it is entirely possible to conceive of selves in accordance with any of the roles that I have attributed to Ross—primarily because each role conceives selves as black boxes—we should not presume that the black-box function of selves serves the same purposes for both of his projects.

[2] These roles will occasionally be equated with explanatory synonyms given how Ross has utilized them in his broader framework. Although I stick to the categories "evolutionary", "narrative", and "economics", sometimes these labels will coincide with alternative labels. For instance, the evolutionary role is sometimes described as "biological"; the narrative role is sometimes described as "biographical", and the economic role is sometimes described as "mathematical".

The bulk of my investigation analyzes arguments that are developed in his *Economic Theory and Cognitive Science: Microexplanation* (Ross, 2005), and subsequent articles (Ross, 2006, 2007a, 2007b) where additional support is provided for the evolutionary and economic basis of selves. In short, the aim of this investigation is not only to clarify what economists can (and can't) expect to do with an account of selves like Ross's, but it further indicates what promising work lies ahead for multiple-self models that are not strictly based on psychological or brain activity.

This paper has the following structure. In Section 1, I provide the background and context for Ross's conception of selves within his anti-individualist framework. Here I flesh out three main roles for selves. In Section 2, I make the case that there are, in fact, separate projects going on here, and I show that not all the resources from one project may be outsourced to another project without violating some of Ross's core convictions. In Section 3, I recommend a few ways to reconcile these different projects, and discuss the ways that an anti-individualistic framework can interface with disciplines outside economics—here I contrast my view against others who have commented on Ross's work. Section 4 concludes.

2. Non-anthropocentric neoclassicism and multiple-selves

Ross' describes his philosophy of economics as both "non-anthropocentric" (2005, pp. 19-22) and "neo-Samuelsonian" (see also Ross, 2014).[3] His project envisions a return to the Samuelsonian tradition in economics where individual psychology is bracketed and excised from the study of markets and their effects on individual behaviors. As a preliminary discussion, this characterization of his project is succinct—it is meant only to provide the groundwork for an investigation of his conception of selves. This will help us to understand why his project does not permit a single interpretation for selves.

2.1 *Economic agency in an anti-individualistic economics*

Ross' non-anthropocentric neoclassicism is predicated on the rejection of two principles commonly associated with microeconomic methodology: these are called "social atomism" and "microeconomic individualism" (2005, pp. 221-223). Social atomism is the thesis that persons are ontologically basic, which means that social phenomena can be understood in terms of the actions of individual persons, and that social reality is irreducible beyond persons (hence they are social 'atoms'). Microeconomic individualism builds upon social atomism by presuming that utility functions are intrinsic properties of persons. This more or less captures the "Robinson Crusoe" picture of

[3] Ross refers to his position as "Samuelsonian" in (2005) but adopted the term "neo-Samuelsonian" in (2014) in light of commentary made by Hands (2008).

economic agency, by which persons are assumed to enter the world with pre-given utility functions for goods prior to encountering a socialized market. "Normative individualism", by contrast, makes no claims about social ontology or the sources of individual utility; it is individualistic only insofar as it views persons as morally autonomous agents, whose intrinsic worth should be taken into consideration for matters of policy or justice (Ross, 2005, pp. 220-222). The problem, Ross tells us, is that economists have tended to conflate microeconomic individualism (which logically implies atomism) with normative individualism: this gives rise to puzzling questions about the essential properties of economic agents. Who or what is an economic agent? —persons are thought to be agents, but what about firms? countries? Also, what cognitive properties do economic agents have? —neoclassical economic models assume they have perfect information and powerful computational abilities, but this is obviously not true of persons in real life. By adopting a *descriptive* (as opposed to normative) *anti-individualistic* approach to economics, Ross argues that any intentional system, human or non-human, can be modeled as an economic agent, and thus denies that there is anything uniquely human about economic agents.

Non-anthropocentric neoclassicism thus draws a sharp distinction between individuals and economic agents. This has two important corollaries in Ross's anti-individualistic philosophy of economics. First, he argues that preferences should not be interpreted as real computations that take place inside the minds (or brains) of persons (2005, p. 108). Utility functions—which are *ad hoc* valuations that numerically represent choices and preferences—are not properties of persons; they are the properties of economic agents which persons may approximate via the regulation of their behavior under specific conditions (cf. Pettit, 1995). While this first corollary follows naturally from Ross's rejection of microeconomic individualism, decades of evidence from experimental and behavioral economic research into intrapersonal choice have also demonstrated that persons are not ideal economic agents given their tendencies to change and/or reverse preferences over time. For this reason, he states that, "…if agents are identified with utility functions, then the biography of a typical person can't be the biography of a single (diachronic) economic agent" (Ross, 2005, p. 156).

If utility functions are just properties (i.e. numerical representations) of economic agents, and if any well-behaved intentional system can be ascribed a utility function, then preferences should be understood as points of reference for the behavioral output of whatever sociological and institutional pressures constrain the behavior of complex systems. For this reason, Ross reminds us that:

> …neoclassical theory, properly understood, is not directly about any specific kind of behavior, and rests on no ontological commitments more definite than the idea that agents can be analytically distinguished from one another (2005, p. 197).

> Agents need not be internally simple—as people are not—so they can, in principle, be firms or households or whole countries or any other sort of unit that acts teleologically…. (2005, p. 198)

Without going into further depth about the general concept of agency, we can surmise that the economic agent, understood as a purely theoretical object, has neither ontological nor psychological properties built into it, and so, warrants the extrication of human properties from it. This allows Ross to reaffirm the neo-Samuelsonian interpretation of preferences as exogenously given: preferences should reflect the aggregative influence of social norms and institutional pressures upon individuals, not their inner cognitive architectures. This anticipates the second corollary of Ross anti-individualism.

The second corollary concerns economics as a science separate from psychology. In justifying the separateness of the disciplines, he argues (citing Lionel Robbins) that economics ought to be viewed as the "abstract logic of choice", not as the study of causal mechanisms of individual choice. He states that, "the implication of the separateness thesis as Robbins justifies it is that choice, as a psychological process, is a black box that, so far as economics is concerned, is supposed to be deliberately left shut" (Ross, 2005, p. 91). Yet, a further justification for the separateness thesis could be linked to Ross's skepticism about the etiology of individual choice behavior as determined by mental content. In adopting Daniel Dennett's intentional stance functionalism (Dennett, 1987), Ross eschews the traditional *internalist* conception of individual choice, which supposes that propositional attitudes have causal power to induce action. Ross, following Dennett, emphasizes that economics is about behavioral regularities, and that we can better study these regularities once we learn how language networks structure and constrain social dynamics (Ross 2005, pp. 61-70).

According to this interpretation of economics as separate science, preferences should be distinguished from the study of the internal mechanics of decision-making as understood by the neuroscientist or behavioral economist.[4] Thus, anti-individualism does not deny that the interaction of real persons gives rise to complex social phenomena, but it emphatically denies that facts about how individual persons make decisions—information about their cognitive architectures and the wiring of their brains—are sufficient to explain the outcomes of their social interactions.[5]

---

[4] While Ross argues that the mechanics of individual choice are idiosyncratic and not generalizable, his claim that the groundwork for a theory of the economic agent should include a commitment to some form of externalism is separately informed by his philosophical commitments.

[5] In bringing together these separate motivations and in justifying the divorce between economics and psychology, Ross sharply claims that he does not *a priori* deny that there is such a thing as faculty introspection, but he denies that it is a stable and direct source of evidence (2005, p. 228). This take on introspection also clarifies his position with regard to the economic

This characterization of Ross's anti-individualistic philosophy of economics is important because it demonstrates why a concept of economic agency that strives to mirror human persons is potentially misleading—viz., because the economic agent, understood as a purely theoretical object, has neither ontological nor psychological properties built into it. None of the tenets of neoclassicism, according to Ross, require that rational agency apply directly to human persons. Likewise, processes and mechanisms that occur 'below' the level of the individual may also be modeled as economic agents, provided that such processes are the sort of unit that act teleologically. Below, I show how we get from descriptive anti-individualism to an account of selves.

2.2 *Three interpretations of selves*

By sharply distinguishing economic agents from flesh-and-blood individuals, Ross is forced to explain how it is that individual persons maintain stable behavior. He argues that individual persons are complex aggregations of behavioral profiles that are determined by social interactions. In contrast with most cognitive-psychological approaches to multiple-self modeling in behavioral economics, Ross argues that selves are not reducible to neural processes or modules in the brain, and so are not the type of object that can be studied in isolation of the social systems of which they are a part. This idea is influenced largely by Dennett's conception of a "real pattern" (1991a)—i.e. mental *and* social constructions generated by our beliefs about ourselves, beliefs that are regulated by sense-making norms, and the actions those beliefs produce in others via public language (Ross, 2005, p. 18). In this way selves are the product of interpersonal experiences that are imbued with meaning through everyday practices and ideals recognized by a society. In this way, selves are the manifestation of both individual and cultural norms that are cultivated over the course of a person's biography, what many philosophers have referred to as "narratives". However, it would be misleading to say that selves are just features of a person's personality or identity that inform a narrative. This would miss out on several important functions that selves play. Below I describe how Ross conceives of selves as narrative constructions; I then compare this role with their evolutionary and economic roles.

Selves as narrative constructions

The term "multiple-self" is a convention familiar to both economists and philosophers. For many philosophers, the term indicates that persons contain multitudes, and that each human biography is rich with personal memories, beliefs and desires, convictions, aspirations, and expectations for the future. All these facets contribute to a

methodology of Robbins and Samuelson, i.e. why he thinks that Robbins's inclusion of introspection is integral to understanding economics as a deductive science.

common theme in philosophy which is that the self is a 'story', or rather, that selves are 'stories' which make up a person's identity. How Ross understands the narrative interpretation of selves is consistent with this theme; however, his contribution relies heavily on exploring how the confluence of personal experiences that make up each biography are intertwined with other biographies. This idea of co-authoring of personal stories is borrowed from Dennett (1991a).

The idea that selves embody real behavioral patterns stems from the assumption that humans are social animals and that our social embeddedness in groups leads to the construction of distinctive narratives that operate much like programs or plug-ins: they represent strategies to act in normatively acceptable ways by guiding behaviors according to the demands of a context or a convention. On this interpretation, selves emerge from a continuous process of enculturation. Ross describes it as follows:

> Selves… facilitate increasing predictive leverage over time by acquiring richer structure as the narratives that produce them identify their dispositions in wider ranges of situations. On this account, individuals are not born with selves; furthermore, to the extent that the consistency constraints on self-narratives come from social pressures, particular narrative trajectories are not endogenous to individuals. (2006, p. 203)

That persons are not born with selves speaks to the importance of enculturation in shaping a human biography. A human biography involves many dialogical modes of being, each of which corresponds to a number of narrative constraints; these narratives contribute to the experience of self that one identifies with. Of course, this experience is contingent upon how one manages and leverages their own narrative (an ideal of how they envision themselves) against the narratives that society imposes against them.

The fluidity and success of my interactions with others thus depends on the mutual, though often implicit, assumption that I will meet others' expectations as dictated by our construal of these shared narratives. Thus, I am the culmination of personal histories with family, friends, and colleagues, and of institutional and public codes of conduct, and I choose—as far as I can choose—how to maintain these personal histories. As Ross opines, "This philosophical account nicely captures the phenomenology and microstructure of selfhood. A personality is experienced to itself, and to others, as a relatively coherent story" (2005, p. 203). This illustrates how selves are narratively constructed and how their biographies are sculpted by everyday social interactions. However, we should keep in mind that this idea of a narrative is an instructive tool that enables academics to make sense of how individuals track their own stories—it's not as if people consciously construct narratives or think of themselves as characters whose actions must be coherent, otherwise they'll violate literary conventions of stability. What we should expect from a narrative approach to selves is a handle for

describing the constraints that are imposed on a human biography. In section 3 I explore how these constraints are realized, both formally and informally.

Selves as evolved mechanisms for social intelligence

Perhaps more important than the narrative construction of selves is the evolutionary role they play for Ross. In order to show that selves are behavioral stabilization devices, Ross argues, it must first be shown that humans, as social animals, were under selective pressure to be good at coordinating and that the biological basis of selves was to facilitate collective endeavors that promoted safety and survival of humans. Ross tells us that the fundamental kinds of games that social animals need to solve, "indeed, the class whose solution is almost constitutive of *sociality*" are coordination games (2005, p. 273). However, a little more needs to be said about the cognitive demands of "sociality", and about what conditions would need to be met in order for selves to emerge in the first place.

Distributed language and the cultural and institutional artifacts it generates are usually taken to be the distinctive marks of human intelligence (Ross, 2007b; see also Zawidzki, 2013). Yet, prior to the enculturation of H. sapiens, it was our perceptual acuity and the capacity to problem solve within ecologically constructed niches that set us apart from other hominids. Our predatory design enables us to process information of magnitudes that are staggering, and contemporary neuroscience reveals that much of what we intake is not consciously registered but is filtered for errors—for perceptual outliers that violate a predictive encoding of our immediate environment. This is the perceptual basis of what Andy Clark has called "biological reason" (1997, see also Clark, 1998, 2001) and it sets the evolutionary stage for improvements in cognition that extend beyond the brain and body. Biological reason can thus be seen as Mother Nature's response to information bottlenecks in the cognitive architecture of individual organisms that needed to communicate to solve joint ventures. Here, bottleneck refers to a physical limitation in computing power that occurs when the quantity of information a system receives exceeds the resources available for 'processing' it. The scare-quotes here are meant to indicate that cognitive processing for humans is not a straightforwardly physical matter as it is for von Neumann computer architectures.[6]

But increased social intelligence precipitates further challenges. New social arrangements enabled by signaling devices and proto-languages would encourage new possibilities to act, and this would have made it difficult to predict the behaviors of conspecifics without some reinforcing norms in place. This introduces a genuine

[6] For more on biological reason and situated agency, see Clark (1998, 2001, 2012); for more on hierarchical predictive encoding and its role in action-oriented perception, see Clark (2015)

possibility that information bottlenecks *would* inhibit collective action based on the sheer number of factors that would need to be considered before making an informed judgement. As such, the capacity for sociality would have generated its own need for adaptive engineering to enable humans to coordinate effectively. This conundrum is well-captured when Ross says that, "increases in nonparametric environmental complexity that arise with sociality put pressure on the power of straightforward economic agency" (2005, p. 277). If natural selection did favor socially intelligent individuals for the sake of computational efficiency it must have been because signaling systems provided an *external apparatus* to distribute the cognitive burden that social interaction would otherwise impose on individuals forced to compute solutions to coordination dilemmas on their own.

The problem that selves emerged to solve was not just the distribution of cognition via signaling systems, but the maintenance and preservation of strategies to protect individuals from exploitation once information became publicly available. Presumably, this is because as coordination drifted away from "purity"—i.e. away from situations of mutual advantage to situations with asymmetric benefits—signaling complexity would have given some individuals advantage over others. This drift and subsequent advancement in signaling phenomena forecasts what's known as the Machiavellian Intelligence hypothesis. The hypothesis suggests that social intelligence emerged as a result of competing pressures to find coordination solutions with partners without ceding strategic advantage to those partners (cf. Byrne and Whiten, 1988, 1997). The importance of the hypothesis for the current discussion is that it establishes the environmental conditions that would have prompted the emergence of selves for stabilizing unpredictable behaviors while also protecting against exploitative competition. In sum, the evolutionary role of selves is to provide a story about how the socio-cognitive capacities of early humans was directly correlated with their ability to signal and coordinate effectively.[7]

Selves as economic agents

If the evolutionary gloss above is approximately true, then we have a description of the conditions that prompted the emergence of selves. But the real challenge we are faced with is showing how the evolution of pre-enculturated biological individuals into socialized H. Sapiens is any indication of their economic function (which is not the same as, but is clearly connected to, the former two roles). In order to meet this challenge, Ross defends the ontological distinction between *pre-socialized biological individuals* (whose behavioral strategies are determined by Mother Nature) and

[7] For further discussion of the evolutionary benefits of language and meta-representational capacities for overcoming strategic exploitation vis-à-vis the Machiavellian Intelligence Hypothesis, see Sterelny (1998, 2007) and Zawidzki (2013).

enculturated H. sapiens (whose behavioral strategies are co-determined by social interaction in market systems). Once this distinction is drawn, selves are shown to emerge as "virtual" economic agents (2005, pp. 276-279—see also Ross, 2006).

As Ross stresses, it would get the ontological story backward if we started by assuming a well-ordered macroeconomy composed of encultured individuals competing for resources *and then* assumed that selves emerged merely "as a technology for improved competitiveness" (2005, p. 275). This would wrongly suppose that (1) individual persons are economics agents, and (2) that H. Sapiens entered the evolutionary scene with well-defined social interests. It's already been argued that the former conjunct is a non-starter; whereas the latter would conflict with the socio-cognitive timeline summarized above. The evaluative capacities needed to interpret and rank options as social interests could only arise once competitive and cooperative demands forced individuals to strategize to achieve their needs. This is why selves are a necessary condition for social interests to arise in the first place. Ross describes this social dynamic process as follows:

> If we so distinguish individual organisms without reference to any economic properties, we can subsequently subject them to economic analysis without introducing circular reasoning into our ontology. Then we want selves to emerge from the social dynamics that can arise when some of these biological individuals become enmeshed in complex… coordination games. (2005, p. 276)

This passage, in opposition to microeconomic individualism, reverses the story that selves emerged from pre-existing economic agents as assumed by Robinson-Crusoe metaphysics. It also avoids the circularity of defining selves as purely mathematical objects, which would be the case if Ross's account didn't go beyond the Samuelsonian framework.

It's important to note how the emergence story defended by Ross, which distinguishes pre-socialized biological individuals from enculturated selves, is taken to be an indication of selves' inherent economic role:

> If complex sociality is negatively correlated with straightforward economic agency, this should lead us to model some biological individuals, those that got enmeshed in complex coordination games with others, as evolving away from such agency. As they develop selves, they become different kinds of individuals, and the coextensivity between them and the biological individuals on which they are historically based breaks down. In the limit, the microeconomic approach with which we logically begin stops applying to them very effectively, and an evolutionary macroeconomics is called for. (2005, p. 277)

Because there is a lot of overlap with the previous roles that selves play, two points need to be unpacked here: The first point is that Ross wants to use the ontological distinction he draws between pre-socialized biological individuals and H. Sapiens

with evolved human selves to distinguish a *formal* economic role for selves. Although we could envision and model any strategic situation previously described in a game-form, only the strategic situations of enculturated selves can be modeled as 'games' in the sense of classical game theory. In this sense, the games that selves play include an information set containing all possible actions that the self can 'choose' from, each of which is designated by a utility function that corresponds to some socially-determined interest. As Ross says, "[i]dentifying a scenario as a game presupposes that players' strategy sets have already been constrained by determination of their specific utility functions" (2005, p. 278). But the whole point of denying microeconomic individualism and its Robinson Crusoe metaphysics is that H. Sapiens can't have well-formed preferences if they haven't yet developed selves. Pre-socialized biological individuals are merely passive recipients of the strategic situation types played by Mother Nature via competitive phylogenetic lineages. As proto-agents, biological individuals are *incapable* of strategizing (hence, they are ideally modeled with evolutionary game theory). On this first point, I don't disagree with Ross.

The second point, however, is that if we are going to model selves as players of games—and not as passive recipients of competitive phylogenetic lineages—then the games they play must be representative of the strategic environment within which they are embedded. This means that games cannot be depicted as *isolated* social interactions, but as consecutive nodes in an interconnected social network. Consequently, a move in one game may count as a simultaneous move in another game or series of games. Recall: this interconnectedness is the basis of the general equilibrium problem described in the evolutionary gloss above. That is, within a densely-connected social network selves function as virtual agents, as behavioral profiles that index information that relevant to the strategies that individuals are likely to play upon engaging one another in different contexts. These behavioral profiles not only reduce the cognitive load required to decipher the actions of others (because it is embedded in the context of the social interaction), but they serve to *reinforce* normative behaviors given their ability to trigger cues or provoke feelings of obligation, sanctioning, and what have you, which are indispensable for stabilizing behavior in a nonparametric choice environment:

> People probably do not literally solve problems, that is, actually find optimal solutions to their sets of simultaneous games (except, sometimes, by luck)… Nevertheless, most people achieve tolerable success as satisficers over the problem space. They do this at the cost of increasingly sacrificing flexibility in the new game situations. (Ross, 2005, p. 204)

This reiterates the economic importance of selves as devices for behavior stabilization since coordination is a solution to optimization problems in both human phylogeny and ontogeny, i.e. in evolutionary history and in contemporary social-psychological

development. While the context of optimization problems will differ as constraints and incentives change, the streamlining of sensible behavior according to norms reduces much of the burden that energy-costly strategizing would otherwise demand of an individual as he or she navigates the social world.

I conclude this section by raising an issue, viz. whether the different roles for selves that I've specified above entail different underlying projects. One the one hand, Ross seems to argue that the economic function of selves as behavior stabilization technologies follows necessarily from their evolutionary-biological function. But, on the other hand, the game-theoretic interpretation of selves would suggest that they are nothing more than strategy profiles of socially-embedded individuals; formally, they are equivalent to their utility functions. While it could be argued that these are merely two complementary roles for selves, one evolutionary-biological, the other methodological, I will show that, after we distinguish what each of these roles entails, it's harder to reconcile how these two roles could be complementary given that Ross's projects pull in different directions.

In the next section I describe how Ross operationalizes selves via a game-theoretic framework. This exegesis makes vivid the tensions underlying the roles I have identified above and points toward a tradeoff that, I will argue, is implicit in his broader philosophy of economics.

3. Social-determination, black boxes, and the externality of intentions

In the previous section I summarized how, according to Ross, the emergence of selves fostered the enculturation of humans; from this we could infer a general economic function for selves which is simultaneously cultural and biological. I now consider whether the formal interpretation of selves that Ross provides (per the economic role) is at odds with the other two functions. I will argue that although the formal interpretation of selves is logically consistent with the rejection of social atomism and microeconomic individualism, this interpretation crowds out the explanatory virtues that selves provide pertaining to their evolutionary-biological role.

In this sub-section I delve further into the nuts and bolts of Ross's formal framework, which he refers to as "game-determination". It is precisely because Ross thinks we *cannot* look to selves for psychological explanations of behavior that he must say something about how to get to such explanations, and he opts to say this: we must look to the situation types that orientate selves with respect to one another in a social network. These situations, when modeled as interconnected games, should tell us something about how individuals will act.

3.1 *Game-determination*

As a framework, game-determination builds upon the narrative-self hypothesis discussed in section 1: it defines the rules of games according to the institutional constraints and normative conventions that undergird a social network, and derives strategies *for* action *from* the narratives that selves are constructed from.[8] However, one obstacle that a game-theoretic account of the coordination of selves needs to overcome is *how* to depict the interconnectedness of social interactions.

It was stipulated above that situation types vary according to the type of player we are interested in modeling (e.g., pre-socialized biological individuals are ideal candidates for evolutionary game theory, whereas social humans with multiple selves requires a game theoretic models that permit diverse strategies). As such, asocial animals and pre-socialized H. sapiens are not the kind of agents that should be modeled with classical game theory. This is because their behavioral traits are determined exogenously by evolutionarily stable strategies rather than by preferences for situation-specific outcomes. As simple proto-agents, asocial animals are incapable of deviating from their natural function and thus exhibit stable behavior from birth to death. By contrast, the utility functions of encultured agents' change over time given that new games are continuously unfolding and the network that connects them grows more and more nebulous. As selves adjust to changing constraints, they can be ascribed new utility functions that are specific to the outcomes of each new situation. This is why players of games can be modeled as new agents each time their strategies change. However, we cannot assume that coordination is captured merely by iterated gameplay between players that are already familiar with one another; selves arise *in order to* reinforce their own narratives precisely by coordinating with new players and by learning which strategies are permitted and which are not. As selves gain new information and develop new methods of coordination, the strategies they play are calibrated and recalibrated. It is for this reason that Ross says, "we can't assume our initial individuation of agents to remain stable as we let socialization feed back into their economic agency profiles" (2005, p. 291). This process of continuous calibration

[8] In response to this claim, it could be argued that although selves are shown to play a narrative role, this does not entail that the features of real narratives need to be included in or formally represented by the *mathematical* strategies that selves stand for in Ross' game-theoretic framework. For instance, Ross states that, "the only *formal* properties needed for selves to play their strategic roles as constraints on sub-personal disorder and the membership of the available sets of G-level games are those properties associated with preference stability. Any entity with sufficiently stable preferences to ensure that stochastic dominance is respected… would do; *this needn't necessarily be a narrative self*" (my emphasis, personal correspondence). I would agree with Ross that the content of peoples' narratives need not be represented by the preference of selves in their games. But one should keep in mind that the issue I am raising here is not that the narrative and economic roles should be continuous, but that these two roles lead to further tensions in Ross' separate projects.

applies to all interacting agents giving rise to a deterministic web of peripherally unfolding games.

To capture the dynamics of a game-determined framework Ross classifies three types, or 'levels', of games that agents are engaged in. *G*-level games depict standard game-theoretic situations like prisoners' dilemmas, assurance games, and pure coordination situations—in these situations the players are modeled *as if* they know what kind of game they're playing, that is, they know what's at stake and have evaluated their alternatives accordingly. *G''*-level games depict evolutionary situations in which phylogenetic lineages compete to transmit genetic information—in these games asocial animals are passive recipients of natural strategies. *G*-level and *G''*-level games correspond to classical game theoretic situations and evolutionary game theoretic situations, respectively.

Ross introduces an additional strategic level, *G'*: this level depicts higher-order games that are played between agents that are already sculpted by cognitive, normative, and institutional pressures, but who are *uncertain* of what game (at the *G*-level) they may play with an opponent or conspecific. So, *G'* games codify the dispositions of players to interact prior to deciding how they each would likely play. Thus, Ross states: *"*G_i' is a game played by two strangers to each other who are already distinctive human selves. Its structure is of course determined by their preengagement utility functions"—these preengagement utility functions are informed by the background and concurrent games that the agent has already played (2005, p. 292). He continues:

> By reference to this game we can state the narrative theory of social self-construction as follows: many engagements involve incremental refinements of the selves of the (nonstraightforward) agents who play G_i' so that they become new agents who, still in S_i, will play G_i. (p. 292)

However, the dispositions that preengagement utility functions represent do not *strictly determine* the outcomes of *G'* games. They merely establish the background conditions (as narrative constraints) that inform selves how they *ought to* approach a strategic situation. What this game-theoretic model provides is a formal platform to depict the sending and receiving of signals to coordinate; this affords modelers the opportunity to visualize or at least theorize about how players evaluate their bargaining position by deciphering subtle physical and rhetorical signs to determine what kind of game shall be played.

A question that arises then is: at what level of strategic interaction do selves emerge such that we can specify them as distinct behavioral profiles? —should selves be identified only with the behavioral outcome that is observable as a move in a game at the G-level? The revealed-preference interpretation of Ross's (neo)Samuelsonian framework would suggest something like this, prima facie. But this can't be right for it would render the concept of selves redundant—they would effectively be no

different than revealed preferences. This leads me to suspect that selves emerge as stable agents at the G'-level, which is where signals are sent, received, and deciphered. Let's flesh this out:

The reason Ross thinks that selves serve a strategic role (within the economic role) is because selves correspond to different behavioral profiles *prior to* engagement at the G-level. If we grant this, then there is some intuitive reason for thinking that pre-engagement utility functions at the G'-level do correspond to dispositions to act, even if those dispositions are not realized (say because an agent perceives deceit and changes its strategy). But this means that strategies depicted as moves in G' games aren't easy to define precisely because their outcomes are what we observe as G-level coordination. One could entertain many possible alternatives for making sense of what actually happens in games at the G'-level: one alternative could be that a prime self (which is determined by its pre-engagement utility functions at the G'-level) chooses among profiles which it 'decides' to deploy in the G-level game. Another alternative is that selves at the G'-level are not yet determined and have to bargain at the G'-level simultaneously as they compete for a position in the G-level game. It's not clear which interpretations we should take. In a later article on the evolutionary basis of selves, Ross clarifies that "…if the subject's own participation in self-narration is a strategic response aimed at coordination with others, then an economic model must interpret selves as products of games played among sets of players that can't include that very self" (2006, p. 205). This leads me to believe that unlike either of the alternatives I propose, Ross envisions *new* selves emerging out of the games played at the meta-strategic level.

The point one should consider here is that if Ross's game-determination framework is to be interpreted as a model of market systems where information about how individuals behave is exogenous to the games that selves play, then we in fact learn little about what selves *do*. Do they merely represent possibilities to act, or do they partake in the selection determination of an appropriate strategy given some incentives and constraints?

It could be argued at this point that there is tension between the different roles Ross envisions for selves, primarily because there is some ambiguity about what takes place at the G'-level. On the one hand, it seems that Ross's illustration of the constraining effects of social networks seems to suppose that selves are, in fact, *not centers of decision-making*, since all strategies in a game-determined framework are externally imposed, at one 'level' or another. This is supposed to demonstrate that selves have only a *virtual* presence. But, on the other hand, this seems to conflict with the evolutionary-biological lessons Ross's also wants to teach us, which are that selves enable pre-socialized biological individuals to become *intentional beings* (i.e. to take the intentional stance toward themselves and others) by sending and receiving signals

via a public language. On this reading, the emergence of selves *is* the emergence of intentional action.[9]

3.2 *Selves as black boxes*

One problem that stands in the way of further analysis is the apparent duality of the projects that selves (in their different roles) are supposed to serve. Ross tells us that selves are ontologically equivalent to persons (2005, p. 318). This is because selves are narrated into being by social interactions, public conventions, and other historically significant episodes. Yet, the above exegesis of the economic (mathematical) role of selves suggests that they are really abstract entities—we don't *see* selves, we infer their presence (or existence) by reflecting on the motivations behind our ordinary behavioral patterns. Game-determination views selves as the culmination of selection pressures and learning opportunities to generate strategy profiles. Strategy profiles are represented by selves' preferences. This includes preferences over the outcomes of single games (G-level games) as well as preferences (meta-strategies) over the outcomes of higher-order games (G'-level games) which influence downward the type of games individual selves will play. What these considerations amount to is the self being treated as a purely mathematical object: selves *just are* whatever enables an entity to maximize a utility function, and so, they are necessarily tautological. There is no method by which to individuate selves *prior to* an individual's engagement (or pre-engagement) with another where coordination demands the taking of a decision. (Recall, this is part of the economic role of selves).

This dilemma should provoke curiosity from readers. The idea that selves cannot formally be individuated (beyond the strategies they represent in G-level games) should raise questions about Ross's overall projects. Recall that Ross has two main projects: (1) is to provide a story about how the emergence of socio-cognitive functions enabled humans to engage in market behaviors via fluid coordination; while (2) is to provide theoretical foundations for an account of economic agency that is not individualistic but still amenable to neoclassical economics. Now, as stated previously, we need not see the three distinct roles for selves as contradictory or in

[9] In fact, it would seem that selves can *only* emerge in the presence of a public language (or public signaling system). Ross continuously extols human language as the primary technology for social learning, and hence, as the primary tool by which selves hold other selves accountable for their actions. After all, public language is what enables selves to first take the intentional stance toward themselves, which is Dennett's primary weapon against Cartesian accounts of cognitive processing (1991b). For Ross, language is the dominant medium by which selves convey information to one another about how they will coordinate, and by extension, how they solve the general equilibrium problem of consistently strategizing with all other selves in a computationally nightmarish social network.

competition with one another. But, if these distinct roles indicate contradictory outcomes for Ross's two projects, then there is need to consider further what each project requires or is committed to.

Before I take this discussion further, I want to consider a possible challenge to my line of inquiry. One could argue that it is, in fact, categorically mistaken to ask what "takes place" at the G'-level of a signaling game because such a question presupposes that what players do at the G'-level is psychological or computational in nature, which is not the case. The formal interpretation of selves explicitly prohibits reading any psychological or computational properties into their behavior because their strategies are already fixed by their situation types. This is why selves—that is, the narrative constraints that distinguish selves—are left as black boxes.[10] In Ross's neoclassical framework, individual actions are produced by virtual economic agents, and selves are narrated to signal those actions for the purpose of making their behavior intelligible (to others as well as to themselves). In response to this disclaimer, I would like to clarify that I do not presume that Ross needs a psychological foundation to account for the behavior of selves if by "psychological" we mean an account that traces decisions back to propositional-attitudes *inside* individuals' heads. But, if we permit that psychological ascriptions of attitudes are just conventions of language which allow individuals to interact and make sense of each other's behaviors (which Dennett certainly does), then this does not count as psychological in the sense that Ross tends to mean it, i.e. as a study of the causal mechanics of individual choice.

This disclaimer about the psychological foundations of selves is important because it illustrates the differences I am trying to draw between the formal interpretation of selves and the real-world economic function of selves which is built upon their biological and narrative roles. The formal interpretation leaves selves as black boxes because they are whatever maximizes an agent's payoffs in a game—this appears to be a *logical* consequence of the definition of selves-qua agents—in a game-determination framework. However, that Ross clearly wishes to externalize agent intentionality via the distributing effects of language and thereby account for the cultural-evolutionary dynamics of signaling phenomena, the black-boxing of selves appears to be a *methodological* consequence of network complexity.[11]

---

[10] In the original passage, Ross advises "let us for now just understand a narrative constraint in the vague operational sense of whatever it is that leads a given group of people to judge some behavioral sequences as ones in which earlier behavioral patterns explain others, and other sequences as ones in which explanation must draw on synchronic factors exogenous to behavioral patterns alone. (2005, p. 286)

[11] The innovation of a G'-level for strategic reasoning is designed to combat a flaw in Frank's (1988) theory of emotional signaling. Although Ross agrees with Frank that emotions are integral for non-conventional (i.e. non-linguistic) signaling, he argues that Frank overestimates their socio-cognitive importance in the broader process of strategic coordination—that is,

Thus far, we've been introduced to a picture of selves that is intuitively plural: selves can only be sensibly understood in the context of other selves. By abandoning all vestiges of Cartesian epistemology, anti-individualism makes it implausible to conceive of a self independently of the structures that give its actions meaning and directedness. While this picture is not intrinsically problematic, it does introduce restrictions on what philosophers of economics can expect to learn about selves. This is why it seems imperative that we focus on the methodological restrictions, since this tacitly permits further study of the dynamics that lead to the regulation of individual behavior via selves. In the remainder of this section I will highlight a few places that we can read Ross as endorsing the view that selves are more like the biological and narrative roles I described in section 1.2. This should signal to readers that the formal interpretation of selves is mostly a dead end if we hope to learn anything about real-life social dynamics.

3.3 *Externalizing intentionality—or, what coordination entails for individuals with selves*

As I discussed above, there is much room for possible misinterpretation about what coordination at the G'-level entails since it is not a visible interaction. For instance, it may seem as if G' games are "binding preplay" for the negotiation of the G-level game. On this reading, social coordination is a cooperative effort since both players seek to match their respective expectations at the G'-level which commits both of them to a mutually optimal G-level game. However, to show why this is the case we need to consider a cluster of related issues:

The first part of the cluster pertains to the reasons why Ross does not to assent to the presumption that higher-order coordination is necessarily cooperative. One reason is straightforwardly strategic: Human selves in the real world may have good reason not to cooperate at the G'-level if they suspect that the resulting G-level game yields vulnerabilities or uncertainties they wish to avoid. This noncooperative thesis follows naturally from the theory of narrative construction and constraint described above since the molding of selves is shown to be an incremental process. It is due to the underlying dynamics of narrative construction that players can't "simply assume self-predictability; [rather] they have to act so as to make themselves predictable" (2005, p. 293).

Another reason why higher-order coordination isn't necessarily cooperative is that it would be implausible for a self to cooperate with all other selves simultaneously. This stems from the inherent complexity and interconnectedness of the social networks that scaffold human biographies. Recall that a move in one game is

---

Frank's account fails to incorporate culturally evolved conventional signals that mediate between G"-level and G-level games (2005, pp. 297–316; see also Ross & Dumouchel, 2004).

simultaneously a move in another game (or series of peripheral games); even if a player intended to negotiate at the G'-level in an attempt to show commitment toward playing a particular G-game that is optimal for both players, this may be interpreted as a display of noncooperation in another G'-level game by a third party, leading to competitive play in a subsequent peripheral G-level game with that third party:

> A person can't keep the various games she simultaneously plays with different people in encapsulated silos, so a move in a game Gi' with the stranger will also represent a move in other games Gk, . . . , n with more familiar partners—because these partners are watching, and will draw information relevant to Gk, . . . , n from what she does in Gi'… Both of these points can be expressed by saying that nature doesn't hand people cards telling them which games they're in when. Games have to be determined dynamically—and determination processes are themselves games. (Ross, 2005, p. 293)

Higher-order coordination compounds the complexity of the general equilibrium problem that selves emerged to solve—i.e. the systems of pressures that underwrite the dynamics of broad social coordination are "computationally intractable" from the perspective of a serial processor. This is why the concept of narrative constraint is integral for Ross's concept of game-determination: the concept of selves is not just useful for understanding how individuals stabilize their behaviors, but also for minimizing (or streamlining) the number of strategies that an individual has to be prepared to deploy. Recall that this is exactly the evolutionary challenge that self-emergence introduced. In response to this challenge, it was argued that people achieve tolerable success as *satisficers* over the general equilibrium problem space. They do this at the cost of increasingly sacrificing flexibility in new game situations. Thus, the general success of coordinating—satisficing rather than maximizing—follows from the tendency of individuals to avoid the kind of destructive games that would require energy-costly computation of the kind likely to cause coordination errors:

> This general fact itself helps to explain the prevailing stability of selves in a feedback relationship. It is sensible for people to avoid attempts at coordination with highly unstable selves. Given the massive interdependency among people, this incentivizes everyone to regulate the stability of those around them through dispensation of social rewards and punishments. As described earlier, this is how and why we get selves, as stabilizing devices, in the first place. (Ross, 2005, p. 294)

In order for selves to develop, that is, in order for the process of enculturation to take place and for coordination problems to be solved by persons, we must presuppose the development of robust cognitive and linguistic tools. At the same time, cognitive and linguistic tools cannot evolve further without stabilizing devices, i.e. selves, to direct and orient their use *as* media for communication.

Moreover, Ross continuously extolls the importance of language as the primary technology for scaffolded learning, and hence, as the primary tool by which selves hold other selves accountable for their actions (2004, 2007b). After all, public language is what enables selves to first take the intentional stance toward themselves; as well, it is the dominant medium by which selves convey information to one another about how they will coordinate, and by extension, how they solve the general equilibrium problem of consistently strategizing with all other selves in a highly complex social network. Public language isn't just some vehicle of information transmission that happens to be useful—it is, from an evolutionary and development perspective, the socio-cognitive tool that allows pre-socialized H. Sapiens to become selves. Ross states that: "For Dennett, narrative structure essentially requires language. This derives not from the implicit analysis of narrative itself… but from the [multiple drafts model of consciousness] …" (2005, p. 286). Moreover, language provides a structure that is "ontologically prior to and wider than" the particular pressures that constrain a narrative self. In this way, public language—understood as a relatively fixed system of information transfer—provides the right kind of external scaffold for judgments to be made (1) by selves about their collective personality, and (2) by other selves for the purpose of policing norms.

4. Social selves versus sub-personal selves

Let's take stock of the discussion thus far. Aside from the primary concern that there are multiple roles for selves, another point of contention concerns what makes up a self – or rather, what gives different selves their identities? I argued that, for Ross, selves are triangulations of brain activity, social interaction, and normative constraints. This idea coincides with the idea that selves *do more* than serve a formal role for game-theoretic representation in a neoclassical framework. It suggests that the narratives that persons rely on to guide their behaviors are stable despite the recalcitrant complexity of their sources. The recurring problem for Ross is that some (most) social facts do not remain constant, and so there is a deep theoretical need to ground our understanding of selves in something firm, something measurable.

4.1 *Against the view that selves are sub-personal*

In trying to get a handle on what anchors selves' identities, philosophers of economics have interpreted Ross' account in one of two polarizing ways, conceiving selves as either *sub*-personal or *supra*-personal entities. As a foil for this discussion, I appeal to Davis's (2011) analysis of selves which interprets them as *sub-personal neural agents*. I contrast my own position against Davis's and show that the differences in

how we interpret Ross illustrate different ways of envisioning future research on the topic. To jump right into it, Davis characterizes Ross as follows:

> Ross' neuroeconomics-based view… treats these different neural systems as relatively independent neural systems and thus as a person's multiple selves. As such, they are sub-personal multiple selves rather than supra-personal ones, and he accordingly investigates what a person is from the perspective of neuroscience rather than from social psychology. (Davis, 2011, p. 125)

As a rough-and-ready description of what selves are, I find this description misses the mark. However, because Davis does provide an otherwise remarkable analysis of Ross' agenda, we should look more carefully at how he understands "sub-personal multiple selves" for it brings additional clarity to Ross' three conceptions of selves.

Davis provides a very clear and concise account of the evolutionary pressures that, for Ross, would drive neural agents – behaving as a semi-cohesive unit – to seek out partnerships with other clusters of neural agents: he states that

> Because evolution has confined sets of sub-personal neural agents to the same individual human bodies, it turns out to be symbiotically in their interest to cooperate with one another in order that the body they jointly inhabit survives. Further, as whole individuals' survival also depends on interaction with other whole individuals (who are similarly the result of internal coordination games). (Davis, 2011, p. 128)

To be fair, the cultural-evolutionary gloss that Davis proceeds to give is a faithful depiction of Ross' account of selves as a technology for behavioral stabilization, so it accords with my analysis above: selves facilitate intrapersonal and interpersonal action through which individuals, conceived as coalitions of neural agents, sculpt and re-sculpt themselves. Where I disagree with Davis is in his presuming that these neural agents constitute selves, and so, are intrinsically sub-personal. This may seem like merely a technicality, a quibbling over proper use of jargon, but, I think a few points are worth fleshing out which will distinguish my contribution as a constructive criticism of Ross.

Davis's intended question "whether a single individual should play any role in a neuro-cellular economics" (2011, pp. 125-132) does not clearly represent either of Ross' projects. How Davis proceeds to answer this question—which envisions Ross trying to unify of two domains of economic inquiry (i.e. the behavior of neurons and the behavior of individuals)—ignores many of the subtleties I've tried to flesh out in this paper. To be clear, Ross does argue that the internal games that neural agents play have an outward effect on the organism as a whole; but, he is adamant that we not conflate this level of activity with the activities that selves engage in, viz. conventional signal-sending. Recall that the goal of introducing cultural dynamics into a game-

determined framework (codified as G'-level strategic play) was to disembark from the phylogenetically determined games of pre-socialized biological individuals.

Perhaps what I've argued thus far is not a radical departure from Davis's interpretation because I essentially agree with him that there is an ambiguity in Ross' argumentation—to quote him again:

> It's one thing to say that individuals have a capacity to reflexively produce self-narratives or discursive representations of themselves, and it is another thing to say that these representations are specifically whole individual representations of themselves: *self*-reports rather than simply representations of different aspects of themselves… Nothing in Ross' analysis of interaction between individual's sub-persona selves quite tells us how they collectively graduate to producing whole individual self-reports. (Davis, 2011, p. 129)

This seems to get to the heart of the problem I raised in section 2, viz. that it isn't clear how self-signaling works at the meta-strategic G'-level prior to selves settling on a course of action. My concern with Davis's interpretation of Ross is that it misrepresents the inherent tension and trade-off that one is confronted with if selves are conceived as black boxes (let alone three of them).

For instance, Davis inquires whether (for Ross) individuals' representations *of themselves* might be alternatively of one neural system, and then another neural system, and so forth, thus indicating that the identity of the whole person is a constant flux of selves (2011, p. 129). This is meant to suggest that Ross' account is problematic because each 'self' is bound up in some set of narratives that *all* depend on equally unstable self-narratives. But Davis's inquiry misrepresents the relationship between selves and persons—it suggests that the potential instability of selves is the result of causal relationship between underlying neural activity and the content of consciousness. To understand why this is wrong-headed, recall the lesson of Dennett's multiple drafts model of consciousness, which was intended to alleviate the temptation to think of mental content (perceptions, judgments) as occupying discrete regions in the brain (1991b). If we were to probe an individual's brain during a perception, we would not find a locus of experience that represents that perception. The experience itself is a stream of multi-track processes that are distributed throughout the brain. Likewise, if we could probe individuals to solicit information about their selves, we would not find collections of discrete stories that correspond to memories and other biographical mental content. The reason selves are black boxes (on any interpretation) is because they afford possibilities to act. Selves do not represent neural information; they represent solutions to coordination dilemmas that are the result of a continuous and predictive updating of their ecological niche. Demanding what makes up a self is like demanding where a propositional attitude is located in the brain. We may distinguish patterns of neural connectivity and on that basis draw correlations with a person's

outward behavior (including their verbal reports of conscious experience). But this does not give way, by analogy, to an account of selves that is sub-personal.

Although Ross sufficiently distinguishes his position from Glimcher-style neuroeconomics,[12] a close examination of his (2006) and (2007b) articles on the evolutionary and ecological basis of human social intelligence further supports a view of selves that is intrinsically rooted in social dynamics, not in sub-personal neural activity:

> …human personalities—selves, that is—have been made phylogenetically possible and normatively central through the environmental manipulations achieved collectively by humans over their history, while *particular* people are ontogenetically created by cultural dynamics unfolding in this context… individual people are themselves systems governed by distributed-control dynamics… and so must for various explanatory and predictive purposes be modeled as bargaining communities. These theses together imply that adequate models of people—and not just of groups of people—will be social-dynamic models through and through. (Ross, 2006, p.200)

What Ross *does say* about the economic study of neural activity does not endorse Davis' reading of selves as rooted in a "neuroeconomics-based" approach. I quote Ross at length:

> Taking account of the way in which people are distinct from their brains in the point of my suggested appeal to neuroscientific control theory… This precisely implies the distinction between brain-level individualism and person-level individualism, especially if one of the advantages people bring to the table by contrast with brains is faster response to the flexibility encoded in social learning. Brains bring compensating advantages of their own, as we should expect. As the discussion of asset valuation above suggests, their reduced plasticity relative to socially anchored selves can help maintain objectivity in circumstances where herd effects occur. *It is just when we don't conflate maximization of utility by brains with goal achievement by selves that we have some hope of using data about the former as a source of theoretically independent constraints on processing models of the latter.* (2006, pp. 207-208 emphasis added)

Now, what Ross means by "using data about the former [utility maximization by brains] as a source of theoretically independent constraints on processing models of

[12] Ross provides a rich analysis of the differences between Glimcher's (2004) neuroeconomics approach (Ross, 2008) and Ainslie's picoeconomics (Ross, 2005, pp. 322-334, 337-353). The lesson to be drawn from this analysis is that his interpretation of selves more closely aligns with Ainslie's account of sub-personal interests, which are not neural agents.

the latter [goal achievement by selves]" is not entirely clear.[13] But, what is clear is that it does not justify Davis' claims that selves are neural agents.

4.2 *Neuroscientific control theory and participatory sense-making*

I argued above that Davis' inquiry whether self-representations alternatively pick out different neural agents misrepresents the relationship between selves and the activities of the brain. Nowhere does Ross (2005, 2006, 2007a, 2007b) refer to his own approach to multiple-selves as "neuroeconomics-based". Moreover, he repeatedly cautions neuroeconomists to keep personal-level information distinct from sub-personal-level content for it otherwise "encourages a slide back into an individualist conception in which people are taken to be mereologically composed out of functional modules that locally supervene on neuronal groups" (Ross, 2006, p. 207). Now, one may ask: if I ultimately agree with Davis that there is an ambiguity in Ross, why does it matter how we differentiate our understanding of selves? Why go to the trouble of arguing that they are not neural if we can't, in the first place, determine what they are?

In clarifying Ross' account of selves we are forced to confront the fact that individual behavior is inextricably tied up in dynamics above and below the personal-level. To this end, however, it is integral to understanding these dynamics that we distinguish the study of *biological individuals*, who are coalitions of neural agents forged from biological evolution, from *persons*, who are products of (some form of) cultural evolution. Even if we cannot identify or agree upon a stable vehicle for the study of selves, a philosophically conservative analysis nonetheless informs us of what possible roles they can (and can't) play, both within economics and in other disciplines. In bringing this paper to a close, I thus consider two ways we can proceed given that selves are left as black boxes. One move involves reading Ross' account as a cautionary tale; the other involves a direct application of the black-box concept.

First, with regard to the study of intrapersonal and intertemporal choice, behavioral economics offers a dizzying array of options for modeling sub-personal selves. One family of models which has gained considerable popularity takes a "dualistic" approach toward the individual, wherein the decision-process is modeled as a game between a long-run "planner" self and short-run "doer" self (this is based on the principle-agent design made famous by Thaler & Shefrin, 1981). Following this format, there have been no shortage of attempts by researchers to map these selves onto

[13] He does say that, "Attention to AI and neuroscience forces us to take seriously *some* limits on the sensitivity of behavior and agency to all the dynamical forces present in an environment. Complex systems can only manifest agency if they achieve stable integration of information in such a way as to shield them, up to a point, from dynamical perturbations" (2006, p. 205). Elsewhere (Ross, 2007a, 2009) he does consider ways of reconciling what he calls "molar" and "molecular" approaches to economic agency. This involves a multi-scale approach to agency that brings neuroeconomics into the picture, but leaves it as an ontologically separate endeavor.

underlying processes in the brain, viz. "controlled" processes and "automatic" processes (cf. Benabou & Tirole, 2002; Benhabib & Bisin, 2005; Loewenstein & O'Donoghue, 2005). The models dictate that the outcome of an agent's choice, when conceived as trade-off between temporally distinct selves, represents endogenous motivations that are *causally determined* by the activation of cognitive systems where these processes take place.[14] Another family of dual-self models takes this idea a step further, attempting to directly observe how the brain optimizes rewards given "budget constraints" over its energy resources. For instance, research conducted by McClure et al. (2004), McClure et al. (2007), and Brocas & Carrillo (2008a, 2008b) indicates that decisions are, in fact, processed in domain-specific systems in the brain, and on this basis, they believe they can isolate the determinants of myopic behaviors.[15]

While there are many reasons to be wary of how both families of models conceive of sub-personal selves, it's possible that the second family of models, which are more explicit about their domain of investigation, could benefit from what Ross refers to as *neuroscientific control theory* (2006, p. 207). Control theory tells us what we can expect to learn about selves *if* we define them as a separate kind of neural agent, which Ross refuses to do. In performing valuations different from intentional selves, brains are accountable for the type and integrity of the information available *to* persons. While control theory does not tell us how to encode information at the level of social learning, it constrains the strategies that intentional selves, as economic agents, can develop insofar as their own signals are translated through a medium that the brain was designed to manage. It is for this reason that Ross's envisions a fruitful partnership between evolutionary game theory and neuroeconomics, with the former providing the methodological scaffolding for social dynamics and the latter defining the (neural) capacities of its agents.

Second, growing interest in the study of distributed cognitive systems has brought philosophers, cognitive scientists, and linguists into close proximity. For instance, embodied and enactive approaches to cognition have speculated about how a community of language-users might achieve social coordination and develop behavioral-linguistic conventions without assenting to an over-arching theory of mental representation (which would require linguists and cognitive scientists to figure out how people

[14] Though, it is a matter worthy of debate *how* behavioral economists envision and model the activation of cognitive processes, and *how* this relates to different categories of decision-making at the individual level. There has been no systematic attempt to understand how dualistic models of this kind conceive of selves with regard to different levels of reward conflict. Put another way, many behavioral economic approaches to dual-self modeling conflate conflict observed at the neural level with experienced conflict at the personal level.

[15] However, alternative research by Glimcher et al. (2007) and Kable & Glimcher (2007) suggests that reward and information systems aren't as discrete as they may appear, and that the decision-making process is distributed throughout the entire connectome, implying a more unitary picture of the brain (cf. Rustichini, 2008; Vromen, 2011).

"read" each other's minds). Accounts such as Hutto (2008), Hutto & Myin (2013), McGeer (2007, 2015), and Zawidzki (2013) suggest that individuals do not read minds, but rather "make" them or "shape" them through commissive speech acts. These speech acts build narratives, reinforce social norms, and enable individuals to *become* intentional beings within a community. The problem with such accounts is they are highly theoretical, they lack a means to quantify the act of sense-making in a community. For instance, De Jaeger & Di Paolo (2007) venture an enactive model of social cognition, by which they represent the process of participatory sense-making as a dyadic interaction between two individuals. While their model is instructive, its abstractness undermines the process of enculturation that we see Ross so carefully trying to construct in his own framework. An account of selves that is black-boxed fits in here because the object of study for enactive social-cognition is not the individual person, but the dyadic relation between social selves. Cast in terms of conditional games (cf. Sterling, 2012) the strategic interactions that lead to intersubjective agreement are the kind of social relationships that Ross's account is poised to explore.

5. Concluding remarks

The motivation for writing this paper was to critically evaluate the concept of human selves, and to locate ambiguity or inconsistency that results from conflicting roles played by selves in Ross's framework. In essence, my argument was that there is a discrepancy between the biographical interpretation of selves and the formal interpretation of selves. The biographical interpretation suggests that selves are a product of social and neural activity—it was for this reason that Ross views selves as "ontologically equivalent to whole people" (2005, p. 318). Under this interpretation I distinguished three distinct roles and fleshed out details of each. By contrast, the formal interpretation of selves was shown to enable modelers to individuate strategies played by selves without needing to individuate selves *per se*. On this reading selves just are the preference profiles of distinct economic agents. While it's entirely possible that the biographical details could serve as inputs for strategies, it's not clear how this can be done. This is because Ross is notoriously critical of behavioral economic programs that seek to isolate and codify dispositions and/or psychological mechanisms that underwrite individual choice-behavior. Most readers familiar with Ross's framework should have a general understanding of these various roles even if they have not thought through the implications themselves.

However, the real issue with which I am concerned, which I've attempted to clarify in this paper, is that selves are not designed for a practical need but a theoretical one, which is to construct (1) an evolutionary story about the cognitive functions of humans, and (2) to show how the concept of economic agency can be salvaged given that humans are not ideal agents. I think this is the reason for ambivalence about their

interpretation which I've described as a separate role: they are mathematical entities insofar as they are individuated according to their utility functions, which is their economic role; and they are behavioral stabilization devices which developed as humans learned to distribute the cognitive burden of resolving coordination, which is their evolutionary role. And spanning both these roles, selves are also biographical entities insofar as they enable people to manage different personas and identities as they participate in market contexts. The problem we are thus faced with is not reconciling these separate roles, but in finding a way to realize Ross' projects which, which seem to demand properties of all these roles simultaneously.

6. Bibliography

Ainslie, G. (2001). *Breakdown of Will*. Cambridge: Cambridge University Press.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3), 871-915.

Benhabib, J., & Bisin, A. (2005). Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions. *Games and Economic Behavior*, 52(2), 460-492.

Brocas, I., & Carrillo, J.D. (2008). The brain as a hierarchical organization. *The American Economic Review*, 98(4), 1312-1346.

Brocas, I., & Carrillo, J.D. (2014). Dual-process theories of decision-making: A selective survey. *Journal of Economic Psychology*, 41, 45-54

Byrne, R., & Whiten, A. (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford: Oxford University Press.

Clark, A. (1997). Economic Reason: The Interplay of Individual Learning and External Structure. *Working paper*. Department of Philosophy, Washington University in St. Louis.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT Press.

Clark, A. (2001). Natural-born cyborgs?. In *Cognitive technology: Instruments of mind* (pp. 17-24). Springer, Berlin, Heidelberg.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181-204.

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Davis, J.B. (2003). *The Theory of the Individual in Economics: Identity and Value*. Routledge

Davis, J.B. (2011). *Individuals and identity in economics*. Cambridge University Press.

Dennett, D.C. (1987). *The Intentional Stance*. Cambridge (MA): MIT Press.

Dennett, D.C. (1991a). Real patterns. *The Journal of Philosophy*, 88(1), 27-51.

Dennett, D.C. (1991b). *Consciousness Explained*. New York: Little Brown & Co.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences* 6(4), 485-507.

Glimcher, P.W. (2004). *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. Cambridge MA: MIT Press.

Glimcher, P.W., Kable, J., & Louie. K. (2007). Neuroeconomic studies of impulsivity: now or just as soon as possible? *American Economic Review* 97(2), 142–147

Hands, D. W. (2008). Introspection, revealed preference, and neoclassical economics: a critical response to Don Ross on the Robbins-Samuelson argument pattern. *Journal of the History of Economic Thought*, 30(4), 453-478.

Hutto, D.D. (2008). *Folk psychological narratives. The Sociocultural Basis of Understanding Reasons*. Cambridge, Mass.

Hutto, D. D., & Myin, E. (2012). *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press.

Jamison, J., & Wegener, J. (2010). Multiple selves in intertemporal choice. *Journal of Economic Psychology*, *31*(5), 832-839.

Kable, J.W., & Glimcher. P.W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12), 1625-1633.

Kincaid, H., & Ross, D. (2009). *Handbook of the Philosophy of Economics*. Oxford: Oxford University Press

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics, 112*(2), 443-478.

Loewenstein, G., & O'Donoghue T. (2005). Animal Spirits: Affective and Deliberative Processes in Economic Behavior. *CMU Working Paper*.

McClure, S.M., Laibson, D.I., Loewenstein, G., & Cohen J.D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695), 503-507.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience*, 27(21), 5796-5804.

McGeer, V. (2007). The regulative dimension of folk psychology. In Hutto, D.D., & Ratcliffe, M. (Eds.), *Folk Psychology Re-assessed* (pp. 137-156). Springer, Dordrecht.

McGeer, V. (2015). Mind-making practices: the social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18(2), 259-281.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review*, 89(1), 103-124.

O'Donoghue, T., & Rabin, M. (2001). Choice and procrastination. *The Quarterly Journal of Economics*, 116(1), 121-160.

Pettit, P. (1995). The Virtual Reality of "Homo Economicus". *The Monist*, 78(3), 308-329.

Ross, D. (2004). Meta-linguistic signaling for coordination amongst social agents. *Language Sciences* 26, 621–642.

Ross, D. 2005. *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Ross, D. (2006). The economic and evolutionary basis of selves. *Cognitive Systems Research*, 7(2-3), 246-258.

Ross, D. (2007a). The economics of the sub-personal. *Economics and the Mind*. Routledge.

Ross, D. (2007b). H. sapiens as ecologically special: what does language contribute?. *Language sciences*, 29(5), 710-731.

Ross, D. (2008). Two styles of neuroeconomics. *Economics & Philosophy*, 24(3), 473-483.

Ross, D., & Dumouchel, P. (2004). Emotions as strategic signals. *Rationality and Society*, 16(3), 251 286

Ross, D., & Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences*, 27(05), 603-627.

Sterelny, K. (1998). Intentional agency and the metarepresentation hypothesis. *Mind & Language*, 13(1), 11-28.

Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480), 719-730.

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, 89(2), 392-406.

Zawidzki, T.W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.

Chapter 4

From selves to systems: On the intrapersonal and intraneural dynamics of decision making[1]

1. Introduction

The idea of a 'divided self' has been the source of folk-wisdom for centuries. However, new research into the cognitive and behavioral foundations of decision-making suggests that this idea is more than just a metaphor. Our minds—and brains—appear to be divided in interesting if unexpected ways. In support of this, evidence suggests that many basic decision errors stem from inner processes over which individuals have no direct control. From the perspective of behavioral decision research, this is monumental. Studying the origins of decision errors, may help researchers to understand more complex 'failures' of rationality, like weakness of will, procrastination, and even addiction. Yet, there are still many uncertainties about the divided mind-brain and its relation to basic decisions errors and self-control problems.

The literature projects two methods for understanding decision errors and self-control problems. One method interprets individual behavior as a dynamic process. Multiple-self models conceive decisions as the outcome of a strategic exchange (a game) between 'selves'. Selves are just a formal representation of a person's competing interests. Multiple-self modeling is a common practice used by economists and decision theorists who wish to understand the conditions that lead to self-control problems. Yet, another method studies how brains process information. Dual-process and dual-system theories provide a multi-purpose framework which differentiates 'higher' and 'lower' cognitive processing. It is believed that some 'systems' are fast and automatic, and therefore error-prone, while others are slow and deliberative. The study of information processing is common among social and cognitive psychologists who wish to explain the causes of decision errors.

Until recently, these two approaches were kept relatively separate from one another. Yet, new trends reveal interesting collaborations between economists and psychologists. Now researchers are investigating how the insights of dual-process and dual-system theories might be used to inform multiple-self economic models. To this end, researchers have tried to integrate various features of these two approaches, and have, in turn, produced a wide array of psychologically sophisticated multi-agent

[1] This chapter is forthcoming in the *Journal of Economic Methodology*, in a special issue on "Interdisciplinary Perspectives on Behavioral Economics". I am grateful to Magdalena Małecka and Michiru Nagatsu of the TINT Center for Excellence in Helsinki for organizing the workshop from which this paper developed, May 22-23, 2017.

models. While different examples of multi-agent models (henceforth, "MaMs") can be found throughout the behavioral sciences, MaMs in particular have two core features: (1) their primary level of analysis is not personal but intrapersonal; (2) their representation of the intrapersonal dynamics of decision-making is based on information processes and systems. These two features set MaMs apart from traditional economic models and psychological theories which only adhere to one of these core features.

While MaMs seem poised to provide better understanding of the causes of decision errors and self-control problems, there are several key issues that pertain to how these models are conceived and how they afford scientific understanding. On the one hand, there is already a great deal of ambiguity surrounding the terms "selves" and "systems". In economics, selves have many meanings and many extensions; these range from the formal to the social to the evolutionary (Elster, 1987; Ross, 2005, 2006; Grayot, 2017). Likewise, in social and cognitive psychology, there have been many debates with regard to what counts as a cognitive system, and what discriminates cognitive systems from one another (Evans, 2006, 2008; Evans & Stanovich, 2013). On the other hand, and more importantly for this paper, there is ambiguity surrounding the very idea of intrapersonal dynamics. This second form of ambiguity is primarily due to the conflating of different levels at which decisions are made.

While some of the above problems are recognized (but unresolved), others seem not to have not been recognized at all. Moreover, there seems to be little concern over whether either of these ambiguities may affect the scientific value of MaMs. In what follows, I provide a systematic analysis of MaMs by way of three separate claims.

*Claim 1***:** MaMs are conceptually ambiguous. There have been some attempts to clarify the meaning of the terms "selves" and "systems" in the philosophical literature on economics and psychology; however, it remains uncertain and therefore contested what the terms refer to and pick out. Claim 1 first establishes that selves and systems are conceptually ambiguous prior to the integration of multiple-self models and dual-process and dual-system theories; it then demonstrates that MaMs perpetuate conceptual ambiguity by bringing these terms into close proximity via integration.

*Claim 2***:** MaMs are ontologically ambiguous. Because it is uncertain what selves and systems refer to or pick out, their roles as intrapersonal and/or intraneural agents may generate ontological ambiguity. This ambiguity is twofold. How selves and systems interact and generate (or resolve) conflict is obscured by the fact that MaMs appeal to both personal-level and subpersonal-level descriptions of agent capacities. In some instances, this conflation reveals a deeper ambiguity over the functional interpretation of conflict.

*Claim 3***:** It is uncertain how MaMs afford scientific understanding. Claim 3 is based on an analysis of three cases from behavioral decision research. The conceptual and ontological ambiguities identified by claims 1 and 2 indicate that MaMs lack explanatory power, and this undermines their scientific value. I argue that this is likely a result of researchers failing to define their target of explanation.

This paper has the following structure: In section 2, I provide an overview of the emergence of MaMs. This lays the groundwork for my analysis and establishes the first claim about conceptual ambiguity. In section 3, I argue that different kinds integrations of multiple-self models and dual-system theories leads to different types of ontological ambiguity, thus establishing the second claim. In section 4, I analyze three examples of MaMs, and I show how that each model exemplifies conceptual and ontological ambiguity. In section 5, I then consider how MaMs afford scientific understanding. I argue that there are fundamental problems which pertain to how they conceive internal conflict and how they explain reasoning errors. I then consider possible rebuttals to my claims. Section 6 concludes.

2. The emergence of multi-agent models: a brief overview

This paper interprets multi-agent models as the integration of multiple-self economic models and dual-system psychological theories. However, integration can take many forms and can mean many things.[2] For this reason, it will be helpful to first examine how economists and psychologists understand their models and theories so that we can differentiate features unique to each. This will help to illustrate why there is so much confusion surrounding the emergence of MaMs.

2.1 *From selves to systems and back*

From the perspective of economics, self-control problems are provocative not just because they can lead to violations of expected utility theory, but also because they blatantly contradict neoclassical conceptions of the economic agent. Multiple-self models thus emerged as a means to resolve these problems. While early intertemporal choice models were not interested in representing the psychological aspects of

[2] Integration is a concept with some philosophical baggage. From the perspective of philosophy of science, integration is frequently associated with interdisciplinarity. It is, however, a matter of debate whether the crossing of disciplinary boundaries – say, through the sharing of concepts and methods – constitutes genuine integration. This paper does not make any strong assumptions about integration. Rather tellingly, it is because decision researchers don't take part in such meta-theoretic debates that I conjecture integration is a vague concept. For philosophical discussions about integration and interdisciplinary in the behavioral sciences, see Grüne-Yanoff (2015, 2016).

decision-making (cf. Samuelson, 1937), innovations by Strotz (1955) and Phelps & Pollak (1968) demonstrated how 'generational' dynamics could lead to preference changes. The idea behind these models is that a person's latent preferences could be modeled as competing interests, which can be distinguished by unique value functions. As a precursor to modern multi-agent models, these generational models cleverly illustrated how myopic and weak-willed behaviors could be rationalized as a tradeoff between short-term and long-term selves.

Thaler & Shefrin (1981; cf. Shefrin & Thaler, 1988) were among the first to capitalize on this idea. Their "dual-self" model interpreted motivational conflict as a game between a long-run "planner" self and a short-run "doer" self. Though this was based on prevailing theories of mental accounting (Kahneman & Tversky, 1979; Thaler 1985), Shefrin & Thaler indicated that dual-self models were consistent with neuroscience evidence of the time (cf. Fuster, 1980). However, it wasn't until the late 1980's that researchers began to make explicit connections between economics and psychology. For instance, Elster (1987) and Loewenstein (1996, 2000) based their interpretations of short-term and long-term selves on psychological models of "hot" and "cold" emotional states. This precipitated early attempts by decision researchers to integrate multiple-self models with dual-processing models of cognition.

In sum, the partitioning of individuals into selves, each of which could be defined by an exclusive value function, enabled economists to make sense of decision anomalies, like impulsive consumption habits and self-defeating preference reversals.[3] However, as intrapersonal and intertemporal choice models have become more psychologically sophisticated, it's become less obvious how selves relate to the underlying causal and physiological processes of decision-making. I return to this point shortly.

By contrast, psychologists interested in the causes of decision errors have sought to understand how information is perceived, organized, and produces action. In a word, they study information processes. Since the 1970's (Schneider & Shiffrin, 1977), dual-process models help to understand how individuals 'attend' to and process stimuli. This has inspired a cottage industry in social and cognitive psychology. The classic interpretation of dual-process theory is that it differentiates "fast", "reactive", and "automatic" cognitive processes from "slow", "controlled", and "deliberative" ones (Schneider & Shiffrin, 1977; Epstein, 1994; Stanovich & West, 2000; Lieberman, 2003). The importance of this distinction is that it allows researchers to distinguish higher cognitive processing, which is associated with the ability to make

[3] For overviews of the history of time-discounting models and analyses of time preferences, see Loewenstein (1992) and Frederick, Loewenstein, & O'Donoghue (2002); For recent surveys on multiple-self modeling with regard to time, see Soman et al (2005) and Heilmann (2010). For a discussion of time-discounting models in relation to disciplinary integration, see Grüne-Yanoff (2015).

deliberative and informed judgments, from lower, more primitive forms of information processing, which are associated with emotional, visceral behavioral responses.

Efforts to distinguish clusters of processes have helped psychologists to identify which processes are involved in different decision situations. In this way, dual-system theories of cognition have emerged as an extrapolation (and perhaps, as a simplification) of dual-process models; principally, they explain how the differential activation of cognitive modes can support more complex mental operations, such as perceptual learning, rule-following and deductive inference, and counter-factual reasoning (Evans, 2006, 2008; Frankish & Evans, 2009; Evans & Stanovich, 2013).

Yet, the concept of a cognitive system remains somewhat ambiguous. Where dual-system theories initially served to connect disparate bodies of evidence in the dual-process literature, the concept of a system has taken on a perplexing array of qualitative and quantitative features by being transplanted into economic optimization models.[4] Notable economic psychologists have used the dual-system approach to test and predict a wide range of decision phenomena, from the specific effects of cognitive load on memory and computation, to the more general effects of priming on task judgment and selection (Strack & Deutsch, 2004; Kahneman & Frederick, 2002, 2005; Kahneman, 2003, 2011; Alós-Ferrer & Strack, 2014). These forays into the analysis of dual-systems upon judgment and decision-making have a rich history in the Heuristics and Biases program, as pioneered by Tversky & Kahneman (1973; 1974; cf. Kahneman, Tversky, & Slovic, 1982).

However, things get complicated when advances in dual-process and dual-system psychology are integrated with the multiple-self modeling techniques of economics. For example, Benabou & Tirole (2002, 2004), Bernheim & Rangel (2004), Benhabib & Bisin (2005), Loewenstein & O'Donoghue (2005) and Fudenberg & Levine (2006) each have sought to characterize the contradictory tendencies of temporally distinct selves by investigating how controlled and automatic processes influence choice behaviors over time. In some instances, the intrapersonal dynamic between sequential selves is taken to establish the limitations on the decision-maker's ability to exhibit self-control (Benabou & Tirole, 2002, 2004; Fudenberg & Levine, 2006). In other instances, the conflict between an individual's desire to consume now or later is interpreted as a "trade-off" between distinct systems, whose aims are regulated by the activation of different cognitive processes (Benhabib & Bisin, 2005; Loewenstein & O'Donoghue, 2005). Where the former integrative approaches presume a dual-self

[4] There are numerous debates about what constitutes a system in the dual-process literature (cf. Evans, 2006, 2008; Evans & Stanovich, 2013). One solution to this problem is to collate processes according to their generalized functions. This has resulted in the use of more neutral terminology: *System 1* vs. *System 2* (Stanovich, 1999; Stanovich & Evans, 2000) and *Type 1* vs. *Type 2* (Evans, 2012).

conception of the decision-maker that is temporally divided, the latter starts with a dual-system conception of decision-maker who is psychologically divided. These are paradigm examples of multi-agent models.

Going one step further, some neuroeconomic approaches to decision-making have modeled brain processes based on what economists perceive to be optimizing procedures. This technique presumes that the brain has limited energy resources and that it must allocate those resources efficiently in order to satisfy rewards. In this way, the brain is modeled as an optimizer with budget constraints. Research conducted by McClure et al (2004) and McClure et al. (2007), and further results obtained by Brocas & Carrillo (2008a, 2008b, 2014), suggest that individual decisions are the outcome of strategic interactions between domain-specific systems. This gives credence to the belief that resource allocation in the brain adheres to economic principles of optimization.[5]

2.2 *Conceptual ambiguity surrounding selves and systems*

While it may appear that multiple-self models and dual-system theories are aligned to provide better understanding of the causes of internal conflict, the terms "selves" and "systems" have much conceptual baggage. For instance, with regard to the status of selves in economics and decision theory, Elster remarked that:

> The conceptual strategies that have been used to make sense of this perplexing notion differ in many ways; with respect to how literally the notion of 'several selves' is taken, with respect to the principles of partition, and with respect to the modes of interaction between the systems. (Elster, 1988, p. 1)

This sentiment has since been echoed in debates in the philosophy of economics about the agency of persons who are internally divided. For instance, Ross (2005, 2010) and Davis (2003, 2011) have argued that individuals are collections of selves: these selves embody social and neural information relevant for making decisions and navigating the social world. Nevertheless, they disagree about how to interpret selves, and about what are the appropriate principles of partition (Grayot, 2017).

Similarly, dual-process and dual-system theories have also received criticism. Evans, a pioneer of the dual-processing movement, has expressed doubts about whether human cognition fit into a two-system framework:

[5] Though, alternative research by Glimcher et al. (2007) and Kable & Glimcher (2007) suggests that reward and information systems aren't as discrete as dual-system theorists make them out to be. They argue that the processes involved in decision-making are so highly distributed throughout the brain that it is better to think of it as a unitary system (Rustichini, 2008; cf. Vromen, 2011).

> Although it is striking that theorists in different areas have proposed dual systems with broadly similar characteristics, it is far from evident at present that a coherent theory based on two systems is possible. (Evans, 2006, p. 206)

> [And that] …my conclusion is that although dual-process theories enjoy good empirical support in a number of fields in psychology, the superficially attractive notion that they are related to the same underlying two systems of cognition is probably mistaken. (Evans, 2008, p. 271)

Nevertheless, there have been a few notable attempts to organize developments in behavioral decision research. One instance of this is Alós-Ferrer & Strack (2014), who map out the theoretical connections between economics and psychology, namely, to show how dual-process models and dual-system theories have provided economics with a "theoretical scaffolding" to interpret human behavior in the context of individual decision-making (Alós-Ferrer & Strack, 2014, p. 1). However, Alós-Ferrer & Strack's overview serves better as a literature review than as a philosophical analysis. Their belief that behavioral economics and economic psychology are distinct disciplines—each developed for the respective needs and purposes of their parent disciplines—has so far blocked them from addressing deeper conceptual and ontological problems that relate to the integration of economic and psychological modeling methods.

Two other instances are Rustichini (2008) and Brocas & Carrillo (2014). Unlike Alós-Ferrer & Strack, both Rustichini and Brocas & Carrillo explore how dual system theories have interfaced with neuroeconomics. I say "interfaced" for both seem committed to the view that the brain is a massively distributed optimizer, and that dual-system theories merely help to understand its optimizing procedures. To this end, Rustichini and Brocas & Carrillo endorse the same convention, namely, that "dual-system models" refer to information processing models, whereas "dual-self models" refer to intrapersonal bargaining models. Although these surveys get closer to the theme of this paper, neither of the authors defends this convention, which is to say, neither investigates how dual-process and dual-system models, as understood by social and cognitive psychology, might have been integrated with multiple-self models in economics.

To summarize, there are two senses in which MaMs are conceptually ambiguous. First, it is not established or easy to determine what the terms "selves" and "systems" refer to, either in economics or psychology. This is a well-known problem; though there is no easy solution. Second, once these terms are brought into close proximity—via attempts at integration in MaMs—further conceptual ambiguity ensues. This establishes claim 1.

3. Agency and ontological ambiguity

To recap, MaMs are not limited to any particular field of economics or decision research: they are utilized by economists and psychologists alike. This explains, in part, why terms like "selves" and "systems" have taken such a wide array of meanings, some of which appear to be coextensive. But even correcting for possible conceptual ambiguities, there are further reasons to believe that MaMs may be ontologically ambiguous. This has to do with the roles that selves and systems play as intrapersonal agents.

3.1 *The uncertain agency of selves and systems*

It's important to remember that MaMs are constrained optimization models—they are constituted by decision agents that have limited resources. These agents are maximizers in the traditional economic sense. But, unlike standard multiple-self models, which conceive selves as virtual solutions to intrapersonal problems, MaMs rely (to varying degrees) on cognitive and neuroscientific evidence to derive motivations for the agents they posit. While these motivations are represented by utility functions like their virtual counterparts, the solutions are determined by information processes in the mind and/or brain. Hence, if one wants to understand how MaMs represent the intrapersonal (or intraneural) dynamics of decision-making, one needs to consider how selves and systems function as economic agents. To do this, however, we need to briefly talk about agency.

According to the *Stanford Encyclopedia of Philosophy*, an agent is "a being with the capacity to act, and 'agency' denotes the exercise or manifestation of this capacity" (Schlosser, 2015). Although this definition is not especially sophisticated, it is a useful point of departure for it calls into question whether economic agency is a special kind of agency, and if so, whether it applies expressly to individual persons? One distillation of this question which continues to divide philosophers and methodologists of economics is whether rational choice models are intended to represent the cognitive capacities of human persons, or whether they're intended to represent an instrumental account of action, one that abides by the rational norms of economic theory. Proponents of the former interpretation are inclined to think that the economic agent portrayed by rational choice models should map one-to-one onto the human person, which is to say, the economic agent is ontologically anchored to the individual. This is the common approach taken by behavioral economists and economic psychologists. Proponents of the latter interpretation argue that economic agency is nothing more than a reference point for the ascription of a utility function, and that, in principle, any entity can be modeled as an economic agent, so long as its behavior, as revealed by its preferences, is consistent.

The rift above can be attributed to unresolved debates concerning the normative and descriptive applications of the concept of economic agency—I will not review them here. Based on this summary, I posit a simple diagnostic that will aid in illustrating the problem of ontological ambiguity in MaMs:

Economic agency implies human agency: It is common for researchers to equate the economic agent with the human person—for both normative and descriptive purposes, researchers regard human persons as prototypical decision agents. This can be, and often is, construed as a one-to-one ontological mapping between human person and economic agent as conceived by rational choice models.

Economic agency implies instrumental rationality: Yet, there is no reason to restrict the concept of agency to humans. Not all economic models require a one-to-one ontological mapping to be mathematically valid or empirically sound. This is what enables economists to posit virtual agents, and to treat non-human entities as instrumentally rational for purposes other than micro-economic evaluation.[6]

The reason why we need to differentiate between kinds of economic agency is because it ceases to be clear what (or rather, where) is the locus of decision-making when individuals are partitioned into selves and/or systems. On the one hand, MaMs may provide solutions to intrapersonal problems that supervene on information processes. For instance, a person may experience conflict as a result of competing urges, and may seek to resolve this conflict by engaging in a bargaining game with temporal selves. We may interpret the bargaining procedure as the virtual embodiment of one cognitive system exerting control over another. On the other hand, MaMs may directly manifest informational conflict between functionally and/or structurally discrete systems. For instance, under the same scenario, what determines whether one system exerts control over another (and resolves intrapersonal conflict) depends on the availability of resources. We may interpret resource limitations as a form of intraneural conflict. Just as the former can be construed in game-theoretic form, so can the latter. The issue, however, is whether the same game-forms apply, and this depends on what selves and/or systems represent. Hence, in moving from one resolution to another, both conceptions of economic agency may come into play.

As I will argue in section 4, it is because selves and systems occupy an uncertain ontological space that they can be harnessed for the expression of both personal and

[6] One should bear in mind that these two conceptions of agency are for pedagogical purposes only. The questions about which researchers are divided far exceed merely normative and descriptive applications of the concept of economic agency. For an analysis of the historical origins of these debates, see Ross (2005, 2010; cf. Davis, 2003, 2011); for further elaboration of Ross' arguments, see Grayot (2017).

subpersonal instances of conflict, which in turn, supervene on both functional and structural properties of the mind and/or brain. Let's now consider further how separate conceptions of agency generate ontological ambiguity.

3.2 *Two types of ontological ambiguity*

The benefit of teasing apart human agency and instrumental rationality is that it affords room to speculate about who (or what) is the primary target of MaMs. This, I argue, constitutes an ontological problem. If economists and decision researchers seek to identify the causes of self-control problems and reasoning errors (and not merely predict when they occur) then it will be in their interest to know and discern which properties are relevant to first-person experiences of conflict versus those properties that are not. This is not merely a philosophical concern: given their wide conceptual latitude, selves and/or systems may take on properties that do not seem to fit with received scientific models of the mind and/or brain. I return to this point in section 5. To put the issue of ontological ambiguity into clearer perspective, consider two lessons from the philosophy of mind and cognitive science.

Lesson 1: personal events are distinct from subpersonal events. Ontological ambiguity can occur when mental entities are identified with physical entities. Mental entities refer to things like thoughts and sensations, whereas physical entities refer to things like brain activity and events in the nervous system. Even if physical entities could explain how mental entities occur (via supervenience relations), physical entities are not accessible to introspection, which mental entities are. Mental entities are thought to describe personal-level events, while physical entities describe subpersonal-level events. This is considered a philosophical problem because personal events and subpersonal events describe different phenomena. (It follows that they require different kinds of evidence to be described as well, though this is a contested issue). The first form of ontological ambiguity (confusing the personal with the subpersonal) has been described by Dennett (1989, 1991; cf. Hornsby, 2000).

Lesson 2: functional design may be separate from physical structure. Ontological ambiguity can also occur when subpersonal-level events are not clearly delineated. For instance, events at the subpersonal level can be attributed to 'functional design' or to 'physical structure'. Functional design descriptions are, as the name suggests, functional: they describe input-output relations. but do not necessarily describe physical behavior of biological mechanisms. This is considered a philosophical problem because descriptions based on functional-design may not accurately represent causal relations, hierarchical organizations, etc. which descriptions at the physical-structural level are thought to represent. This second form of ontological ambiguity (confusing

functional design with physical structure) has been described by Marr (1982; cf. McClamrock, 1991).[7]

The above problems characterize familiar problems in the philosophy of mind and cognitive science. Agents are defined by their capacities to act, and because we are interested in knowing how selves and/or system interact, the generic cases above illustrate nicely how descriptions of capacities relate to or give rise to ontological ambiguity. In the next section I investigate how MaMs conceive of selves and/or systems as intra-personal agents, which means that I flesh out their capacities to (inter)act.

4. Three examples of multi-agent models in behavioral decision research

In the last section, I proposed two ways to think about the agency when individuals are partitioned into intrapersonal agents. In this section, I will defend my second claim, viz. that the uncertain status of selves and/or systems begets ontological ambiguity. This analysis is organized according to three questions: firstly, how are selves and/or systems conceived in each account; (2) how do selves and/or systems interact and represent internal conflict; (3) how does this dynamic lead to ontological ambiguity?

4.1 *A model of heuristic judgment (Kahneman & Frederick, 2005)*

The work of Daniel Kahneman and Shane Frederick draws inspiration from various sources in cognitive psychology (cf. Sloman, 1996; Chaiken & Trope, 1999; Gilbert, 1999, 2002). Yet, their model of heuristic judgment relies heavily on a distinction drawn by Stanovich & West (2000). According to the latter, System 1 and System 2 stand as labels for collections of cognitive processes that can be distinguished "by their speed, their controllability, and the contents on which they operate" (Kahneman & Frederick, 2005, p. 268). The Kahneman & Frederick (2005) model, like its predecessors (cf. Kahneman & Frederick, 2002; Kahneman, 2003) seeks to understand how the interactions of System 1 and System 2 give rise to judgment errors, which result in unsound decisions. While the dual-system approach to human reasoning has gained considerable traction, I argue that Kahneman & Frederick's particular model does a

[7] Dennett (1989) distinguishes three levels of abstraction—called "stances"—by which to understand human behavior. The *physical stance* understands behavior in terms of physiological processes; the *design stance* understands behavior in terms of a system's purposes; and the *intentional stance* understands behavior in terms of mentalistic, or folk-psychological explanations. Similarly, Marr (1982) distinguishes three ways of characterizing information processing. The most basic level is the biological level, or *implementation level*; this is followed by the *algorithmic level*, which pertains to functional descriptions; and lastly is the *computational level*, which describes the programs run by information systems.

poor job of characterizing the interaction of Systems 1 and 2. Their portrayal of the activation of cognitive and affective processes which correspond to these systems is not adequate to understand how conflict between systems leads to decision problems.

In the Kahneman & Frederick (2005) model, the heuristics of accessibility and representativeness reflect the rapid, automatic, and effortless nature of the processes of System 1. Accessibility refers to the means or ability of an individual to retrieve information. As a decision heuristic, it highlights the ease or naturalness with which the mind registers content and attributes of objects of choice—it is thus associated with memory-based judgments where frequency of experiences determines the likelihood of accessing relevant information. For instance, accessibility may explain how individuals quickly identify outliers in a group of physically similar objects without the aid of a measurement tool or guidance of a rule. Likewise, it may also explain how individuals respond to emotionally charged language or repulsive images before they consciously register them. Such responses are useful for avoiding danger and for making rapid judgments. However, when a task is too complex to be immediately comprehended, lack of accessibility may lead one to substitute-in information. "Attribute substitution" is the heuristic process by which individuals simplify a task or choice dilemma through retrieval of information that is present in mind; it typically involves replacing the key attributes of an object or proposition with attributes of another, more familiar object. Kahneman & Frederick (2005, pp. 269-74) argue that many of the systematic biases uncovered by previous research into static choice violations are due to attribute substitution.

Prima facie, Kahneman and Frederick's dual-system approach is not integrative in the same way that other behavioral economic and neuroeconomic accounts are. Yet, their model is predicated on a divided self, which interprets individual behavior as the outcome of the interaction of System 1 and System 2. Kahneman & Frederick endorse what is known as the "default-interventionist" model of dual-system theory, which posits that System 1 and System 2 are arranged sequentially. Under this view, System 1 and its concomitant processes are activated by default—the individual has no control over initial responses to external stimuli. System 2 is thought to intervene on System 1 when it detects errors in judgment. This is what allows it to subdue some impulsive behaviors. They describe this dynamic as follows:

> In the particular dual-process model we assume, system 1 quickly proposes intuitive answers to judgment problems as they arise, and system 2 monitors the quality of these proposals, which it may endorse, correct, or override… We assume system 1 and system 2 can be active concurrently, that automatic and controlled cognitive operations compete for the control of overt responses, and that deliberate judgments are likely to remain anchored on initial impressions. (Kahneman & Frederick, 2005, p. 267)

However, an unresolved problem with this model is that it is underdetermined how System 1 and System 2 interact—or rather, that it is unclear what it means to say that they "interact". Kahneman & Frederick claim that the "effect of concurrent cognitive tasks provides the most useful indication of whether a given mental process belongs to system 1 or system 2" and further, that, "Because the overall capacity for mental effort is limited, effortful processes tend to disrupt each other, whereas effortless processes neither cause nor suffer much interference when combined with other tasks…" (2005, p. 268). Accordingly, they interpret the monitoring function of system 2 to be dependent on the effort required to inhibit System 1. Thus, in order for System 2 to monitor and override System 1, it must have the resources to do so. But it is unclear whether resources refer to functional capacities (e.g., alertness or willpower) or whether they refer to physiological resources (e.g., GABA and dopamine). Kahneman & Frederick defer to the neurosciences to flesh this out—but this merely sidesteps the issue. In supposing that System 2 has some limited control over the automatic and unconscious processes of System 1, System 2 would, in some sense, have to constitute System 1. It is, in fact, a common criticism of system-based interpretations of dual-process theory that cognitive systems cannot be construed as discrete since many processes operate on a continuum. Kahneman & Frederick caution readers not to think of systems as "autonomous homunculi", and clarify that the term "system" is merely a "label for collections of cognitive processes that can be distinguished by their speed, their controllability, and the contents on which they operate" (Kahneman & Frederick, 2005, p. 267). But this simply begs the question—their account must presuppose that System 1 and System 2 have the functional characteristics they do because cognitive processes operate on a continuum.[8]

The issue that arises here is that, without a clearer understanding of how System 1 and System 2 operate and interact, it becomes uncertain what, exactly, Kahneman & Frederick's model of heuristic judgment tells us about the intrapersonal dynamics of decision-making (aside from the obvious fact that people sometimes lack the focus or training to avoid biases in judgment). In focusing on the functional characteristics of System 1 and System 2, their account straddles an ontological divide which requires System 2 perform both personal-level and subpersonal-level functions. On the one hand, System 2 is responsible for capacities which support conscious control, reflection, and rational deliberation—things we attribute to persons; but, on the other hand, System 2 must frequently perform subpersonal-level tasks which allow it to function

[8] Although most proponents of the System 1 / System 2 distinction endorse the "default-interventionist" model, it is not agreed what the appropriate neuroanatomical correlates of System 1 and System 2 are, or would be, and for this reason the story of their interaction is mired in theoretical disputes about the functional design dual-system models (cf. Osman, 2004; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2009; Keren, 2013; Mugg, 2016). Furthermore, there isn't sufficient empirical evidence to validate either the default-interventionist or the parallel-competitive interpretation of system interaction. (Sinayev, 2016; Pennycook, 2017).

as monitor of System 1. What this means is that the Kahneman & Frederick model is ambiguous with regard to the personal–subpersonal ontological distinction. So, although there is no intrinsic problem with how they conceive of System 1 (it is solely and unambiguously comprised of subpersonal processes), there is a problem with how they conceive of System 2. I return to discuss implications for this ontological ambiguity in section 5.

4.2 *The brain as hierarchical organization (Brocas & Carrillo, 2008a)*

Brocas & Carrillo's (2008a) neuroeconomic framework is predicated on a modular interpretation of the brain. This means that they take different systems in the brain to literally compute or process information in line with some biological function. To this end, each biological function requires the intervention of several other systems, whose network connections impose constraints on the availability of energy. Insofar as these modules can be interpreted as having independent goals, Brocas & Carrillo (2008a) argue that the brain can be modeled as an organization of hierarchical systems, in which the hierarchy is determined by the flow of information between regions of the "reflective" system and "impulsive" system. Thus, their framework derives several models, each of which purports to show how conflict between brain systems gives rise to time preferences and related decision errors. I argue, however, that the notion of physiological conflict is contentious in Brocas & Carrillo's (2008a) framework. It is often not clear what is the resolution, or level, at which conflict occurs and at which information is constrained.

In their (2008a) publication, Brocas & Carrillo present three ways that conflict in the brain gives rise to decision errors. For the sake of space, I will concentrate on what they call information asymmetry. Information asymmetry, as the name suggests, refers to physiological constraints on information flow between brain regions. The flow is determined to be asymmetrical precisely because neural connectivity is a limited resource and most brain areas are unidirectionally linked to others. This results in limited awareness of individuals' motivations for their decisions (2008a, p. 1315).[9]

With regard to modeling physiological conflict, Brocas & Carrillo adopt several principal-agent configurations to represent the interaction of systems in the brain. Each configuration corresponds to a different cognitive operation. For illustrative purposes, let's consider a sample model that formally represents the interactions of the impulsive system and the reflective system in the brain. Space constraints prohibit me from giving full attention to the formal results provided in (Brocas & Carrillo, 2008a) which distinguish consumption and labor behavior under full information and

[9] The other two forms of conflict are *temporal horizon* and *incentive salience*. Here I concentrate specifically on information asymmetry for it is crucial for understanding the physiological basis of their endogenous discounting model.

imperfect information. This discussion is limited to the basic formal exposition provided in (Brocas & Carrillo, 2008b).

Suppose an individual lives an infinite number of periods $t \in \{1, 2, \ldots T\}$; she works $n_t \in [0, \bar{n}]$ and consumes $c_t \geq 0$. Each unit of labor worked entails that the individual has an additional unit to spend. The individual is thus divided into separate systems: the agent (which corresponds to the impulsive system) is myopic and informed, i.e. it 'knows' the relative desirability of a consumption package. It's preferences at t are depicted as

$$U_t = \theta_t u(c_t) - n_t$$

where $u' > 0$ and $u'' < 0$. θ_t is privately known and represents the marginal value of consumption at time t. Likewise, the principal, (which corresponds to the reflective system) is forward-looking and uninformed, i.e. it does not 'know' the value of θ,. The principal weighs the utility of all agents under a budget constraint that links lifetime consumption and lifetime labor

$$S = \sum_{t=1}^{T} E\left[\theta_t u(c_t) - n_t\right]$$

where S captures the intertemporal utility of the principal from the perspective of t.

The first caveat of this formulation is this that if the principal knew θ_t the 'existence' of agents would be irrelevant. Thus, presuming informational asymmetry, the principal at each date proposes a menu of incentive compatible pairs to the agent:

$$\{c_t(\theta_t), n_t(\theta_t)\}_{\theta_t \in [\underline{\theta}, \bar{\theta}]}$$

Where $c_t(\theta_t)$ denotes a consumption package and $n_t(\theta_t)$ denotes the labor the agent is incentivized to work if she wishes to consume it (this is comparable to a contract with hidden information).

Brocas & Carrillo determine that the optimal strategy for the principal, given that she does not know the private value of θ_t, is to restrict the agent's choices at each period so as to maximize her own utility. This result gives rise to a self-disciplining rule of "work more today to consume today" (2008a, 4). This allows agents (i.e. regions of the brain that make up the impulsive system) to pursue immediate rewards within the restrictions set by the principal (i.e. regions of the brain that make up reflective system). In essence, this configuration portrays a precommitment technology set by the reflective system.

It's important to keep in mind that this formulation—i.e. the interaction of agent and principal—are representations of neural activity that is not accessible to the individual by introspection. Thus, when Brocas & Carrillo claim that the principal does not "know" the value of θ_t, what they actually mean is that neural network connections do not allow the reflective system to receive information; information reaches the impulsive system sooner than the reflective system can monitor it. According to Brocas & Carrillo, the reflective system prohibits the impulsive system from seeking immediate gratification by regulating the information it receives. By analogy, the principle ensures that what the agent consumes is within a menu of the principal's choosing. But this is not very illuminating given that everything a brain does depends on the flow of information.

To recap, Brocas & Carrillo use the terms "reflective system" and "impulsive system" to identify regions of the brain that process information relevant to the achievement of different cognitive functions. (The impulsive system is analogous to System 1 processes, whereas the reflective system is analogous to System 2 processes). The principal-agent model depicts a formal relationship, one which is based on the principal having imperfect information. In reality, this relationship is not based on imperfect information but on asymmetrical information, which incurs intraneural conflict. But this raises an important question, which is whether the principal-agent model depicts one system ("the automatic system", comprising both reward prediction and motor preparation for consumption), within which information flow is disrupted, or two systems (a "reward system" and a "motor preparation system") between which information delivery is prevented. Brocas & Carrillo do not seem to be very concerned with the distinction, as they are more interested to show that economic theory can be useful for understanding how the brain acts like an optimizer:

> The methodology used in neuroeconomic theory is in fact quite close to the methodology economists rely on to represent the choices of an individual assuming he is a coherent entity. We are simply taking one step back: the coherent unit is not the individual but rather the cells (and perhaps the systems) that compose him" (2008b, p. 46).

However, the clause "and perhaps the systems" turns out to be an important bit of information that could drastically change how their model is interpreted. Because of precisely this, it is uncertain whether intraneural conflict occurs at the level of systems or at the level of cells.[10] Given their conception of systems as brain regions, the former

[10] Compare this with the following claim: "In our 'as if' methodology, each system wants to pass reliable information given its objective. However, this information may contradict the information passed by a different system. A third system may then inhibit the activity of one of the systems to distort the decision in favor of the other. Overall, behavior can be represented as the result of an interplay between systems with different objectives, and the particular nature of the interaction will vary across decision problems" (Brocas & Carrillo, 2014, p. 47).

would constitute an understanding of optimization based on the brain's functional design, whereas the latter would constitute and understanding of optimization that is based on its physical structure. Because they waver between the two, this indicates that Brocas and Carrillo's neuroeconomic framework is ontologically ambiguous. I return to this point in section 5.3, where I show how this ontological ambiguity leads to explanatory problems.

4.3 *Deliberative vs. affective systems (Loewenstein & O'Donoghue, 2005)*

Loewenstein & O'Donoghue's account can be read as an attempt to generalize results obtained by other intertemporal dual-self models. In particular, their model interprets differently motivated selves based on the functions of the "affective" and "deliberative" systems (these are analogous to System 1 and System 2). When an individual makes a choice, the interaction of the affective and deliberative systems result is a tradeoff—or 'effort cost'. The effort cost expresses an individual's 'quantity' of will power, which dictates their self-control. However, what will power is, and how it informs their conception of an effort cost creates ambiguity about the very idea of conflict. I will argue that these ambiguities arise because Loewenstein & O'Donoghue do not clearly establish the roles of systems as intrapersonal agents.

When an agent makes a choice, x (within some set of choices, $x \in X$), the interaction of the affective and deliberative systems results in a tradeoff—what they call an 'effort cost'—between the affective optimum, which describes the 'choice' the affective system would make free of influence from the deliberative system, and the deliberative optimum, which describes the 'choice' the deliberative system would make free of any influence from the affective system. The affective optimum is represented $x^A \equiv \arg\max_{x \in X} M(x,a)$, where $M(x,a)$ is the motivational function which captures the affective system's desire for x. The deliberative optimum is represented as $x^D \equiv \arg\max_{x \in X} U(x)$, where $U(x)$ is the utility of the deliberative system's choice. The interaction between the two systems is thus represented:

$$V(x) \equiv U(x) - h(W,\sigma) * [M(x^A,a) - M(x,a)]$$

where $h(W,\sigma) *[M(x^A,a) - M(x,a)]$ represents the cognitive effort exerted by the deliberative system over the affective system (with $h(W,\sigma)$ representing the cost to mobilize willpower). In short, the value function computes the deliberative optimum, measured in utility, minus the effort it takes to regulate the affective system. Loewenstein & O'Donoghue claim that this model captures the effort cost it takes for a person to exert control over their impulses.

Notice, however, that valuation for the deliberative system is measured as a function of *utility*, whereas valuation for the affective system is measured as a function of

motivation. The motivation function (captured by the affective optimum x^A) is taken to be exogenous to the deliberative system's utility function—presumably, this is because the processes associated with the affective system are activated by parts of the brain that are inaccessible to introspection (2005, p. 3). This could be read as an indication of their ontological stance regarding the target and explanatory aim of their model—viz. that it seeks to explain individual-level behavior through the effects of sub-personal processing.

To illustrate this point, Loewenstein & O'Donoghue describe how the deliberative system values a single choice with an outcome spread over time—i.e. an action x has an immediate pay-off $z_1(x)$ and a future pay-off $z_2(x)$. The affective system, being myopic and driven to consume immediately, has a motivational function $M(x) = z_1(x)$, whereas the deliberative system, which values both immediate and future rewards, has a utility function $U(x) = z_1(x) + z_2(x)$. A choice which maximizes x given *both* values can be represented as:

$$V(x) = [z_1(x) + z_2(x)] - h^*[z_1(x^A) - z_1(x)]$$

where the inclusive value of $z_1(x)$ and $z_2(x)$ are diminished by the effort cost to regulate the affective system. Given that the affective optimum is exogenous, the pay-off $z_1(x)$ effectively tips the weighted sum of the two pay-offs toward the immediate reward. This, Loewenstein & O'Donoghue argue, is equivalent to maximizing

$$\tilde{V}(x) = z_1(x) - [1/(1+h)]^* z_2(x)$$

which depicts a natural discounting function. In this reformulation, $[1/(1+h)] < 1$ indicates that the deliberative system will devalue future pay-offs, not because it has time preferences of its own, but because the joint attention toward immediate rewards by both systems will outweigh any interest the deliberative system has for separate future pay-offs.

Loewenstein & O'Donoghue seem to commit the same initial error as Kahneman & Frederick, viz. they attribute higher cognitive functions to the deliberative system, while portraying the affective system as a collection of automatic processes. Their justification for this is twofold: Firstly, given that the deliberative system is associated with the operations of the prefrontal cortex, only the deliberative system is capable of making decisions. This explains why, by contrast, they refer to the affective system's optimum as a motivation function, not as a utility function. However, unlike Kahneman & Frederick (but like Brocas & Carrillo), Loewenstein & O'Donoghue portray the interaction of systems by way of a principal-agent formalism, not by explicit descriptions of either functional or physiological processes. The methodology for depicting this interaction, whereby the principal trades off its own utility to restrict the

choices of the agent, obscures what intrapersonal conflict is and how cognitive control is achieved. This is partly due to the fact that the concept of a system is left open-ended. The claim, "We refer to the two processes as 'systems' simply to underline the fact that they can generate divergent motivations, not to suggest that they are operate independently or are physiologically distinct" (2005, p. 9). But this admission doesn't help their cause. Even if Loewenstein & O'Donoghue contended that their understanding of systems is purely functional, they still abuse the concept of a higher cognitive system by expecting it to do all sorts of things that persons could not consciously do, i.e., monitor processes of the affective system, calculate the utility costs to exert control. Like, Kahneman & Frederick, the ontological status of the deliberative system is ambiguous; and this status is only exacerbated by the fact that they introduce willpower as its primary cognitive resource. So, not only is Loewenstein & O'Donoghue's model ignorant of how the deliberative system and affective system interact physiologically, but it is ontologically ambiguous with regard to intrapersonal conflict because it involves both personal-level and subpersonal-level events.

5. Implications for scientific understanding

In the introduction of the paper I presented three claims: The first claim is that selves and systems are ambiguous concepts. The second claim is that MaMs can be ontologically ambiguous. However, it could be argued that these are philosophical problems that have limited scientific import. This leads to this paper's third and final claim, viz. that decision researchers should take conceptual and ontological ambiguity seriously because they possibly undermine scientific understanding.

In support of this claim, I argue that MaMs may lack explanatory power, and this undermines their scientific value. To demonstrate this, I will revisit the cases above. My inquiry is organized around two questions: What does the model purport to explain? and How does the model achieve this goal? In each case, there is a disruption between the purported aim and the means to achieve that aim. I attribute this disruption to the fact that each of the above MaMs fails to define its explanandum.

5.1 *What does the model purport to explain?*

Recall that each MaM discussed above purports to understand how different decision errors arise, and in some instances, this is used to make sense of self-control problems. To recap, here is how each MaM pursues this goal.

Kahneman & Frederick's model of heuristic judgment is designed to show how the functioning of System 1 and System 2 relate to various reasoning techniques, called heuristics, and failures of reasoning by way of biases. The goal of the model is to provide a map which identifies different causes of reasoning errors, those which

provoke System 1 into action, and those which prevent System 2 to override System 1. Their explanation of how this happens amounts to a description of the conditions which can lead System 2 to fail to 'intervene' and stop System 1 from carrying out irrational behaviors.

Brocas & Carrillo's model, the brain as hierarchical organization, seeks to identify an endogenous discounting function in the brain which explains how individuals reverse preferences. Their goal is to utilize neuroscientific insights about information asymmetries in the brain to explain how intraneural conflict arises, and how it leads to decision errors. Their explanation is based on a model of the brain which views neural systems in a hierarchical relation to one another, which can be depicted as if they are utility optimizers.

Loewenstein & O'Donoghue's model of deliberative and affective systems tries to provide a generalized decision model which demonstrates how the deliberative system 'decides' to intervene on the affective system. Unlike Kahneman & Frederick's dual-system approach, they attempt to quantify the effort it takes the deliberative system to override the impulses of the affective system. This primary aim of their model is to improve both the predictive and explanatory power of dual-self models which utilize both psychological and neuroscientific evidence.

Having stated the purported scientific goals of each model, we can now consider the second question, "how does the model achieve its goal?" Below I answer this 'how' question by demonstrating that MaMs are not sufficiently explanatory.

5.2 *How does the model achieve this goal?*

A model of heuristic judgment. Because Kahneman & Frederick's scientific goals are modest compared with the other two, their problems are simpler. In short, Kahneman & Frederick do not adequately explain how System 1 and System 2 interact. To reiterate section 4.1, I consider how System 1 and System 2 function as intrapersonal agents. System 1 is the default system, which means that its capacities are not accessible to introspection, whereas System 2 is the intervening system, which means that it monitors system 1 to prohibit rapid judgments from implementing bad decisions. But, System 2 also is described as a "rational" system, one which is involved in careful deliberations. For this reason, it's not clear whether System 1 and System 2 are really separate entities. This ontological ambiguity is at the heart of the some important—and well documented—explanatory problems.[11] For instance, Kahneman and Frederick can't explain how the 'monitoring' and 'intervening' operations of System 2 upon System 1 actually work. In fact, rather it seems that these descriptions are metaphorical, not scientific (which Kahneman has alluded to elsewhere, cf. 2011).

[11] See, e.g., Osman (2004); Keren & Schul (2009); Kruglanski & Gigerenzer (2009); Keren (2013); Mugg (2016).

There is a litany of reasons to question the System 1 / System 2 distinction. However, to be charitable to the scientific aims of Kahneman & Frederick's model, we should evaluate it on whether it achieves its purported aim to explain how System 1 and System 2 relate to reasoning errors. To this end, the model works perhaps as a loose framework. But as an explanatory model for the purposes of understanding how decisions are made, it is insufficient.

The brain as hierarchical organization. B&C believe that the asymmetric flow of information between systems in the brain is the cause of some reasoning errors. Yet, it's difficult to tell whether this flow of information constitutes a causal relationship or a merely functional one, which they flesh out with formal optimization models. It is quite uncertain whether their goal is to explain what brains actually do or to justify the use of optimization models to organize brain functions. This lack of a clear explanatory target seems to follow from the ontological ambiguities discussed above.

Because their model focuses on the brain (and not the individual), it was established that their model conflates the functional characteristics of systems with their neural signaling pathways. This gives the impression that there is more going on in the brain than two systems competing for energy resources. It suggests that there may be multiple systems with varying degrees of control over the flow of information, with some central mechanism governing the flow of energy resources in a strategically optimal way. In fact, Brocas & Carrillo cite evidence for the existence of a central resource allocation system, which they posit as a candidate for a third system (cf. Brocas & Carrillo, 2014).[12] But if this is the case, then we encounter a tension between the functional characteristics of the reflective and impulsive systems wherein it is uncertain how conflict arises, and the physical structure of these systems, which are understood as distributed regions of neural networks whose functions vary according to features of the individual's choice environment.

Though, one may be inclined to think that there is no ontological ambiguity here since, even though Brocas & Carrillo believe that the "systems" to which they refer are neural signaling pathways, it's unproblematic to read them also as providing a functional conceptualization of systems. I am sympathetic to this line of reasoning. But, if one interprets them as providing an economic interpretation of the functional characteristics of neural signaling pathways, it's then unclear whether these characteristics are artifacts of the principal-agent model (and other economic formalisms which they use), or whether they identify functions unique to the brain's physical structure (as opposed to brain functions that are incidentally picked out by their

[12] They state, "some areas of the lateral prefrontal cortex play an active role when attention is divided, for instance when two tasks have to be completed at the same time. This points to the existence of what has been called a 'Central Executive System' whose role is to coordinate the systems involved in the different tasks" (2014, p. 50).

model). Hence, If Brocas & Carrillo's justification for adopting the neuroeconomic approach is that it provides a descriptively superior model of optimization procedures as they occur in the brain, then we should not expect there to be ambiguity about the how functional characteristics supervene on the brain—otherwise why adopt the 'as-if' approach in the first place? Why not stick to the conventional methods of functional neuroscience?

In sum, while Brocas & Carrillo may provide sophisticated and explicit descriptions of the antecedents of physiological conflict, it's not certain what they intend to explain. The ontological tensions between their interpretation of the functional design and physical structure of neural signaling pathways in the brain can be seen as a downstream effect of their failure to define where intraneural conflict arises within systems.

Deliberative vs. affective systems. In section 4.3, I argued that Loewenstein & O'Donoghue's model is ontologically ambiguous in two ways: firstly, it wavers between the personal and sub-personal level in its portrayal of intrapersonal conflict. Secondly, it invokes both functional-design and physical-structure descriptions to justify the use of a principal-agent model, though it does not carefully distinguish these. We can extract two explanatory problems from these ambiguities.

Recall that the deliberative system is thought to 'calculate' an effort cost which is based on some quantity of will power. They closest thing to a non-metaphorical explanation they give is a quick and conceptually vague description of will power (cf. Baumeister & Vohs, 2003). They liken will power to an energy source which the deliberative system needs to perform its function. But our question is, how does will power inform their conception of intrapersonal conflict? How does the deliberative system 'monitor' and 'intervene' upon the affective system? Like Kahneman & Frederick, the interaction between systems is explained away as a topic for the neuroscientist. Even if we were to grant this, the question about how will power relates to cognitive effort, and how this is 'calculated' by the deliberative system, is left unexplained. By substituting descriptions of physiological processes (which waver between functional design descriptions and physical structural descriptions) with optimization models, the inter-system dynamic is effectively relegated to a black box. This, rather than improving explanatory power, diminishes it. Ultimately, it is unclear what Loewenstein & O'Donoghue think intrapersonal conflict consists in, or how it is generated.

5.3 *Rebuttals and reconsiderations*

One could argue that this paper's diagnosis of MaMs betrays a conservatism that is, in reality, not very interesting to the cognitive or behavioral scientist, and that my emphasis on ontological ambiguity relies too much on a philosophically nuanced

critique of functional explanations. I would like to address this concern by contrasting the above accounts with a family of models from functional neuroscience that do not share these problems. The study of addiction is an apt example here as it coincides with this paper's theme on reasoning errors and self-control problems.

There is increasing evidence to suggest that the neurochemical basis of addiction lies in the human midbrain (striatial / dopamine circuit) is relatively autonomous from frontal systems (orbitofrontal and pre-frontal cortex), which are typically associated with executive functioning and cognitive control. the striatum, which projects from the midbrain, can be treated as if it were external to the agent because its valuations of attention and motor cuing occur prior to the activation of frontal systems. an economic model can then represent a striatum that has learned to consume addictively as imposing an exogenous cost on the agent's efforts to optimize, and the agent remains unambiguously virtual and functional (much like the account discussed in section 4). As argued in Ross (2012), "such models provide algorithms by which the reward system is taken to estimate the expected opportunity costs of attending to one stimulus rather than another and of preparing one motor response rather than another" (p. 719; cf. Montague & Berns, 2002; Ross, 2008). Hence, what has led behavioral economists and neuroeconomists discussed in this paper get into trouble is that they address, by way of functional models, intracortical processes for which neurochemical specifications are not yet in hand.

The economic models are thus, in part, speculations about intracortical mechanisms. But, the reason that functional and neurochemical models of addiction cohabit comfortably in some neuroeconomic models (i.e. "neurocellular economics"—Ross, 2008) is they don't confuse functional characteristics of intracortical agents with interpersonal conflict. This is, in summation, why the dual-system and dual-self models of behavioral economists tend to fall into ontological ambiguity. Whether an account like Brocas & Carrillo's dodges this general critique is hard to say because they are far less concise in their depiction of the functional characteristics of neural signaling pathways.

This consideration heeds another possible rebuttal, which is whether MaMs really are about intrapersonal and/or intraneural conflict? It could be argued that perhaps I am putting too much stock in the notion of conflict, or that I (wrongly) interpret MaMs to be solely about this issue. Admittedly, the paper uses the metaphor of the divided self as way of motivating my diagnosis of MaMs, so I do, in a sense, presuppose that which I analyze. However, even if the accounts I cite are not interested in conflict per se, the act of partitioning individuals into intrapersonal and intraneural agents—whether construed as selves or systems (or both)—suggests that reasoning errors arise due to complicated internal dynamics.

It could be argued that I am trying to make comparisons among models that are not so easily comparable. Along these same lines, it could be argued that my concept

of MaM is too generic, that it gives the wrong impression about what each model tries to explain relative to the others. I have two responses:

First, the issue of comparability is important. The reason for investigating MaMs is that there is not enough information about models that synthesize or integrate economics and psychology, especially with regard to models that partition individuals into simpler agents. The few authors that have attempted to review this literature, who I discussed in section 2, came up short of the goal I seek here. What I provide is a philosophically precise analysis that links reasoning errors to internal conflict, which so far has been sorely missing from the decision research literature.

Second, the issue of comparability unmasks an inherent challenge to writing this paper. One cannot, it seems, embark on such a complicated analysis without also getting tangled in debates about interdisciplinarity and integration. These are, without a doubt, important debates – especially as it pertains to the long and complicated histories of economics and psychology. Yet, these are different debates. My interest here is not in dictating which discipline should hold ownership of MaMs, but in bypassing this question. Behavioral decision researchers are not historians of science and do not make modeling decisions on the basis of their disciplinary loyalty. While the case could be made that economists borrow more from psychologists, it would put the cart before the horse to claim that MaMs are just instances of behavioral economic modeling. The question I have pursued in this paper is how to make sense of models that posit selves and systems as intrapersonal and intraneural agents, and my response was to compare and contrast three unique cases which utilize these concepts in similar but subtly different ways. It is because there are no meta-theoretic rules dictating which methods researchers can use that we need a wider purview to begin analyzing the different ways economics and psychology have been integrated. This, it seems to me, should come before we set limitations on the concept of integration.

6. Concluding remarks

While the divided mind is familiar metaphor, this paper argues that how researchers conceptualize and implement this idea with formal models and theoretical language has led to confusion about how to represent the intrapersonal dynamics of decision-making. Although multi-agent models would seem to be a boon for interdisciplinary decision research, the rapid integration of multiple-self modeling techniques with dual-system theories has led to confusion about what, exactly, causes internal conflict. Attempts at integration have shown researchers assimilating dynamical processes that are not only conceptually quite different, but also involves crossing ontological boundaries. I discussed three instances of this, from economic psychology, behavioral economics, and neuroeconomics. I concluded from this investigation that conceptual and ontological ambiguities are not merely philosophical problems. They are

scientific problems, insofar as decision researchers desire to explain how reasoning errors and self-control problems are generated by intrapersonal or intraneural conflict.

7. Bibliography

Alós-Ferrer, C., & Strack, F. (2014). From dual processes to multiple selves: Implications for economic behavior. *Journal of Economic Psychology*, 41, 1-11

Baumeister, R.F., & Vohs, K.D. (2003). Self-regulation and the executive function of the self. *Handbook of Self and Identity*, 1,197-217.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3), 871-915.

Benhabib, J., & Bisin, A. (2005). Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions. *Games and Economic Behavior*, 52(2), 460-492.

Bernheim, B.D., & Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review*, 94(5), 1558-1590.

Brocas, I., & Carrillo, J.D., (2008a). The brain as a hierarchical organization. *The American Economic Review*, 98(4), 1312-1346.

Brocas, I., & Carrillo, J.D., (2008b). Theories of the Mind. *The American Economic Review*, 98(2), 175-180.

Brocas, I., & Carrillo, J.D. (2014). Dual-process theories of decision-making: A selective survey. *Journal of Economic Psychology*, 41, 45-54.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process Theories in Social Psychology*. Guilford Press

Davis, J.B. (2003). *The Theory of the Individual in Economics*: *Identity and Value*. Routledge.

Davis, J.B. (2011). *Individuals and Identity in Economics*. Cambridge University Press.

Dennett, D.C. (1987). *The Intentional Stance*. Cambridge (MA): MIT Press.

Dennett, D.C. (1991). *Consciousness Explained*. New York: Little Brown & Co.

Elster, J. (Ed.). (1987). *The Multiple Self*. Cambridge University Press.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8), 709.

Evans, J.S.B.T. (2006). Dual system theories of cognition: Some issues. *In Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 202-207.

Evans, J.S.B.T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255-278.

Evans, J.S.B.T., (2012). Dual process theories of deductive reasoning: facts and fallacies. In Holyoak, K., & Morrison, R. (Eds.), *The Oxford handbook of Thinking and Reasoning*, pp.115-133. Oxford University Press.

Evans, J.S.B.T., & Stanovich, K.E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2), 351-401.

Fudenberg, D., & Levine, D.K. (2006). A dual-self model of impulse control. *The American Economic Review*, 96(5), 1449-1476.

Fudenberg, D., & Levine, D.K. (2011). Risk, delay, and convex self-control costs. *American Economic Journal: Microeconomics*, 3(3), 34-68.

Fuster, J.M. (1988). Prefrontal cortex. In *Comparative Neuroscience and Neurobiology* (pp. 107-109). Birkhäuser Boston.

Gilbert, D.T. (1999). What the mind's not. In Chaiken, S., & Trope, Y. (Eds.), *Dual-process Theories in Social Psychology*, (pp.3-11). Guilford Press.

Glimcher, P.W., Kable, J., & Louie. K. (2007). Neuroeconomic studies of impulsivity: now or just as soon as possible? *American Economic Review* 97(2), 142–147.

Grayot, J. (2017). The Quasi-economic agency of human selves. *Œconomia. History, Methodology, Philosophy*, 7(4), 481-511.

Grüne-Yanoff, T. (2015). Models of temporal discounting 1937–2000: An interdisciplinary exchange between economics and psychology. *Science in Context*, 28(4), 675-713.

Grüne-Yanoff, T. (2016). Interdisciplinary success without integration. *European Journal for Philosophy of Science*, 6(3), 343-360.

Heilmann, C. (2010). *Rationality and time* (Doctoral dissertation, PhD Thesis, London School of Economics and Political Science).

Hornsby, J. (2000). Personal and sub-personal; A defense of Dennett's early distinction. *Philosophical Explorations*, 3(1), 6-24.

Kable, J.W., & Glimcher. P.W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12), 1625-1633.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5), 1449-1475.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In Kahneman, D., & Gilovich, T. (Eds.), *Heuristics and biases: The psychology of Intuitive Judgment*, 49-81.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In Holyoak, K., & Morrison, R. (Eds.), *The Cambridge Handbook of Thinking and Reasoning*, (pp. 267-293). Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263-292.

Kahneman, D., Slovic, P., & Tversky, A. (1982). (1982) Judgment under uncertainty: Heuristics and biases. *Judgement under uncertainty: Heuristics and biases*, (pp. 3-20). Cambridge University Press.

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3), 257-262.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4(6), 533-550.

Kruglanski, A.W., & Gigerenzer, G. (2011). Intuitive and deliberative judgments are based on common principles. *Psychological Review*, 118, 97–109

Lieberman, M. (2003). Reflexive and reflective judgment processes: A social cognitive neuroscience approach. In Forgas, J.P., Williams, K.D., & von Hippel, W. (Eds.), *Social Judgments: Explicit and Implicit Processes* (pp. 44–67). New York: Cambridge University Press.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3), 272-292.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *The American Economic Review*, 90(2), 426-432.

Loewenstein, G., & O'Donoghue T. (2005). Animal Spirits: Affective and Deliberative Processes in Economic Behavior. *CMU Working Paper*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, New York

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, 1(2), 185-196.

McClure, S.M., Laibson, D.I., Loewenstein, G., & Cohen J.D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695), 503-507.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience*, 27(21), 5796-5804.

Montague, R., & Berns, G. (2002). Neural economics and the biological substrates of valuation. *Neuron* 36, 265–84.

Mugg, J. (2016). The dual-process turn: How recent defenses of dual-process theories of reasoning fail. *Philosophical Psychology*, 29(2), 300-309.

Osman, M., (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11(6), 988-1010.

Phelps, E.S., & Pollak, R.A. (1968). On second-best national saving and game-equilibrium growth. *The Review of Economic Studies*, 35(2), 185-199.

Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Ross, D. (2006). The economic and evolutionary basis of selves. *Cognitive Systems Research*, 7(2-3), 246-258.

Ross, D. (2008). Two styles of neuroeconomics. Economics & Philosophy, 24(3), 473-483.

Ross, D. (2012). The Economic Agent: Not Human, But Important. In Mäki, U. (Ed.), *Philosophy of Economics* (pp. 691-735). Amsterdam: Elsevier.

Rustichini, A. (2008). Dual or unitary system? Two alternative models of decision making. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4), 355-362.

Samuelson, P.A. (1937). A note on measurement of utility. *The Review of Economic Studies*, 4(2), 155-161.

Schlosser, M.. (2015). Agency. In Zalta, E. (Ed.) *The Stanford Encyclopedia of Philosophy*, Stanford University. Fall 2015 Edition.

Shefrin, H.M., & Thaler, R.H. (1988). The behavioral life-cycle hypothesis. *Economic Inquiry*, 26(4), 609-643.

Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127.

Soman, D., Ainslie, G., Frederick, S., Li, X., Lynch, J., Moreau, P., Mitchell, A., Read, D., Sawyer, A., Trope, Y. and Wertenbroch, K. (2005). The psychology of intertemporal discounting: Why are distant events valued differently from proximal ones?. *Marketing Letters*, 16(3), 347-360.

Stanovich, K.E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning*. Psychology Press.

Stanovich, K.E., & West, R.F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94(4), 672.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8(3), 220-247.

Strotz, R.H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3), 165-180.

Sinayev, A. (2016). *Dual-System Theories of Decision Making: Analytic Approaches and Empirical Tests*. (Electronic Thesis or Dissertation). Retrieved from https://etd.ohiolink.edu/

Thaler, R.H. (1985). Mental accounting and consumer choice. *Marketing Science*, 4(3), 199-214.

Thaler, R. H., & Shefrin. H.M. (1981). An economic theory of self-control. *The Journal of Political Economy*, 89(2), 392-406.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.

Vromen, J. (2011). Neuroeconomics: two camps gradually converging: what can economics gain from it?. *International Review of Economics*, 58(3), 267-285.

Chapter 5

Why behavioral economics needs to revise its faith in dual process theories

1. Introduction

Dual process theory has been playing a prominent role in both the cognitive and behavioral sciences. Dual process theory (DPT) posits a duality between two kinds of mental processes. "Higher" mental processes are often associated with energy-intensive cognitive tasks like deductive and hypothetical reasoning. "Lower" mental processes are associated with perceptual and affective operations, like attentional cueing and motor-response preparation. The standard view of DPT —also known as the "received view" (Evans & Stanovich, 2013b; cf. Mugg 2015)—contends that different aspects of human cognition, such as critical thinking and decision-making, can be categorized according to and/or understood as the result of these two mental processing types. The higher mental processes are depicted as *slow*, *controlled*, *reflective*, *serial*, *rule-based*, *effortful*, and *conscious*—this category is commonly referred to as "System 2" or "Type 2" processing. The lower mental processes are depicted as *fast*, *reactive*, *automatic*, *intuitive*, *heuristic*, *associative*, and *unconscious* (or *preconscious*)—this category is commonly referred to as "System 1" or "Type 1" processing.[1]

Over the last few decades, behavioral economists have sought empirical support for their models by appealing to psychology and neuroscience, often employing the concepts and rhetoric of DPT. The first, and perhaps best-known example of this is the study of judgment and decision-making under risk and uncertainty in the Heuristics and Biases tradition (Tversky & Kahneman, 1973, 1974; Kahneman, Tversky, and Slovic, 1982; Kahneman & Frederick, 2005). The second is the development of intrapersonal and intertemporal choice models in relation to choice consistency and self-control (Loewenstein, 1996, 2000; Benabou & Tirole, 2002; Bernheim & Rangel, 2004; Benhabib & Bisin, 2005; Loewenstein & O'Donoghue, 2005; Fudenberg & Levine, 2006). Unlike neoclassical decision models that presume agents to be faultless utility-maximizers and which often resort to *ad hoc* explanations to justify deviations from expected utility theory, dualistic models influenced by DPT have provided

[1] For classic texts on dual processing, see Shiffrin & Schneider (1977), Evans (1989), Epstein (1994), Sloman (1996), Stanovich & West (2000), Lieberman (2003). For recent developments and modifications, see Stanovich (1999, 2004, 2009, 2011), Evans (2006, 2008, 2009, 2011, 2012), Evans & Frankish (2009), Evans & Stanovich (2013a, 2013b), and De Neys (2017).

behavioral economists with a novel and seemingly realistic way to interpret choice-behavior.

DPT has come under intense scrutiny during the last two decades. Cognitive scientists and philosophical psychologists alike have criticized the theoretical foundations of the standard view of DPT and have argued against the validity and relevance of evidence used to support it (Gigerenzer & Reiger, 1996; Osman, 2004; Keren & Schul, 2009; Kruglanzki & Gigerenzer, 2011). Moreover, recent modifications of DPT in light of these criticisms have generated additional concerns regarding DPT's applicability and irrefutability (Keren, 2013; Mugg, 2015; Pennycook, 2017; Bonnefon, 2018). This should raise concerns for behavioral economists who see DPT as providing psychologically realistic foundations for their models. In particular, it raises the possibility that dualistic models are not as descriptively accurate or reliable as behavioral economists presume them to be. In fact, the case can be made that the popularity of DPT in behavioral economics has less to do with the empirical success of dualistic models, and more to do with the convenience that the dualistic narrative provides economists looking to sort out decision anomalies (Grüne-Yanoff, 2017; cf. Angner & Loewenstein, 2012). I will argue that the growing number of criticisms of DPT leaves behavioral economists with something of a dilemma: either they stick to their purported ambitions to give a realistic description of human decision-making and modify their use of DPT, or they stick to DPT and modify their ambitions. To illustrate this dilemma, this paper raises two challenges:

The first challenge pertains to how dualistic models represent choice as the outcome of dual processes and/or systems and the tensions that arise therein. This challenge is twofold: (1) Some dualistic models appeal to neuroscientific evidence to determine how different parts of the brain evaluate prospects; this sometimes involves positing specialized mechanisms or sub-systems to further partition the decision process. However, because DPT is not *per se* a mechanistic theory, dualistic models which appeal to neuroscientific evidence may exaggerate or distort the roles that mechanisms or sub-systems play in the execution of decisions. On the one hand, this may lead to confusion about what, exactly, DPT is meant to represent; on the other hand, it may render DPT redundant. (2) However, the majority of dualistic models don't rely on neuroscientific evidence to represent the formation of choice; rather, in most cases, DPT serves as a functional framework which provides approximate descriptions of the mental processes that underpin decisions. While this is consistent with the standard view of DPT, it also means that functional dualistic models cannot explain how choice emerges from the interaction of dual processes and/or systems without a supplemental story. This leaves behavioral economists with two options: (i) bite the bullet and leave the black-box closed; (ii) use formal constructs from economics to work out the details of how dual processes and/or systems interact. The problem

with the latter option is that truth of dualistic models is hostage to the truth of DPT in general.

The second challenge addresses a more fundamental question with regard to DPT, namely, how well does it capture humans' rational faculties? Despite the belief that System 2 (Type 2) is necessary for, or synonymous with, rational action and System 1 (Type 1) is non-rational, there are both theoretical and empirical considerations to suggest that this dichotomy is not reliable, and perhaps in need of revision. Hence, behavioral economists are overly optimistic if they think that DPT, construed as a functional framework, provides dualistic models with the normative foundations they seek. While this does not warrant abandoning a dual process view *in toto*, it certainly calls into question whether in vogue dualistic models can reliably predict and/or explain reasoning errors outside the lab, since what is conducive to triggering System 1 (Type 1) or System 2 (Type 2) may depend entirely on the choice setting.

Lastly, it should be made clear that the aim of this paper is not to disparage behavioral economists for their attempts to develop more realistic models. Quite the opposite: the aim is to address open questions which concern DPT as a psychological theory and, to that end, to better understand its limitations in the behavioral sciences.

This paper is structured as follows: Section 2 provides an overview of the theoretical and empirical criticisms of DPT from the perspective of cognitive science and philosophical psychology: it discusses the main differences between system-based and type-based theories of information processing. Section 3 surveys the explicit and implicit influences of DPT upon behavioral economics and differentiates two styles of dualistic modeling. Section 4 considers the tensions for each of the two styles of dualistic models and discusses possible limitations of each with regard to DPT. Section 5 then considers whether DPT can sustain the normative foundations of dualistic models given that neither system-based nor type-based interpretations can cleanly distinguish rational from irrational choice. Section 6 concludes.

2. Recent developments in dual-process research

2.1 *Taking a closer look at System 1 and System 2*

Nearly all versions of DPT subscribe to the same basic idea, which is that human minds rely on distinct types of mental processing to accomplish different tasks in daily life.[2] It's widely believed that these processes evolved for specific purposes and are designed and attuned to respond to features of the external environment. As previously mentioned, the standard view distinguishes mental processes that are fast, reactive, automatic, intuitive, heuristic, associative, and preconscious from mental processes

[2] For overviews of DPT's applications and interpretations across psychology See Evans (2006, p. 208) and Pennycook (2017).

that are slow, controlled, reflective, serial, rule-based, and conscious. Mugg (2015) refers to this as the "standard menu" of mental processes; I will refer to the standard menu throughout this paper.

On the standard menu, processes come in two types. As previously mentioned, the former are believed to be evolutionarily old and directly linked to autonomic functions, such as 'fight or flight' responses and stimulus-bound perceptions. The latter set of processes are believed to have evolved more recently and aid in higher cognitive functionings that draw upon working memory and require sustained effort and attention (Evans & Over, 1996; Stanovich, 2004). Classic experiments, such as the Wason Selection Task (Wason & Evans, 1975; Evans, 1989) and Stroop Test text (Stroop 1935; Osman 2004), demonstrate how reasoning errors and biases depend largely on the allocation of cognitive resources, which is determined by the automaticity of information processing protocols (this is believed to be a sign of the strength or proximity of neural pathways). Certain processes—typically those associated with older evolutionary structures—are easily primed and often trigger responses before an individual can, say, consult a rule or deliberate about a problem. In the domain of reasoning and decision-making, the effects of automatic and rapid processing can be observed through misapplied decision heuristics and faulty reasoning, as well as computational errors. These experiments are believed to give credence to DPT.

If this sounds somewhat vague, however, it's because DPT *is* vague. The constellation of theories that make up DPT are better thought of as a generic framework than a unified theory (Evans & Stanovich, 2013). Hence, a well-known defect of the standard view is that it's not obvious what distinguishes mental processes from one another, aside from the labels the theory ascribes to their functional roles. Moreover, it's difficult to know whether token theories utilizing the standard menu of mental processes refer to the same thing. (This is also a problem for functional explanations in psychology—Levin, 2017). Consequently, the standard view of DPT does not actually provide an account of how reasoning tasks are accomplished, and decisions made; what it provides is a generic theory about the potential origins of reasoning and decision errors. This, it would seem, is a major deficiency for the theory: if it cannot explain how the mind inhibits or overrides bad judgments that are generated by rapid or automatic mental processes, then what is the point of making the distinction to begin with? After all, we don't always submit to our biases—we are often able to restrain gut-reactions and to recognize hasty errors for what they are. But the fact that we frequently make reasoning errors under conditions of risk and uncertainty indicates that much of our mental processing is not under conscious control.

Theorists have responded to this issue by positing separate modes of processing, paring the standard menu into discrete systems. These are most commonly known as System 1 and System 2 (Stanovich, 1999; Kahneman, 2003a, 2003b, 2011; Kahneman & Frederick, 2005); though, some authors have opted for less neutral terminology,

referring to them as the Heuristic system and Analytic system, respectively (Stanovich 2004; Evans, 2006; cf. Stanovich & Evans, 2013a). For an extensive overview of the clusters of attributes that are said to belong to System 1 and System 2, see Evans (2006, 2008).

While the idea of separate cognitive systems has helped to synthesize many token theories in the DPT literature, alleviating some of the worry about terminological discrepancies, perhaps the most important—and arguably the most controversial—aspect of the System 1 / System 2 framework is the way the two systems are believed to *interact*. One interpretation is that the systems are arranged sequentially: System 1 operates autonomously, with System 2 monitoring and intervening when it has the power (i.e. resources) to do so. This is known as the "default-interventionist" model. Another interpretation is that the two systems are arranged in parallel and must compete for control over our behavior. This "parallel-competitive" model is appealing given new evidence about the distribution of brain processes (Sinayev, 2016; Lurquin & Miyake, 2017; see also Pennycook, 2017). Yet, the consensus among many researchers, at least in the areas of reasoning and decision-making, is that this latter view is untenable (cf. Evans, 2008; Keren & Schul, 2009).[3]

The reason why many seem to resist the parallel-competitive model of system interaction is because it requires a more complex explanation about how System 1 and System 2 cooperate and reconcile conflict. For proponents of the default-interventionist model, there is no conflict *per se*; System 1 operates autonomously and System 2 either intervenes or it doesn't. But on the parallel-competitive model, both systems are thought to generate responses to input, and although System 2 can overrule System 1, the associative force of System 1's responses my counter System 2's attempts to intervene. While it is still a debated issue which model better approximates the interaction of System 1 and System 2, proponents of both agree that System 2 could not operate without System 1 because the higher cognitive functionings of System 2 depend on information received by System 1 (cf. Evans & Stanovich, 2013a).[4] However,

[3] Arguably, one could make the case for a third arrangement, wherein systems process information simultaneously (in parallel), but they are allowed to influence each other in complex, feedback interactions (Sinayev, 2016). For the sake of space, I will not consider this interpretation of system interaction here. However, for reasons that will become evident in section 4, I suspect that this is not conducive to behavioral economic modeling.

[4] It's important to keep in mind that "parallel" has different interpretations and can range over different operations within a system. In this instance, parallel is meant to encompass all operations of System 1 and System 2, respectively. Systems organized in parallel are designed to respond to unique inputs and do not overlap or share functional characteristics with one another—in a word, they operate autonomously. This can be contrasted from other instances of "parallel and distributed" processing which occur at the intra-system level. For instance, some who endorse the default-interventionist model readily acknowledge that within System 1 there exist many autonomous sub-systems which operate in parallel (isolated from one another, but

some critics have speculated that if System 1 and System 2 operated independently of one another (which the parallel-competitive model suggests), then it seems the only way they could meaningfully interact and compete for control over our behavior is by way of a third system, which has access to the inputs and processes of both System 1 and System 2. Indeed, there is growing neuroscientific support for the existence of executive control functions in the brain, and some supporters of the system-based interpretation ascribe this function to System 2. But the range of processes that this executive function has control over appears to be limited (cf. Pennycook, 2017); moreover, if executive control were a feature of System 2, this would indicate that System 1 and System 2 are not isolated from another (otherwise System 2 couldn't perform its role as executor—Keren & Schul, 2009). For this reason, the default-interventionist model is, at least in the domain of decision-making, the more plausible of the two models of system interaction. Yet, despite the proliferation of DPTs that use the terminology of System 1 and System 2, there are several outstanding criticisms of the systems-based interpretation—many of which have not received due attention outside the cognitive sciences. Consider the three criticisms:

Criticism 1: Systems are not distinct / discrete. It is reasonable to think that System 1 and System 2 roughly correspond to neuroanatomical differences in the brain; and, it's been suggested by many (with the support of brain imaging software) that some functions of System 1 and System 2 can be correlated with domain-specific modules and/or neural circuits (Mars et al., 2011; Botvinick & Cohen, 2014; Botvinick & Braver, 2015). However, there is not sufficient evidence to warrant *identifying* System 1 and/or System 2 with any fixed neural architecture (Osman, 2004; Keren, 2013). Rather, evidence indicates that many processes associated with both systems "crosscut" each other (Mugg, 2015: cf. Keren & Schul, 2009; Evans & Stanovich, 2013a). This has two important consequences for DPT: firstly, it indicates that System 1 and System 2 are not discrete—in fact, they may share or utilize the similar neural pathways for the completion of certain tasks. This is not so surprising when one considers that the standard menu of processes is characterized according to the *functions* of System 1 and System 2, and these functions may be multiply realized depending on the task at hand or the circumstances surrounding a task. Secondly, as Keren & Schul (2009) point out, the contrastive nature of System 1 and System 2 really is a matter of degree, as mental processing occurs along a continuum (e.g., the dividing line between "controlled" and "automatic" processing, as for many mental processes, is fuzzy and indistinct).

Criticism 2: Inter-system interactions are underdetermined by evidence. The issue as to whether systems are arranged in a sequential or parallel fashion is very much a contingent matter: it depends entirely on how one defines the concept of a cognitive

not from higher reflective process). This is the basis of Stanovich's concept of The Autonomous Set of Systems (TASS) which comprises the Heuristic System (Stanovich, 2004, 2011).

system and how this is fleshed out in terms of its functional characteristics. Because it is not agreed upon what the appropriate neuroanatomical correlates of System 1 and System 2 are, the story of their interaction is mired in theoretical and terminological disputes. Nevertheless, although most researchers believe System 1 and System 2 to be arranged sequentially, another possible criticism is that there isn't sufficient empirical evidence to validate either the default-interventionist model or the parallel-competitive model of system interaction. Recent meta-analyses and replication studies indicate that neither model is singularly equipped to predict and explain how individuals' reason and make decisions (Sinayev, 2016; Lurquin & Miyake, 2017; cf. Pennycook, 2017). The evidence and counter-evidence to support both models could be interpreted as a fundamental flaw in the theory itself.

Criticism 3: Evidence for dual systems is confined to the laboratory. Finally, there is growing consensus among critics that system-based interpretations of DPT is predictive only insofar as it predicts behavior in highly controlled, laboratory settings (Keren, 2013; Buturovic & Tasic, 2015). On the one hand, it has not been proven that either system is solely responsible for reasoning errors; on the other hand, the case has been repeatedly made that proponents of the system-based interpretation (viz. Kahneman and Frederick) presuppose norms of rationality that focus solely on rules of logic and statistical prowess. This emphasis on testing peoples' abilities to solve puzzles and perform tasks in artificial conditions says little about their day-to-day reasoning abilities (Gigerenzer & Reiger, 1996; Gigerenzer & Brighton, 2009; Kruglanski & Gigerenzer, 2011). It has been further argued that the system-based interpretations of DPT relies on biased results, that experimenters are selective in their reporting of evidence (Gigerenzer, 2015). It goes without saying that the above criticisms have generated much controversy.[5] As I argue in section 4, however, each of these criticisms plays a part in my argument that not all irrational decisions are attributable to System 1 (and, conversely, that System 2 does not always produce rational decisions).

2.2 *Why the Type 1 / Type 2 distinction doesn't escape criticism*

It could be argued that the above criticisms, while valid, do not undermine the theoretical significance of DPT; rather, they merely demonstrate the limitations of particular models and particular applications of it. For instance, Evans & Stanovich (2013a, 2013b) now acknowledge that the systems-based interpretation of DPT has many deficiencies. Nevertheless, they maintain that such criticisms also betray a confusion by critics between theory and meta-theory, and they maintain that DPT—construed as a

[5] Although Evans and Stanovich have elected not to use the terms "System 1" and "System 2" to characterize DPT, they acknowledge that most of the criticisms presented above apply to their own conception of Heuristic System and Analytic System.

meta-theory—has not been, or rather, cannot be refuted (Evans & Stanovich 2013b; Pennycook, 2017). What Evans & Stanovich mean by "meta-theory" is not altogether clear. They claim that, "Broad frameworks, like dual-process theory, have a very important role to play in psychology, and there are numerous examples of research programs organized within and around such frameworks… What we can expect at this level is general principles, coherence, plausibility, and the potential to generate more specific models and the experiments to test them" (2013b, p. 263). As such, Evans and Stanovich have since abandoned the system-based interpretation, arguing instead that DPT is most plausible if mental processes are organized by a single dichotomy, namely, their type; hence they adopt the terminology Type 1 and Type 2 to distinguish the processes which they formerly attributed to the Heuristic System and Analytic System, respectively.

While proponents of this new type-based interpretation are cautious not to overstate the discreteness of Type 1 and Type 2 processes, they utilize many of the same attributes to differentiate the two: Type 1 consists of autonomous processes that are automatic and not under an individual's conscious control, whereas Type 2 consists of reflective processes that are typically associated with activities like language use, mental simulation, and complex problem solving. However, what really sets Type 1 and Type 2 apart from System 1 and System 2 is the role of working memory in higher-order functionings (Evans & Stanovich, 2013a).

Stanovich (2009) further modifies the type-based interpretation, positing that in addition to the set of autonomous processes (or TASS) that make up Type 1, Type 2 processes can be further bifurcated into distinct stages: the first stage involves what he calls "algorithmic" processing, which initiates many of the monitoring and executive functions that are associated with the Analytic System. It is only after algorithmic processing that the second stage of Type 2 is engaged, where genuine "reflective" processing takes place (Stanovich, 2011). The algorithmic stage of mental processing is an important innovation in this model, as it is intended to mediate between the autonomous processes of TASS while effectively priming information for conscious manipulation. The significance of positing an algorithmic "level" is that it is thought to account for discrepancies in the application of DPT, such as individual differences in intelligence and cognitive ability. Nevertheless, there remain a number of problems for this type-based interpretation of DPT. Consider three more criticisms:

Criticism 4: Types do not distinguish mental processes. The restructuring of DPT based on processing types was largely intended to solve the cross-cutting problem by using the continuity of mental processes to its advantage. This works insofar as it sidesteps the issue of having to carefully demarcate separate systems; but it essentially pushes the problem back a level and does not provide a solution to the ambiguity surrounding mental processes (Keren, 2013). While proponents like Evans and Stanovich may argue that working memory is a sufficient criterion to distinguish

autonomous processes (TASS) from non-autonomous ones, this does little to improve understanding of the putative menu of processes which make up Type 2 functionings. This "stripped down" version of DPT makes the overall framework less precise, which makes one wonder whether it is not simply a theory about working memory instead of a theory about reasoning and decision-making (Keren, 2013; Mugg, 2016).

Criticism 5: The criterion of "rule-based" reasoning is obscured. Both system-based and type-based interpretations of DPT appeal to processes that are "rule-based". As argued by Kruglanski & Gigerenzer (2009), there seems to be much equivocation in the use of the term "rule-based" as a criterion to distinguish types of processes. On the one hand, "rule-based" could refer to the conscious effort of an individual to adhere to rules (e.g., rules of normative conduct, rules of a game, rules of arithmetic); but, on the other hand, "rule-based" could refer to unconscious "computational" processes that aid in or underwrite cognitive tasks. For some, namely those interested explicitly in the psychology of deductive reasoning, this equivocation may not be much a problem, the property "rule-based" refers typically to higher-order capacities to reason abstractly and perform mental simulations. However, for those interested in the mental processes that support the learning of implicit skills and other preconditions for reasoning and decision-making, this can get confusing very quickly, as it's not obvious whether the criterion refers to a conscious ability of the individual to reason analytically, or whether it refers to the ability to model some aspect of cognition according to rules.[6] Evans & Stanovich (2013) are not convinced this is a major issue, but the criticism (6) indicates further why it may turn out to be a problem.

Criticism 6: What does algorithmic processing refer to? The idea of algorithmic processing was introduced to alleviate confusions about where and how Type 2 processes are initiated; however, Stanovich (2009, 2011) has argued that this innovation has been very useful for explaining how implicit skills are developed and for accounting for individual differences in cognitive ability and intelligence. The problem is that it's anything but clear how algorithmic processes are realized, and how they differ—at the neuroanatomical level—from other Type 2 processes, if they do at all. Stanovich uses primarily functional terminology to portray the rule-based nature of algorithmic processes; though there is little indication that it relies on or can be attributed to any specific mechanisms or neural pathways. As such, this doesn't alleviate the problem of mental processes cross-cutting each other, nor does it clarify what it means to

[6] This discrepancy about what it means to describes mental processes as 'rule-based' runs parallel to debates in philosophy of cognitive science concerning the meaning and interpretation of "computation" in computational theories of cognition (van Gelder, 1995; Thompson, 2007; Hutto et al., 2018*;* cf.; Piccinini, 2013; Piccinini & Bahar, 2015). The bone of contention is whether, or on what grounds, it makes sense to say that the mind "computes" information, i.e. at the level of representations or at the level of neurophysiology. Skeptics of computational theories of mind reject claims that human thinking is "computational" because there is, as of yet, evidence that anything akin to symbol-manipulation happens when thought is produced.

describe some mental processes as rule-based. We are told that even though individuals are not conscious of algorithmic processes, they are still considered Type 2 processes because they depend on working memory and are representational in nature.

In sum, these additional criticisms suggest that type-based interpretations DPT may obscure rather than clarify the idea that human reasoning and decision making is *inherently* dualistic. Part of the reason for this is that many of the same reasoning processes can be also described by a 'one-system' model (Osman, 2004; Kruglanski & Gigerenzer, 2011). This can be seen as a further justification for the claim that rational decision-making cannot be reduced to the operations of single system, or in this instance, a single type: Type 2 processes do not guarantee rational decision-making, and likewise, Type 1 processes do not necessarily produce irrational decisions.

3. How has dual process theory influenced behavioral economics?

Behavioral economics has earned a reputation for being psychologically realistic and for providing new insights into the hidden processes that govern individual decision-making. This is due in large part to mass market publications, such as Dan Ariely's *Predictably Irrational* (2008), Richard Thaler's *Nudge* (co-authored with Cass Sunstein, 2008), and Daniel Kahneman's *Thinking, Fast & Slow* (2011), which portray behavioral economics as an exciting new discipline that has the potential to unlock the mysteries of the human mind. This image has been further reinforced by Kahneman and Thaler receiving Nobel Prizes for their contributions to economics. This reputation has, in part, strengthened economists' faith in the truth or factivity of DPT. In this section I make the case that at least two research programs in behavioral economics have been inspired or influenced by DPT (or core features of it). The first, and perhaps best well-known, is the study of judgment and decision-making under risk and uncertainty in the Heuristics and Biases tradition. The second is the development of intrapersonal and intertemporal choice models for the study of self-control.[7] The question this section addresses is how DPT, and related forms of psychological dualism, have influenced behavioral economics.

3.1 *Explicit and implicit examples of psychological dualism in behavioral economics*

The heuristics and biases research program is exemplary for understanding how DPT has permeated behavioral economics and for understanding how other research

[7] While Kahneman is perhaps the most vocal proponent of DPT, Thaler has also contributed to the popularity of DPT—see Thaler & Sunstein (2003), Thaler, Sunstein, and Balz (2012), Kahneman & Thaler (2006). Hence, it could be argued that research into libertarian paternalism and nudge policy constitute a third branch of behavioral economics that is heavily influenced by DPT (cf. Camerer et al., 2003; Loewenstein & Haisley, 2008; Heilmann, 2014).

programs have come to rely on similar forms of psychological dualism. In order to appreciate how the heuristics and biases program influenced behavioral economics, one needs to make a careful distinction between its economic and psychological applications.

Concerning its psychological applications, it is well understood that one of the major threads in DPT, application to judgment and decision-making under risk and uncertainty, has its origins in the behavioral decision research of Tversky & Kahneman (1973, 1974) and Kahneman, Tversky, and Slovic (1982)—see, e.g., Sent (2004) and Heukelom (2012) for detailed overviews. The goal of this research was, primarily, to understand why individuals tend to make mistakes when forming probabilistic judgments and how to predict when cognitive load may compromise one's reasoning ability. The principal discovery in this research is that individuals often resort to shortcuts and other rules of thumb to facilitate decision-making; sometimes these shortcuts are helpful, but often they can be biased in ways that undermine one's reflective or computational abilities. Original studies posited three primary heuristics—*accessibility*, *representativeness*, and *anchoring* (Tversky & Kahneman, 1973; Kahneman, Slovic, and Tversky, 1982)—as the likely cause of reasoning errors; and it was clear that Tversky & Kahneman saw these heuristics as the result of differential mental processing, consistent with research in social and cognitive psychology of the 1970's (cf. Shiffrin, & Schneider, 1977; Nisbet & Ross, 1980). Subsequent research on psychological applications of heuristic and biases research have reinforced the idea that human decision-making is involves the interplay of two types of mental processes (Sloman, 1996; Chaiken & Trope, 1999; Gilbert, 1999, 2002).

The most vivid demonstrations of the mind's two systems are explicated in Kahneman & Fredrick (2002, 2005, 2007) and Kahneman (2003a, 2003b, 2011) who adopt the terminology (from Stanovich and West, 2000) of "System 1" and "System 2". For all intents and purposes, this understanding of System 1 and System 2 bears all the core characteristics of other system-based interpretations of DPT. For instance, Kahneman & Frederick (2002, 2005) claim that System 1 corresponds to "intuitions", which are directly informed by the perceptual system, whereas System 2 corresponds to "reflective judgments", which cautiously assess intuitions and formulate responses to them. Kahneman (2011) further elaborates the various roles that System 1 and System 2 play in both economic and non-economic choice-settings.[8]

However, not all dualistic models in behavioral economics are as explicit about (or aware of) the influence of DPT. For instance, psychological dualism is a recurring theme in the study of intrapersonal and intertemporal choice. Common afflictions,

[8] Kahneman has since remarked that the cognitive features of prospect theory correspond to functions of System 1 and System 2, viz. that an individuals' reference point is said to be their default valuation position, which is contingent upon their interpretation of the choice context. However, there is no mention of DPT in either Tversky & Kahneman (1992) or Wakker (2010).

such as temptation, procrastination, myopia and addiction may give rise to imprudent and self-defeating behaviors which aren't well represented with the standard tools of utility theory. Yet, there is plenty of evidence to suggest that the core features of DPT have carried over to the realm of intrapersonal and intertemporal choice. The table below (see figure 1) provides a list of dualistic nomenclature that has been used to characterize the structures that give rise to internal conflict and produce decision anomalies.

Dualistic label	Reference
"*Cold"* vs. *"hot" modes*……	Loewenstein, 1996, 1999; Benabou & Tirole, 2002; Bernheim & Rangel, 2004; 2007; Ainslie, 2001, 2005; Soman, 2005
Cognition vs. *emotion*……..	Sanfey et al., 2003
Deliberative system vs. *affective system* ……	Loewenstein & O'Donoghue, 2005
Cognitive system vs. *affective system*……	Camerer, Loewenstein, & Prelec, 2005
Reflective vs. *impulsive system*……	McClure et al., 2004, 2007; Strack & Deutsch, 2006; Fudenberg & Levine, 2006;
Controlled vs. *automatic processing*……..	Loewenstein, 1996, 1999; Camerer, Loewenstein, & Prelec, 2005; Benhabib & Bisin, 2005
Executive control vs. *conflict monitoring*…..	Benhabib & Bisin, 2005; Brocas & Carrillo, 2008; 2014

Figure 1. – *list of dualistic models that do not use terminology of DPT*

One will notice that the references cited in the table cover a wide range of theoretical approaches and modeling techniques, from mathematical psychology to neuroeconomics. What these examples all have in common is that they (i) employ a dualistic rhetoric to characterize the competing interests or motivations or physiological processes that are believed to generate tension within the individual decision maker, and (ii) that they codify the competing interests, motivations, and physiological processes through some model of constrained optimization (e.g., principal-agent model or limited information game).

There is, to be sure, a continuity which links recent intrapersonal and intertemporal choice models with prototypical "planner–doer" and "multiple-self" models

(Thaler and Shefrin, 1981; Schelling, 1984; Shefrin & Thaler, 1988; Ainslie, 1992; Ainslie & Haslam, 1992). The particular arrangement of the planner-doer model indicates that the planner self is not just farsighted, but that it has rational authority, which is realized when it exerts control over the doer self. This is a clear indication that economists have sought to resolve the problem of dynamic inconsistency by invoking asymmetrical reasoning processes, which may supervene on real psychological or physiological processes. So, while the case can be made that traditional planner-doer and multiple-self models presuppose some kind of psychological dualism, the behavioral economic models listed in the table above place more emphasis on the cognitive and neurological basis of willpower and self-control.[9]

3.2 *The increasing popularity of dualistic models in economics*

How do we reconcile the increasing popularity of dualistic models in behavioral economics with the controversies surrounding DPT? One answer is that economists simply aren't familiar with the debates in cognitive science and philosophical psychology about DPT, and thus shouldn't be expected to know each and every criticism against it (especially if those criticisms are still up for debate). However, it could also be the case that behavioral economists think these criticisms do not apply to them, perhaps because their descriptive aims are not so tied up in the truth or factivity of DPT. As we saw above, while some dualistic models are explicit about their relationship with DPT (such as in the heuristics and biases tradition), many dualistic models are not explicit about their relationship to DPT (either because they don't recognize the historical connections, or because they adopt alternative nomenclature for dual processes and systems). For this reason, it's difficult to know whether the controversies and criticisms of DPT carry over and have implications for behavioral economics.

[9] One explanation for why DPT has remained implicit in the recent development of intrapersonal and intertemporal choice models may have something to do with how the literature on time preferences and choice inconsistency has developed. For example, Loewenstein (1996, 2000) and Metcalfe & Mischel (1999) are among the most frequently cited in the behavioral economics literature with regard to the psychological underpinnings of inconsistent preferences. Loewenstein and Metcalfe & Mischel both emphasize the significance of visceral factors (e.g., cravings, sexual arousal, pain) and heightened emotional states (e.g., "hot" states) as responsible for impulsive or unreflective choices. Thus, the hot-cold heuristic has percolated throughout behavioral economics and has been adopted by many as the default psychological model for the study of self-control problems. In this way, the hot-cold heuristic serves nearly the same function that DPT does—in fact, the only discernable difference between the hot-cold heuristic and DPT is the disciplines from which they emerge: The hot-cold heuristic is more closely associated with the psychometrics of willpower and delay-gratification, which has close links to utility theory as it is employed in mathematical psychology. But the hot-cold heuristic is no more enlightening with regard to its descriptive accuracy or explanatory power. Rather, the case could be made that DPT encompasses the hot and cold states; hence, if DPT is descriptively inadequate, then the hot-cold heuristic is by entailment.

To answer this question, we need say a bit more about the possible roles that DPT plays in dualistic models.

Some dualistic models make use of neuroscientific evidence and appeal to executive mechanisms and/or specialized sub-systems in the brain; these mechanisms and sub-systems are believed to play a pivotal role in the execution of decisions and thus are thought to be highly relevant to the analysis of rational choice. In these cases, DPT not only provides empirical support for dualistic modeling, but it possibly opens the proverbial "black box" by pin-pointing physiological structures that have been left out of economic analysis in the past (Camerer, Loewenstein, & Prelec, 2005; Camerer, 2007; cf. Bernheim, 2009). Yet, the proportion of dualistic models which rely on neuroscientific evidence is small. By contrast, the majority of dualistic models in behavioral economics can be understood as taking a functional (non-physiological) approach to the analysis of decisions. In these cases, DPT could be thought to provide a menu of mental processing types, which would allow behavioral economists to make broader inferences about the etiology of decision phenomena.

So where does this leave us, or rather, behavioral economists? As I stated in the introduction, the controversies and criticisms of DPT give rise to a dilemma of sorts, one which requires either a revision of DPT or a revision of behavioral economists' descriptive aims. But we've now seen there are at least two descriptive roles that DPT could play with regard to dualistic modeling: one role concentrates on the physiological aspects of decision-making; the other role focuses on the functional (non-physiological) characteristics of decision-making. And this is a rather crude generalization; many dualistic models above seem to want to take advantage of both descriptive roles. With this in mind, the next section will be dedicated to showing that, in fact, both roles are susceptible to criticisms raised in section 2, and this calls into question the descriptive aims of behavioral economists who invoke DPT.

4. Two styles of dualistic modeling in behavioral economics

In this section, I investigate the use of DPT in dualistic models on two fronts: First, I question how behavioral economists utilize neuroscientific evidence for the purpose of dualistic modeling. I will argue that, because the standard view of DPT is not bound to any particular neural structures, dualistic models which appeal to neuroscientific evidence may exaggerate or distort the role that specialized mechanisms or sub-systems play in the execution of decisions—this puts into perspective and exemplifies criticisms 1 - 3 raised in section 2. Hence, *neuroscience-based dualistic models* run the risk of confusing what, exactly, DPT represents, possibly rending it redundant. Second, I question the use of DPT construed as a functional framework. I argue that *functional dualistic models* cannot explain how choice emerges from the interaction of dual processes and/or systems without a supplemental story. This leaves behavioral

economists with two options: (i) bite the bullet and leave the black-box closed; (ii) use formal constructs from economics (principal-agent formalism, bargaining and information games) to work out the details of how dual processes and/or systems interact. The problem with the latter option is that truth of dualistic models is hostage to the truth of DPT in general.

4.1 *Neuroscientific evidence in dualistic models*

Camerer, Loewenstein, & Prelec (2005) (CLP) have championed the use of psychology and neuroscience to improve standard economic theory. They emphasize that standard economic theory is inadequate because it is unable to account for decision anomalies that result from "automatic" and "emotional" processing which governs an extensive portion of human behavior. "Human behavior," they argue, "requires a fluid interaction between controlled and automatic process, and between cognitive and affective systems. However, many behaviors that emerge from this interplay are routinely and falsely interpreted as being the product of cognitive deliberation alone" (CLP, 2005, p. 11). It's important to realize that, for CLP, the contrastive functions of automatic versus controlled processing (and the emotional versus cognitive systems which underlie this processing) aren't merely incidental to understanding how individuals make decisions; they are presumed to be neuroscientific fact. Interestingly, one finds many references to core texts from the DPT literature in CLPs "two-dimensional" cognitive framework, e.g., Schneider & Shiffrin (1977), Sloman (1996), Kirkpatrick & Epstein (1992), Lieberman et al (2002), Gollwitzer, Fujita, and Oettingen (2004), Kahneman & Frederick (2002), to name a few.

Unlike other dualistic models in behavioral economics which merely pay lip-service to neuroscience (Bernheim & Rangel, 2004; 2007; Benhabib & Bisin, 2004; Loewenstein & O'Donoghue, 2005; Fudenberg & Levine, 2006), CLP attempt to derive criteria for rational choice from the interactions of neural systems, as do others—see, e.g., McClure et al., 2004, 2007; Brocas & Carrillo, 2008, 2014. Where the latter neuroeconomists are explicit about their focus on eliciting preferences from the deliberative processes of neural systems, the former are not.[10] An outcome of this focus on the neural basis of rational choice is that dualistic models have placed greater emphasis on what are sometimes referred to as "executive control" and "conflict monitoring" systems. These specialized systems are taken to be *causally* responsible for operations that are associated with automatic versus controlled processing, such as the initiation of an "override" function to suppress automatic impulses. Consider the following passages:

[10] However, CLP understand the notion of "rational behavior" as indicating a particular sort or kind of processing, namely as deliberate and under conscious control. This is different from understanding "rational behavior" as behavior being consistent with a set of axioms.

> Controlled processes occur mainly in the front (orbital and prefrontal) parts of the brain. The prefrontal cortex (pFC) is sometimes called the "executive" region, because it draws inputs from almost all other regions, integrates them to form near and long-term goals, and plans actions that take these goals into account. (Camerer, Loewenstein, & Prelec, 2005, p. 17)

> We have observed that choices between lesser immediate and greater delayed rewards elicit activity in distinct neural systems that appear to favor different choice outcomes. That is, intertemporal choice under these conditions elicits decisional conflict. A growing body of evidence suggests that a dorsocaudal region of the ACC [dorsal anterior cingulate Cortex] responds to conflicts in processing… This is consistent with findings from the current study in which we observed activity in a similar area of the ACC that was greater for decisions involving choices between immediate and delayed rewards than for choices between only delayed rewards. Such findings have been taken as evidence for a conflict-monitoring function of ACC, which serves to detect conditions requiring the recruitment of cognitive control mechanisms subserved by prefrontal cortex and associated structures… (McClure et al., 2007, pp. 5803).

For CLP and McClure et al., the mechanisms in question not only provide a clearer picture of how the brain works, but they serve as a vehicle to track valuation procedures among separate neural systems. It stands to reason that they look to these mechanisms and sub-systems to mediate internal conflict, and therefore, to serve as the fulcra against which rational analysis can be set.

However, because CLP (2005) and McClure et al (2007) place such emphasis on executive control and conflict monitoring sub-systems within broader valuation, it's difficult to know how these function among and fit into their respective "two-dimensional" frameworks. Presumably the cognitive system (or reflective system) is where conscious and sustained deliberation takes place for the individual decision-maker; if executive control is what prevents automatic behavior from running wild, one would presume that the mechanism responsible for it would be housed in the same system, viz. the cognitive system.

This may sound like a pedantic issue, but it is illustrative of a tension that can be found throughout certain parts of neuroeconomics, namely, that the rhetoric of DPT is used to motivate a duality which slowly disappears from view. It leads one to wonder whether the cognitive-emotional duality of CLP's framework is really necessary to their purposes—after all, it is the conflict monitoring and executive control systems which are doing the work, not the individual who must contemplate having immediate or delayed rewards.

To take another example, Brocas & Carillo (2014) argue that the "central executive system" (CES) literally coordinates the tasks of separate sub-systems by governing the flow of information between regions of the brain (cf. D'Esposito et al., 1995; Szameitat et al., 2002). Like CLP and McClure et al., Brocas & Carrillo regard the

CES as imperative to rational functioning because it literally monitors and resolves conflict between neural systems. Yet, the CES is not under the individual's control, which is to say, it operates autonomously, on the basis of chemical signals. Like CLP, Brocas & Carrillo don't adopt the standard view of DPT, so it's difficult to locate (both conceptually and physiologically) the CES among the other cognitive or reflective systems which are typically associated with higher-order and deliberative thinking. These examples both illustrate an important conceptual gap between how DPT is conceived as a descriptive framework, and how neuroscientific evidence fits into that framework and fleshes out the gritty details (details which would otherwise remain in the black box). And this, as one will recall, is the main point of criticism 1.

Now, just because the standard view of DPT is not committed to a particular neural interpretation doesn't preclude one from speculating about the neural foundations of certain functions, such as cognitive control. But if one invokes specialized mechanisms or sub-systems then we ought to have some idea how these support DPT. CLP's and McClure et al's dualistic models are ambiguous with regard to what is *really* under investigation—that is, the cognitive and emotional systems? The controlled and automatic processes which comprise those systems? Or the executive control and conflict monitoring sub-systems which govern the controlled and automatic processes? There is clearly some sort of hierarchy here; but how it relates to and informs the economic analysis of rational choice is left unspecified. Rather, CLP say things like this:

> Neuroscience is shot through with familiar economic language—delegation, division of labor, constraint, coordination, executive function—but these concepts are not formalized in neuroscience as they are in economics. There is no overall theory of how the brain allocates resources that are essentially fixed (e.g., blood flow and attention). An "economic model of the brain" could help here. Simple economic concepts, like mechanisms for rationing under scarcity, and general versus partial equilibrium responses to shocks, could help neuroscientists understand how the entire brain interacts. (Camerer, Loewenstein, & Prelec, 2005, p. 56)

There is surely merit to this claim as neuroeconomics is developing into an independent sub-field of behavioral economics. But the question that arises is, if the logic of economic theory is suited (better suited?) to analyze decisions as they are realized in the brain, as CLP argue, then where—or at what point in the decision process—ought behavioral economists to hone their focus? At the penultimate moment of execution? Or at a more removed, interstitial point? This question is important not just for understanding how dualistic models envision the use of neuroscientific evidence, but also for understanding what rational choice consists in for dualistic models like CLP's.

It is useful to introduce a distinction here between two styles of neuroeconomics. The approach of CLP and McClure et al., which uses neuroscientific methods to characterize individuals' preferences, has been dubbed "behavioral economics in the

scanner" (BES) (Ross, 2008; Harrison & Ross, 2010). This approach can be contrasted with what has been alternatively called "economics of neural activity" (ENA) (Vromen, 2011) which interprets brain activity with econometric methods. It is the former style of neuroeconomics that we are here concerned using as this gives rise to questions about where rational choice is executed (when it is executed properly). According to proponents of ENA, the idea of partitioning the brain into spatially distinct valuation systems, as CLP and McClure et al do, is misguided—there is really one single valuation system (cf. Parker and Newsome, 1998; Schall and Thompson, 1999; Glimcher, 2003).

However, as Vromen (2011) suggests, the disagreement between BES and ENA over how many valuation systems there are can be reconciled by instead thinking "in terms of upstream and downstream phases in the total causal chain of decision-making in the brain" (2011, p. 278). Hence, it may after all be the case that there are multiple regions and circuits responsible for "upstream" valuation which fit CLP and McClure et al's dualistic picture of the decision-maker; but along the way, these valuations converge to a single phase or node, which requires an alternative picture of the execution processes, and hence, an alternative model of the mind / brain.

The message to take away here is not that dualistic models cannot interface with neuroscience to produce more realistic models of the decision-making process; but doing so runs the risk of distorting DPT. This speaks also to the significance of criticism 2, which suggests the interaction of systems is underdetermined by evidence. For CLP and McClure et al., the valuation systems that comprise the cognitive and emotional systems which in turn produce controlled and automatic processes are only part of the story, i.e. part of the total causal chain of decision-making. And, if they agree with proponents of ENA that the final phases of decision-making streamline to a single system, then they must acknowledge that DPT is only partly true, or perhaps, only useful for understanding *certain stages* of the decision process. The question this raises is why behavioral economists like CLP and McClure et al need DPT at all? While the contrastive features of controlled and automatic processing have intuitive appeal to behavioral economists looking to reform standard economic theory, the story CLP and McClure et al seem to want to push is that the brain is where the magic of decision-making happens. And the closer we zoom in, the more complicated things become. Hence, we reach a point of resolution at which DPT is too crude a framework to provide added descriptive value—if anything, it begins to pull in the opposite direction.

4.2 *Functional dualistic models*

The section above can be seen as an investigation of a limiting set of cases: it shows how DPT generates descriptive tensions when dualistic models focus on the brain.

But the case could also be made that these tensions only constitute a problem for neuroscience-based dualistic models, whereas the majority of dualistic models, in fact, don't rely so heavily on neuroscientific interpretation. What's more likely for the remaining dualistic models is that behavioral economists view DPT as a functional framework, one which gives approximate descriptions of decision procedures according to the standard menu of mental processes. The issue with these functional dualistic models is that they generate more, not less, confusion about the role that DPT plays with regard to their descriptive ambitions. There are two aspects to this.

First, how functional dualistic models understand the interaction of dual processes and/or systems is complicated by both criticism 1 and criticism 2. Recall that, in addition to the lack of neural evidence for demarcating dual systems, criticism 1 also raises the issue that mental processes operate on a continuum and are not always or even typically discrete—which is just to say, the dichotomy presumed by the standard menu of mental processes is (purposefully) generic (Evans & Stanovich, 2013a). Hence, this makes criticism 2 more potent, as it as it calls into question how, exactly, dual processes and/or systems interact when the functional interpretation leaves them, by definition, in a black-box. Some proponents have been willing to bite the bullet in regard to this challenge (e.g., Kahneman & Frederick, 2005).

However, some economists have attempted to overcome this issue by interpreting the interaction of dual processes and/or systems according to the rules of certain formal economic constructs and working out the interaction in terms of bargaining or constraining power—what determines whether one system or type of processes overrides another is contingent on "willpower" or "cognitive control", which is derived from the mathematics of the models (e.g., principal-agent model, limited-information game, etc.). Not surprisingly, the use of formal constructs to work out the internal dynamics of choice is found mostly in intrapersonal and intertemporal choice models—see, e.g., Benabou & Tirole (2002); Bernheim & Rangel (2004); Benhabib & Bisin (2005); Loewenstein & O'Donoghue (2005); Fudenberg & Levine (2006). The details of *how* DPT's functional characteristics *may* map onto these formal constructs is discussed in Grayot (*forthcoming*), so I will not repeat them here. Yet, the question raised by Grayot, which is applicable here, is whether these formal constructs possibly misconstrue how dual systems and/or types of produce choice. To understand how this relates back to criticism 2, consider again the two (most common) candidate theories for depicting system interaction, i.e. the parallel-competitive model and the default-interventionist model.

Among the functional dualistic models listed above, Loewenstein & O'Donoghue (2005) adopt a model of system interaction which is quite similar to the default-interventionist model. There model presumes that "effective system" is active by default, whereas the "deliberative system" requires energy resources to play the part of monitor and intervener. Now, why they adopt this model and not something closer to the

parallel-competitive model is unknown, since there is not much background discussion about it. One reason might be that the default-interventionist model presumes the rational authority of the deliberative system, whereas parallel-competitive models do not. In fact, Loewenstein & O'Donoghue (2005) state that their model relegates the impulses and motivations of the affective system to an exogenous variable, which is fixed independently of the valuations of the deliberative system (2005, p. 6). One gets the impression the default-interventionist model is popular among behavioral economists not because it is the most realistic, but because it is the most flexible with regard to realistic choice models. Here's why:

The criteria by which DPT distinguishes the functional characteristics of mental processes are few and open-ended. While there are any number of interpretations of DPT in the cognitive sciences and psychology, the case has been made that what distinguishes mental processes which are slow, controlled, reflective, serial, rule-based, effortful, and conscious from those which are fast, reactive, automatic, intuitive, heuristic, associative, and unconscious rests on just basic two criteria: *autonomy* and *working memory* (Evans & Stanovich, 2013b; cf. Thompson, 2013). It's thus easy to see how DPT can be leveraged to characterize any number of decision anomalies.

But this gives rise to a more fundamental question, namely, if autonomy and working memory are all that is required to justify the ascription of some dual processes and/or systems, what does it take to possibly refute DPT? Consider the following passage by Pennycook:

> The observation that the distinction between intuition and reflection is irrefutable is foundational because it means that dual-process models should not be concerned with justifying this claim. That is, dual-process models must take this distinction as a given and build from there. If we know with a reasonable degree of certainty that the mind has this capacity for two different types of processes (autonomous and non-autonomous), where do we go from there? […] Thus, the mere distinction between intuition and reflection based on autonomy is sufficient for the claim that dual-process theory is irrefutable, but not sufficient for the claim that the theory is worth anyone's time. (2017, p. 8)

Although Pennycook does not regard this foundational aspect of DPT to be inherently problematic—he states, the "true test of a good theory is whether it can be applied successfully to problems and generate hypotheses" (2017, p. 8)—this should raise at least some alarms with regard to the descriptive aims of behavioral economists. One need not be card-carrying Popperian to see that irrefutability may run contrary to the aims of behavioral economics, which are "to improve the realism and psychological assumptions underlying economic theory" (Camerer, 1999). If DPT is thus construed as a functional framework, its theoretic structure allows for virtually unlimited interpretation.

Of course, one could retort that DPT is better construed as a meta-theory than a first-order theory. As Evans & Stanovich continue to argue, "such frameworks [meta-theories] cannot be falsified by the failure of any specific instantiation or experimental finding. Only specific models tailored to the tasks can be refuted in that way… " (2003b, p. 263). According to Evans & Stanovich, this is precisely why the System 1 /System 2 distinction is misleading—it gives the impression that the various dichotomies underlying DPT are strict and consistent across token theories—which we now know not to be the case. Hence, they abandon the system-based interpretation of DPT and commit to (what they believe is) a single, more coherent, dichotomy between autonomous processes and those which require working memory—this, recall, is the Type 1/Type 2 distinction, which they endorse as the least problematic interpretation of DPT (cf. Evans, 2017).

5: DPT and the myth of the inner rational agent

Where does this leave us? It seems that both styles of dual system modeling have limitations given their use or understanding of DPT. What justifies the continued reliance upon DPT by behavioral economists? In this section, I consider an alternative reason for DPT's popularity, one which trades on its fundamental irrefutability. I speculate that DPT appeals to behavioral economists because it satisfies modeling needs which are normative in origin: the role of DPT is to provide a psychological narrative that is seemingly realistic but does not give up one of the core features of economic theory, viz. the "myth of the inner rational agent".[11] However, I will argue that even this role has its limitations. In particular, DPT cannot satisfy certain rationality requirements. This final set of considerations draws upon the fact that even type-based interpretations of DPT may not suit the purposes of behavioral economics. This puts into perspective and exemplifies criticisms 4 - 6 raised in section 2

In everyday life, what distinguishes rational from irrational action is a matter of degree. The same cannot be said of economics—rationality is judged according to whether choice-behavior satisfies expected utility theory (or some variation of it). For this reason, the contrastive features of DPT appear to be tailor-made for economists: having the ability to distinguish between mental processes that generate reasoning errors and mental processes that don't is a critical tool for the analysis of rational choice. As already indicated, many consider System 2 to be the "rational system": this is not just because it supports higher cognitive functionings, like hypothetical and

[11] I borrow this term from Infante, Lecouteux, and Sugden (2016) who likewise observe that some behavioral economists rely on dualistic models of the human being, wherein an "inner rational agent" is trapped inside an "outer psychological shell". This model, they argue, has strong ties to neoclassical economics. However, where Infante, Lecouteux, and Sugden are concerned with what they call "preference purification" in welfare economics, I take the inner rational agent concept to be equally illustrative for the current purposes.

counter-factual thinking, but also because it is associated with the detection and inhibition of biased judgments and impulsive behaviors. But is it safe to presume that System 2 always produces rational outcomes? The answer is no; but to understand why, we need to clarify things.

Firstly, when proponents refer to the rational capacities of System 2, this is based on a standard of rationality that is rooted in the norms of deductive logic and probability theory (Evans & Over, 1996; Gigerenzer, 1996; Kahneman & Tversky, 1996; Stanovich, 1999; Stein, 1996). Although this normative standard has been a point of much contention in the philosophy and psychology of human reasoning (cf. Gigerenzer & Goldstein, 1996; Samuels, Stich, & Bishop, 2002; Samuels, & Stich, 2004; Over, 2004), we can set it aside for now. Secondly, System 2 is often identified with *critical thinking*. Critical thinking, according to the American Philosophical Association, is defined as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual consideration upon which that judgment is based" (Facione, 1990). With this in mind, we can then ask what it means to identify System 2 with successful critical thinking: (i) It could mean that System 2 is a necessary condition (prerequisite) for critical thinking; or (ii) it could mean that System 2 is sufficient for critical thinking, which is to say, that all instances of System 2 processing are instances of critical thinking (Bonnefon, 2018). It's not difficult to imagine why the latter interpretation isn't realistic. Not only is System 2 reserved for ordinary activities that don't meet requirements of critical thinking (reading a book engages many System 2 processes, like mental simulation and hypothetical thinking, but may not count as an instance of critical thinking). More importantly, there are instances where higher cognitive functions, like reflection and deliberation, cause people to make mistakes they may not have otherwise made. Bonnefon (2018) gives two examples of errors that result solely from System 2 processing: one is related to false justification or what he calls "pseudo-rational" answers. Often people *follow* their initial impulses and seek to justify them through clever rationalizations (this would support classic studies by Nisbett & Ross (1980) which reveals that individuals may give false verbal reports to justify actions they had no control over—often they do this without realizing the report is false). Another example of System 2 failures are the result of over-thinking, in which a person may mix-up or confuse relatively simple information by deliberation.[12] While these examples of System 2 errors are not as

[12] It's important to distinguish errors from over-thinking from errors due to cognitive fatigue. An important caveat of many system-based interpretations of DPT is that individuals make reasoning errors when System 2 does not have the resources to monitor or inhibit System 1 functions. The point I am making here is that this presupposes that System 2 is, by default, rational. But the fact that a well-functioning System 2 could produce errors is a separate possibility not often considered in the literature.

systematic as those advertised by heuristics and biases research, they illustrate an important point, which is that System 2 processing does not guarantee critical thinking.

How does this relate to DPT and the myth of the inner rational agent? In essence, the passage above drives a wedge between critical thinking and rational decision making. A successful completion of System 2 thinking is said to pass three stages, *conflict detection*, *sustained inhibition*, *explicit resolution* (De Neys and Bonnefon, 2013, Pennycook et al., 2015, Stanovich & West, 2009). Yet, if one thinks that rationality is defined according to the norms of logic and probability theory, then it reveals that System 2 does not guarantee rational choice—in fact, because rationality is based on coarser standards than critical thinking, it's likely that violations of rationality by System 2 are more common than instances of non-critical thinking (One can very easily make calculation errors while deliberating about a decision; this would constitute critical thinking by engaging System 2). Of course, this only addresses the strictest associations of System 2 with rational choice. One could still hold that System 2 is the rational authority and this agent sometimes fails to realize its potential.

Let's now consider the former disjunct, that System 1 is the primary cause of reasoning errors. While the above argument implies that System 1 is not the *sole* cause of reasoning errors (viz. because System 2 is sometimes involved), proponents of "ecological rationality" (Gigerenzer, Todd, and the ABC Group, 1999; Gigerenzer, 2004, 2007, 2008; Gigerenzer & Brighton, 2009) argue that it is inappropriate to presume that rationality is constrained by the norms of logic and probability theory. Many useful heuristics are generated by mental processes associated with System 1, not all of which lead to reasoning errors. Even if these processes are evolutionarily old and not conducive to modern life, automatic and implicit processing is critical for many higher-reasoning tasks. Yet, the reason why decision researchers and economists treat *these* processes as inherently irrational is because they help to predict a small range of decision phenomena that are relevant for some economic purposes. Hence, supporters of ecological rationality maintain that DPT *presupposes* its normative assumptions: if the only reason for adopting the terminology of System 1 and System 2 is that it justifies the normative standards of logic and probability theory, this also begs the question (Gigerenzer & Brighton, 2009; Gigerenzer & Sturm, 2012).

But perhaps this is only a problem for the system-based interpretation of DPT. What if behavioral economists, taking a cue from the god-fathers of DPT—Evans & Stanovich—adopt a subtler, more coherent framework for the analysis of rational choice? For instance, would the type-based interpretation of DPT help alleviate these issues? The answer is no.

Firstly, even recent modifications of DPT do not agree on the source(s) of reasoning errors. In fact, critics of DPT have doubled-down on the Type 1/Type 2 distinction for it further obscures the functional characteristics that are typically thought to differentiate systems (recall criticism 4). To see why, consider a more sophisticated

version of DPT, viz. Stanovich's tri-process model (2009, 2011): Presume that Type 2 processes are comprised of two stages, the algorithmic stage and the reflective stage: Can this be leveraged to better capture humans' rational faculties? The answer is still no, and here's why: the burden of explanation for how individuals solve higher-order reasoning tasks—the kind of solution that is relevant to economic reasoning and decision making—is compounded, not simplified. As criticisms 5 and 6 show, the algorithmic stage is, technically, not accessible to introspection; it operates below the threshold of conscious awareness and therefore is primarily responsible for initiating override procedures that the individual experiences as conflict (whereas the experience of conflict is at the reflective stage). Although there are plenty of candidates at the neuroanatomical level to host algorithmic processes, Stanovich uses primarily functional terminology to portray the rule-based nature of algorithmic processes. So again, the details of override initiation and the like are black-boxed among those reflective Type 2 processes. While it is tempting to say that the algorithmic stage is, perhaps, responsible for reasoning errors that critics attribute to System 1, there is no evidence that this is what actually happens.[13] The only reason why one would attribute, say, computational errors that result in decision anomalies, is because the theory of type-based processing says so. In fact, there is no theoretical reason why the reflective stage of Type 2 couldn't cause reasoning errors (just as System 2 was shown to). When individuals (unknowingly) justify mistakes or attempt to assimilate impulsive behaviors with pseudo-rational answers, this must pass through the reflective stage. This serves as a further reason for not giving Type 2 processes full rational authority.

It thus seems that the theoretical affinity DPT's System 2 (Type 2) is misplaced, despite the appeal of having a psychological basis for inner rational agency. Even if behavioral economists concede that the interactions between System 1 and System 2 are only an approximation of the inner dynamics of decision making, it would require significant deviations from DPT to continue to promote dual system models which presume that somewhere, within the individual, is a rational agent.

[13] That is to say, there is no clear neuroscientific evidence to support this. Rather, Stanovich (2011) uses psychometric data to make a compelling case for the role of the algorithmic stage in Type 2 processing (even if it's not clear what underwrite those processes). His argument rests on claims that individual differences in cognitive ability are indications that something, prior to reflective processing, has override control of Type 1 processes. The reason why these processes simply aren't relegated to Type 1 is because they are not intrinsically autonomous; rather, they are learned and internalized with practice. Reading comprehension and arithmetical skills are examples of learned skills. The case has been made that such data relies on question-begging assumptions and driven by confirmation bias (Polonioli, 2014).

6. Concluding remarks

I've argued that behavioral economists are faced with a dilemma and must revise their faith in DPT. In section 2, I provided an overview of the current status of DPT from the perspective of (philosophy of) psychology and cognitive science; I elaborated six criticisms of DPT, three of which pertain to system-based interpretations (upon which the standard menu of mental processes is commonly defined), and three more which pertain to type-based modifications of DPT. The take-away message of this section is that DPT is not as descriptively accurate as it is often portrayed to be, and therefore, not a reliable model of the mind/brain: it posits structures that it cannot articulate and processes that it cannot track. Section 3 surveyed both the explicit and implicit influences of DPT upon behavioral economics posited that there are two general styles of dualistic modeling in behavioral economics. Section 4 considered whether the use of DPT generated tensions for each of the two styles of dualistic modeling and discussed their possible limitations. In Section 5 I argued that, at best, DPT provides a narrative upon which behavioral economists can structure their models. But even this minimal role would require justification given that DPT cannot sustain the normative foundations of dualistic models. This is because neither system-based nor type-based interpretations cleanly distinguish rational from irrational choice.

However, one can envision a number of reasons why behavioral economists may not wish to give up on DPT, and further, may not wish to adopt a new psychological model to reinterpret classic (genre-defining) experiments. They may argue that existing DPT-based economic models offer novel predictions—and that descriptive accuracy does not imply or require psychological realism (cf. Angner & Loewenstein 2012). In fact, behavioral economists may not be willing to give up on DPT precisely because it's way of partitioning the mind is naturally attuned to the needs of economic analysis. The question that apologists are tasked with, then, is how to retain the framework of DPT while mitigating its theoretical and empirical inadequacies.

7. Bibliography

Angner, E., & Loewenstein, G. (2012). Behavioral economics. In Mäki, U. (Ed.), Philosophy of Economics (pp. 641-689). Amsterdam: Elsevier.

Ainslie, G. (1992). *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Cambridge University Press.

Ainslie, G. (2001). *Breakdown of Will*. Cambridge University Press.

Ainslie, G. (2005). Précis of Breakdown of will. *Behavioral and Brain Sciences*, 28(5), 635-650.

Ainslie, G., & Haslam, N. (1992). Self-control. In Loewenstein, G., & Elster, J. (Eds) *Choice over time* (pp. 177-209). New York, NY, US: Russell Sage Foundation.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3), 871-915.

Benhabib, J., & Bisin, A. (2005). Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions. *Games and Economic Behavior*, 52(2), 460-492.

Bernheim, B.D., & Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review*, 94(5), 1558-1590.

Bernheim, B.D. & Rangel, A. (2007). Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review*, 97(2), 464-470.

Bernheim, B.D. (2009). The psychology and neurobiology of judgment and decision making: What's in it for economists?. In Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A (Eds.), *Neuroeconomics: Decision Making and the Brain* (pp. 113-125). Academic Press.

Berg, N., & Gigerenzer, G. (2010). As-if behavioral economics: Neoclassical economics in disguise?. *History of Economic Ideas*, 133-165.

Bonnefon, J.F. (2018). The pros and cons of identifying critical thinking with System 2 processing. *Topoi*, 37(1), 113-119.

Botvinick, M.M., & Cohen, J.D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, 38(6), 1249–1285.

Botvinick, M., & Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annual Review of Psychology*, 66, 83–113.

Brocas, I., & Carrillo, J.D. (2008). The brain as a hierarchical organization. *The American Economic Review*, 98(4), 1312-1346.

Brocas, I., & Carrillo, J.D. (2014). Dual-process theories of decision-making: A selective survey. *Journal of Economic Psychology*, 41, 45-54.

Buturovic, Z. & Tasic, S. (2015). Kahneman's failed revolution against economic orthodoxy. *Critical Review*, 27(2), 127-145.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism". *University of Pennsylvania Law Review*, 151(3), 1211-1254.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1), 9-64.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process Theories in Social Psychology*. Guilford Press.

Craver, C.F., & Alexandrova, A. (2008). No revolution necessary: neural mechanisms for economics. *Economics & Philosophy*, 24(3), 381-406.

Daugman, J. G. (1993). Brain metaphor and brain theory. In *Computational Neuroscience* (pp. 9-18). MIT Press.

D'Esposito, M., Detre, J.A., Alsop, D.C., Shin, R.K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, 378, 279–281.

De Neys, W. (Ed.). 2017. *Dual process theory 2.0*. Routledge.

De Neys, W., & Bonnefon, J.F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences*, 17(4), 172-178.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8), 709.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytic-rational thinking styles. *Journal of Personality and Social Psychology*, 71, 390–405.

Evans, J.S.B.T. (1989. Bias in human reasoning: Causes and consequences. *Lawrence Erlbaum Associates*, Inc.

Evans, J.S.B.T. (2006). Dual system theories of cognition: Some issues. *In Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 202-207.

Evans, J.S.B.T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255-278.

Evans, J.S.B.T. (2009). How many dual-process theories do we need? One, two, or many? In eds. Evans, J.S.B.T., & Frankish, K. (Eds.), *In two minds: Dual processes and Beyond* (pp. 33-54). Oxford University Press.

Evans, J.S.B.T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, 31(2-3), 86-102.

Evans, J.S.B.T., (2012). Dual process theories of deductive reasoning: facts and fallacies. In Holyoak, K., & Morrison, R. (Eds.), *The Oxford handbook of Thinking and Reasoning*, pp.115-133. Oxford University Press.

Evans, J.S.B.T. (2017). Dual process theory: perspectives and problems. In De Neys, W. (Ed.), *Dual Process Theory 2.0* (pp. 145-164). Routledge.

Evans, J.S.B.T., & Frankish, K.E. (Eds.). (2009). *In Two Minds: Dual Processes and Beyond*. Oxford University Press.

Evans, J.S.B.T., & Over, D.E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.

Evans, J.S.B.T., & Stanovich, K.E. (2013a). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241.

Evans, J.S.B.T., & Stanovich, K.E. (2013b). Theory and metatheory in the study of dual processing: Reply to comments. *Perspectives on Psychological Science*, 8(3), 263-271.

Evans, J.S.B.T., & Wason, P.C. (1976). Rationalization in a reasoning task. *British Journal of Psychology*, 67(4), 479-486.

Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. *The Delphi Report*.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2), 351-401.

Gigerenzer, G., 2001. Decision making: Nonrational theories. In *International Encyclopedia of the Social and Behavioral Sciences* (pp. 3304-3309). Elsevier Science.

Gigerenzer, G. (2004). Striking a blow for sanity in theories of rationality. In M. Augier & J. G.March (Eds.), *Models of a man: Essays in memory of Herbert A. Simon* (pp. 389–409). MIT Press.

Gigerenzer, G. (2007). *Gut Feelings: The Intelligence of the Unconscious*. New York: Viking Press.

Gigerenzer, G. (2008). *Rationality for mortals*. New York: Oxford University Press.

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, 6(3), 361-383.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107-143.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4), 650.

Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin*, 119, 23–26.

Gigerenzer, G., & Sturm, T. (2012). How (far) can rationality be naturalized?. *Synthese*, 187(1), 243-268.

Gigerenzer, G., Todd, P.M., & The ABC Research Group. (1999). *Simple Heuristics That Make us Smart*. New York: Oxford University Press

Gilbert, D.T. (1999). What the mind's not. In Chaiken, S., & Trope, Y. (Eds.), *Dual-process Theories in Social Psychology*, (pp.3-11). Guilford Press.

Gilbert, D.T. (2002). Inferential correction. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 167-184). New York, NY, US: Cambridge University Press

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The Psychology of Intuitive Judgment*. Cambridge university press.

Glimcher, P.W. (2003). The neurobiology of visual-saccadic decision making. *Annual Review of Neuroscience*, 26(1), 133-179.

Gollwitzer, P., Fujita, K., & Oettingen, G. (2004). Planning and the implementation of goals. In *Handbook of Self-regulation*: *Research, Theory, and Applications*. Guilford Press.

Grayot, J., (*forthcoming*). From selves to systems: on the intrapersonal and intraneural dynamics of decision making. *Journal of Economic Methodology*.

Grüne-Yanoff, T. (2017). Reflections on the 2017 Nobel Memorial Prize Awarded to Richard Thaler. *Erasmus Journal for Philosophy and Economics*, 10(2), 61-75.

Harrison, G., & Ross, D. (2010). The methodologies of neuroeconomics. *Journal of Economic Methodology*, 17(2), 185-196.

Heilmann, C. (2014). Success conditions for nudges: a methodological critique of libertarian paternalism. *European Journal for Philosophy of Science*, 4(1), 75-94.

Heukelom, F. (2014). *Behavioral economics: A history.* Cambridge University Press.

Hutto, D. D., Myin, E., Peeters, A., & Zahnoun, F. (2018). Putting computation in its place. In Sprevak M., & Colombo M. (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 272-282). London: Routledge.

Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1), 1-25.

Kahneman, D. (2003a). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5), 1449-1475.

Kahneman, D. (2003b). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, 58(9), 697.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In Kahneman, D., & Gilovich, T. (Eds.), *Heuristics and biases: The Psychology of Intuitive Judgment*, 49-81.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In Holyoak, K., & Morrison, R. (Eds.), The Cambridge Handbook of Thinking and Reasoning, (pp. 267-293). Cambridge University Press.

Kahneman, D., & Frederick, S. (2007). Frames and brains: Elicitation and control of response tendencies. *Trends in Cognitive Sciences*, 11(2), 45-46.

Kahneman, D., & Thaler, R.H. (2006). Anomalies: Utility maximization and experienced utility. *Journal of Economic Perspectives*, 20(1), 221-234.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263-292.

Kahneman, D., Slovic, P., & Tversky, A. (1982). (1982) Judgment under uncertainty: Heuristics and biases. *Judgement under uncertainty: Heuristics and biases*, (pp. 3-20). Cambridge University Press.

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3), 257-262.

Keren, G., & Teigen, K.H. (2004). Yet another look at the heuristics and biases approach. In Koehler, D.J., & Harvey, N. (Eds.), *Blackwell handbook of judgment and decision making*, (pp.89-109). John Wiley & Sons.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4(6), 533-550.

Kirkpatrick, L.A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, 63(4), 534.

Kruglanski, A.W., & Orehek, E. (2007). Partitioning the domain of social inference: Dual mode and systems models and their alternatives. *Annual Review of Psychology*, 58, 291-316.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberative judgments are based on common principles. *Psychological Review*, 118, 97–109.

Lieberman, M.D., Gaunt, R., Gilbert, D.T. and Trope, Y. (2002). Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. In *Advances in experimental social psychology* (Vol. 34, pp. 199-249). Academic Press.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3), 272-292.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *The American Economic Review*, 90(2), 426-432.

Loewenstein, G., & O'Donoghue T. (2005). Animal Spirits: Affective and Deliberative Processes in Economic Behavior. *CMU Working Paper*.

Loewenstein, G. and Haisley, E.C. (2007). The economist as therapist: Methodological ramifications of 'light' paternalism. In *Handbook of Economic Methodologies.*

Lurquin, J.H. and Miyake, A. (2017). Challenges to ego-depletion research go beyond the replication crisis: a need for tackling the conceptual crisis. *Frontiers in Psychology*, 8, 568.

Mars, R., Sallet, J., Rushworth, M., & Yeung, N. (Eds.). (2011). *Neural Basis of Motivational and Cognitive Control*. Cambridge, MA: MIT Press.

McClure, S.M., Laibson, D.I., Loewenstein, G., & Cohen J.D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695), 503-507.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience*, 27(21), 5796-5804.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychological Review*, 106(1), 3.

Mugg, J. (2016). The dual-process turn: How recent defenses of dual-process theories of reasoning fail. *Philosophical Psychology*, 29(2), 300-309.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11(6), 988-1010.

Over, D. (2004). Rationality and the normative/descriptive distinction. In Koehler, D.J., & Harvey, N. (Eds.), *Blackwell handbook of judgment and decision making*, (pp. 3-18). John Wiley & Sons.

Parker, A.J., & Newsome, W.T. (1998). Sense and the single neuron: probing the physiology of perception. *Annual Review of Neuroscience*, 21(1), 227-277.

Pennycook, G. (2017). A perspective on the theoretical foundation of dual process models. In De Neys, W. (Ed.), *Dual Process Theory 2.0* (pp. 13-35). Routledge.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80, 34-72.

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68(1), 29.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, *37*(3), 453–488.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Ross, D. (2005). Economic Theory and Cognitive Science: Microexplanation. MIT Press.

Nisbett, R.E. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall.

Samuels, R., Stich, S., & Bishop, M. (2012). Ending the rationality wars. *Collected Papers, Volume 2: Knowledge, Rationality, and Morality*, 1978-2010, 2, 191.

Schelling, T.C. (1984). Self-command in practice, in policy, and in a theory of rational choice. *The American Economic Review*, 74(2), 1-11.

Sent, E.M. (2004). Behavioral economics: how psychology made its (limited) way back into economics. *History of Political Economy*, 36(4), 735-760.

Shefrin, H.M., & Thaler, R.H. (1988). The behavioral life-cycle hypothesis. *Economic Inquiry*, 26(4), 609-643.

Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127.

Sinayev, A. (2016). Dual-System Theories of Decision Making: Analytic Approaches and Empirical Tests. (Electronic Thesis or Dissertation). Retrieved from https://etd.ohiolink.edu/

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3.

Stanovich, K.E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning*. Psychology Press.

Stanovich, K.E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. University of Chicago Press.

Stanovich, K.E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory. In eds. Evans, J.S.B.T., & Frankish, K. (Eds.), *In two minds: Dual processes and Beyond* (pp. 55-88). Oxford University Press.

Stanovich, K.E. (2011). *Rationality and the Reflective Mind*. Oxford University Press.

Stanovich, K.E., & West, R.F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences*, 23(5), 645-665.

Stanovich, K. E., & West, R. F. (2009). *What Intelligence Tests Miss. The Psychology of Rational Thought*. Yale University Press.

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford, England: Oxford University Press.

Strack, F., Werth', L., & Deutsch, R. (2006). Reflective and Impulsive Determinants of Consumer Behavior. *Journal of Consumer Psychology*, 16(3), 205-216.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643.

Szameitat, A.J., Schubert, T., Muller, K., & von Cramon, D.Y. (2002). Localization of executive function in dual-task performance with fMRI. *Journal of Cognitive Neuroscience*, 14(8), 1184–1199.

Thaler, R.H., & Shefrin. H.M. (1981). An economic theory of self-control. *The Journal of Political Economy*, 89(2), 392-406.

Thaler, R.H., & Sunstein, C.R. (2003). Libertarian paternalism. *American Economic Review*, 93(2), 175-179.

Thaler, R.H., & Sunstein, C.R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press

Thaler, R.H., Sunstein, C.R., & Balz, J.P. (2012). Choice architecture. In Shafir, E. (Ed.), *The Behavioral Foundations of Public Policy*, (pp. 428-439). Princeton University Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Thompson, V. (2013). Why It Matters: The Implications of Autonomous Processes for Dual Process Theories—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8, 253–256.

Thompson, K.G., & Schall, J.D. (1999). The detection of visual signals by macaque frontal eye field during masking. *Nature Neuroscience*, 2(3), 283.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Scien*ce, 185(4157), 1124-1131.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92(7), 345-381.

Chapter 6

Looking back and looking ahead…

1. Looking back

In economics, *agency* and *choice* are concepts that are inextricably linked. Agents make choices, and choices represent agents' preferences. One concept cannot persist without the other. Hence, in recognizing that individuals are boundedly rational, indeed even boundedly individual, economists and behavioral decision researchers who use rational choice theory have been faced with a decision: either ignore anomalies produced by decades of experimental research and interdisciplinary collaboration and continue on with the standard tools and operative concepts of orthodox economics; or, address anomalies and revise those tools and operative concepts. The literature on philosophy and methodology of economics, as well as on behavioral economics and behavioral decision research indicates that economists are divided over how to proceed.

Chapters 2 – 5 project two main approaches to reconciling the tension between agency and choice. One approach views individual persons as the primary objects of study for economics, and as such, look to psychology and the neurosciences to identify more appropriate loci for the study of choice (either in the brain, or within functional structures that support decision-making). The other approach views individual persons not as the primary object of study (economic agents are the primary objects, and they are ontologically distinct from persons given that they are purely theoretical entities). As such, choice should be construed as the outcome of external pressures like markets, institutions, and social norms, which impart constraints as well as socio-cognitive support. So where does one go from here? My intuition is to follow the trajectory of the two approaches down separate paths.

2. Looking ahead

Let's suppose that behavioral economists wish to achieve greater descriptive power concerning models of the internal dynamics of decision-making (perhaps they don't—but suppose they do), it seems that they have two options: option 1 is they can wait to see what comes of the on-going debates between *behavioral economics in the scanner* and *economics of neural activity* approaches in neuroeconomics. This could prove promising given that the fields do seem to be converging in unexpected ways (Vromen, 2011). But it could be argued that the convergence of neuroeconomic approaches is bringing more rather than less complexity concerning where and how decisions are made, and choices executed (Fumagalli, 2011, 2016). This mixture of

added complexity and lack of clear limitations about explanatory benefits of neuroeconomics may not be appealing to behavioral economists who were not already committed to the neural enhancement of economics models. Thus, Option 2:

Below are two alternative psychological frameworks which provide some relief from the theoretical and empirical inadequacies of dual process theory while also offering novel ways to predict and explain decision phenomena.

Alternative 1: "Decision field theory" (cf. Busemeyer & Johnson, 2004, 2008) is a computational model of decision making which uses a connectionist, neural network framework to represent preference formation. Rather than represent decisions as a deterministic set of cognitive processes, decision field theory represents choice options via a network of actions with interconnected property nodes; the value of a given action is affected by the attention weight which links an action to a given property. Attention weights are influenced by background beliefs and information, but are inherently stochastic. A preference state is achieved when the accumulation of attention weights reaches a threshold and induces an action.

The primary benefit of computational models of cognition such as decision field theory is that offer a legitimately computational basis for human learning and inference by way of mathematical modeling and computer simulation (and, of course, behavioral experiments). When applied to the study of decision making, such models provide a means of tracking utility optimization procedures in a way that can track preference formation. This would constitute a more realistic interpretation of the information processing metaphor that behavioral economists use.

The limitation of such a model is that it's not evident how individuals' *mental states* mediate the distribution of attentional weights to actions when decision field theory is interpreted as an artificial neural network—in this way, it is comparable to functionalist accounts of dual process theory which black-boxes processes like override and conflict monitoring functions which prevent automatic and impulsive behavior from occurring. Yet, when applied directly to the study of the brain, the computational basis of decision field theory is better able to accommodate the "noise" associated with stochastic attentional shifting and this has great potential to explain both the causes of reasoning errors, and hence capture decision anomalies that concern behavioral economists, while also providing a realistic depiction of underlying decision processes. Individuals' choices are not formed through linear reasoning procedures, as dual-process-based economic models presuppose; real decision-making is messy and fragmented, and this is ignored by current dual process models (even by neuroeconomic applications of dual process models.

Alternative 2: While Bayesian models traditionally offer little insight into the psychological basis of decision making, certain "enlightened" Bayesian models of cognition have the potential to unite rational analysis of the Bayesian program with cutting edge knowledge of cognitive mechanisms which do underwrite decision

procedures. In Jones & Love (2011), several candidate models are proposed, each of which identifies a different area of cognition and/or perception that is integral to the decision process. While it remains to be seen how well these models predict novel decision phenomena (many candidate models are being currently tested), there is reason to believe that a Bayesian model of cognition applied to local or specific cognitive and perceptual tasks could explain how decision anomalies occur without adverting to "bargaining games" or "tradeoffs" between dual systems whose underlying functional characteristics aren't well-defined. Enlightened Bayesian models of cognition seek to ground optimization procedures in the very mechanisms that cognitive science recognizes to be complicit in reasoning errors. If it can be shown that certain mechanisms, or clusters of mechanisms, abide by Bayes's rule and "compute" optimization procedures, this potentially avoids many of the conceptual and ontological confusions generated by dual-process-based economic models.

Further, unlike computational models of cognition, which are most descriptive and hence most illuminating when applied directly to the brain, Bayesian models of cognition claim to apply to multiple-levels of analysis (to use Marr's distinction). Although there are different models on the market, and it will take time to determine which are amenable to the purposes of behavioral economics, some Bayesian models of perception claim to adequately bridge computational, algorithmic, and implementation levels in a way that does not conflate their functional characteristics. If true, this could provide a remarkable basis for grounding rational analysis that behavioral economists seek. But, this, like the former alternative, is speculative and requires testing in economic conditions before it can be declared viable or not viable…

3. Bibliography

Busemeyer, J.R., & Johnson, J.G. (2004). Computational models of decision making. *In Koehler, D.J., & Harvey, N. (Eds.), Blackwell handbook of judgment and decision making,* (pp.133-154). John Wiley & Sons.

Busemeyer, J.R., & Johnson, J.G. (2008). Micro-process models of decision making. *Cambridge handbook of Computational Psychology*, 302, p.321.

Fumagalli, R. (2013). The futile search for true utility. *Economics & Philosophy*, 29(3), 325-347.

Fumagalli, R. (2016). Five theses on neuroeconomics. *Journal of Economic Methodology*, 23(1), 77-96.

Jones, M., & Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169-188.

Vromen, J. (2011). Neuroeconomics: two camps gradually converging: what can economics gain from it?. *International Review of Economics*, 58(3), 267-285.

Samenvatting

In dit proefschrift geef ik een filosofisch perspectief op verschillende opvattingen over sleutelbegrippen als actorschap, rationaliteit en preferentie, en de relatie die ze hebben met keuze. Het filosofische perspectief dat ik inneem is tweeledig: enerzijds kan filosofische analyse de dubbelzinnigheden van definities en concepten die in interdisciplinair onderzoek kunnen ontstaan verhelderen. Dit is belangrijk gezien het feit dat traditionele filosofische concepten als geest, cognitie en intentionaliteit een rol spelen in hedendaagse economische studie van keuze. Een van de doelen van dit proefschrift is dan ook om hedendaags onderzoek naar actorschap en keuze aan een dergelijk filosofisch onderzoek te onderwerpen. Anderzijds kunnen de vragen en onderwerpen die in dit proefschrift worden besproken worden opgevat als wetenschapsfilosofie: ze gaan over wetenschappelijke praktijken, zowel theoretisch als empirisch. Om dit te doen richt het proefschrift zich op economisch en gedragsonderzoek. Dit omvat traditionele micro-economische disciplines, zoals besluitvorming en speltheorie, maar het heeft ook betrekking op nieuwe interdisciplinaire vakgebieden die tussen economie en cognitieve wetenschappen inliggen, zoals gedragseconomie, neuro-economie en experimentele psychologie.

Hoofdstuk 2 geeft een brede karakterisering van de controversiële relatie tussen actorschap en keuze door zich te richten op een centraal debat in de filosofie van de economie. Behaviorisme, breed opgevat, is de positie dat mensen stimulus-respons machines zijn, en dat gedrag kan worden beschreven en verklaard zonder verwijzing naar mentale gebeurtenissen of interne psychologische processen. Gedragsdeskundigen hebben de neiging om individuele acties te beschouwen als geconditioneerde reacties op externe krachten. Mentalisme daarentegen is het standpunt dat mensen meer zijn dan stimulus-responsmachines, en dat economen, om de beslissingen en het keuzegedrag van individuen te begrijpen, wellicht het reilen en zeilen van de geest en/of de hersenen moeten onderzoeken. Hoofdstuk 2 evalueert de relevantie van de mentalisme-behaviorisme (MB) tweestrijd in de economie in het licht van recente debatten en de daaruit volgende argumenten ten gunste van het mentalisme. Ik beargumenteer dat er twee problemen zijn met de huidige opvattingen over de MB-onderscheid zoals het van toepassing is op de manier waarop economen en beslissingsonderzoekers bewijsmateriaal interpreteren en verzamelen. Ten eerste is het onduidelijk waar het MB-onderscheid precies over gaat of betrekking op heeft —dat wil zeggen, economen en beslissingsonderzoekers kunnen verschillende motivaties hebben om het mentalisme te onderschrijven en/of om zich te verzetten tegen behaviorisme. Ten tweede, en nog belangrijker, is het onduidelijk hoe het MB-onderscheid verondersteld wordt empirisch onderzoek in de economie en het beslissingsonderzoek te verbeteren of te bevorderen. Met name aanhangers van het mentalisme hebben de moeilijke taak om te verduidelijken wat mentalisme inhoudt.

Met betrekking op het eerste probleem beschouw ik twee veelvoorkomende motivaties om het mentalisme te steunen: de ene motivatie doet een beroep op de keuzetheoretische grondslagen van de economie; de andere doet een beroep op de wetenschappelijke praktijk in de economie. Met betrekking op het tweede probleem beargumenteer ik dat het MB-onderscheid waarschijnlijk geen vooruitgang of verbetering van de wetenschappelijke praktijk in de hedendaagse economische context zal opleveren, omdat noch het mentalisme (noch het behaviorisme) in staat is om verklarende problemen te analyseren en op te lossen die uniek zijn voor niet-keuzedata, d.w.z. psychologische en neurowetenschappelijke data. Ik besluit met het bespreken van de beperkingen van het functionalisme, de steunpilaar van het mentalisme, en stel ten opzichte van het MB-onderscheid alternatieve onderscheiden voor, waarvan sommigen al gebruikt worden in de naburige cognitieve en gedragswetenschappen.

Hoofdstuk 3 gaat in op de vraag of, d.w.z. onder welke omstandigheden, mensen zich gedragen als economische actoren. In tegenstelling tot de debatten die in hoofdstuk 2 worden besproken, die een overwegend individualistische benadering tot economische concepten en beslissingsfenomenen hanteren, laat dit hoofdstuk zien hoe externe krachten zoals sociale instituten en informatiestructuren individueel gedrag zowel ondersteunen als beperken. Ik beargumenteer dat individualisme problematisch is als basis voor het onderzoek naar sociale interactie. Daarbij bestudeer ik de theorie van Don Ross (2005, 2006) over 'multiple-selves' als een manier om de beperkte rationaliteit van individuen te verzoenen met hun beperkte individualiteit. Ross stelt dat individuele personen complexe samenvoegingen van “zelven” zijn, die ontstaan als reactie op druk van buitenaf om individueel gedrag te reguleren en het volgen van publieke normen en conventies mogelijk te maken. Ik onderzoek dus de verschillende rollen die “zelven” spelen in Ross' bredere filosofie van de economie en ik identificeer afzonderlijke projecten die zich daarin voordoen. Ik onderscheid drie verschillende rollen voor “zelven”, een evolutionaire, narratieve en een economische. Ik stel dat deze rollen bijdragen aan twee verschillende, maar overlappende projecten. Ik beargumenteer dat, hoewel het niet problematisch is om “zelven” te begrijpen op basis van hun verschillende rollen, we er niet van uit moeten gaan dat hun functies of eigenschappen in één rol dezelfde doelen kunnen dienen voor verschillende projecten.

Na het belang van externe krachten voor het begrijpen van de quasi-economische actorschap van de mens te hebben uitgewerkt, keert hoofdstuk 4 terug naar het domein van de individuele besluitvorming. De vraag die hier wordt gesteld is: hoe integreren interdisciplinaire benaderingen van beslissingsonderzoek psychologische inzichten met economische methoden? En, wat zijn de conceptuele en ontologische problemen voor een dergelijke integratie? Hierin kijk ik kritisch hoe “multiple-self” modellen van intrapersoonlijke en intertemporele keuze zijn geïntegreerd met “dual-process” en “dual-system” theorieën uit de sociale psychologie en cognitieve wetenschap.

Multiple-self modellen van intrapersoonlijke en intertemporele keuzes kwamen naar voren in de beslissingstheorie en speltheorie om economen te helpen de dynamiek van interne conflicten beter te begrijpen en om anomalieën en inconsistenties die zich voordoen te voorspellen en – hopelijk – te verklaren. Dual-proces theorieën over redeneren en oordelen zijn een ander middel om interne conflicten vast te leggen. Het stelt onderzoekers in staat om "hogere" cognitieve processen te onderscheiden, die geassocieerd worden met doordachte oordelen en het vermogen om logisch te redeneren, van "lagere", meer primitieve informatieprocessen, die meestal geassocieerd wordt met affectieve toestanden en emotionele reacties. Ik gebruik de term 'multi-agent model' om modellen aan te duiden die gebruik maken van meerdere actoren met contrasterende psychologische vaardigheden. Dergelijke modellen lijken steeds populairder te worden gezien hun vermeende vermogen om redeneerfouten en beslissingsanomalieën als gevolg van interne conflicten of gebrek aan zelfcontrole te voorspellen en te verklaren. Ik analyseer hoe multi-actorenmodellen "zelven" en "systemen" opvatten en gebruiken om intrapersoonlijke en intraneurale conflicten voor te stellen. Het hoofdstuk is gestructureerd aan de hand van drie beweringen. De eerste en tweede bewering stellen vast dat multi-agentmodellen zowel conceptueel als ontologisch ambigu zijn. De derde bewering stelt dat deze ambiguïteiten kunnen leiden tot problemen in het wetenschappelijke streven naar begrip van keuzefenomenen. Het onderzoek van multi-agentmodellen is niet alleen cruciaal om te begrijpen hoe economenen en psychologen zelfcontrole interpreteren en modelleren, maar biedt ook een belangrijke kans om de effecten te bestuderen van de wederzijdse beïnvloeding van verschillende disciplines.

Hoofdstuk 5 bouwt voort op de argumenten uit hoofdstuk 4 en gaat in op het succesverhaal van de gedragseconomie. Het onderzoekt de rol die de dual process-theorie (DPT) heeft gespeeld in de gedragseconomie en gaat over de vraag wat de beschrijvende beperkingen zijn van psychologisch dualistische modellen. Zowel cognitieve wetenschappers als filosofische psychologen hebben kritiek geuit op de theoretische fundamenten van de standaardvisie van de DPT en hebben beargumenteerd dat het bewijs dat wordt gebruikt om deze theorie te ondersteunen niet geldig is. Bovendien hebben recente wijzigingen van de DPT naar aanleiding van deze kritiek tot extra zorgen geleid over de toepasbaarheid en onweerlegbaarheid ervan. Ik beargumenteer dat dit tot bezorgdheid zou moeten leiden bij gedragseconomen die de DPT zien als een psychologisch realistische basis voor hun modellen. In het bijzonder verhoogt het de mogelijkheid dat dualistische modellen niet zo nauwkeurig of betrouwbaar zijn als gedragseconomen veronderstellen. In feite kan men stellen dat de populariteit van de DPT in de gedragseconomie minder te maken heeft met het empirische succes van dualistische modellen, en meer met het gemak dat het duale verhaal economen biedt die op zoek zijn naar het oplossen van beslissingsanomalieën. Ik beargumenteer dat de groeiende kritiek op DPT

gedragseconomen met een dilemma achterlaat: of ze houden vast aan hun vermeende ambities om een realistische beschrijving van menselijke besluitvorming te geven en wijzigen hun gebruik van DPT, of ze houden vast aan DPT en wijzigen hun ambities.

Hoofdstuk 6 besluit mijn proefschrift. In dit hoofdstuk denk ik na over de vraag hoe nu verder? In hoofdstukken 2-5 komen twee hoofdbenaderingen naar voren om de spanning tussen actorschap en keuze te verzoenen. De ene benadering beschouwt individuele personen als de primaire studieobjecten voor de economie, en als zodanig kunnen psychologie en neurowetenschappen helpen bij het bepalen van de juiste benadering naar dit studieobject. De tweede benadering beschouwt individuele personen niet als het primaire studieobject (economische actoren zijn de primaire studie, en ze zijn ontologisch verschillend van personen). Als zodanig moeten keuzes worden opgevat als het resultaat van externe (markt)druk. Dus, voor elk van deze benaderingen duiken er er nieuwe ontwikkelingen en nieuwe filosofische vragen op.

Summary

In this thesis, I offer a philosophical perspective on the different conceptions of key notions such as *agency*, *rationality*, and *preference*, and their relation to *choice*. The philosophical perspective I offer is two-fold: on the one hand, philosophical analysis can clarify ambiguities of definitions and concepts that can arise in interdisciplinary research. This is of particular importance given how traditional philosophical concepts such as *mind*, *cognition*, and *intentionality* feature in contemporary economic studies of choice. Hence, one project of this thesis is to subject cutting-edge research on questions of agency and choice to such philosophical scrutiny. On the other hand, the questions and topics discussed in this thesis can be understood as an exercise in philosophy of science: they deal with questions and topics that pertain to the practices, both theoretical and empirical, of scientists. To this end, the thesis targets economics and behavioral decision research. This includes traditional microeconomic disciplines, such as decision and game theory; but it also can be extended to new interdisciplinary syntheses between economics and the cognitive sciences, such as behavioral economics, neuroeconomics, and experimental psychology.

Chapter 2 provides a broad characterization of the contentious relationship between agency and choice by focusing on a pivotal debate in the philosophy of economics. *Behaviorism*, broadly construed, is the position that humans are stimulus-response machines, and that behavior can be described and explained without making reference to mental events or to internal psychological processes. Behaviorists tend to regard individual actions as patterned—or conditioned—responses to external forces. *Mentalism*, by contrast, is the position that humans are more than stimulus-response machines, and that in order to understand individuals' decisions and choice-behaviors, economists may need to investigate the goings-on of the mind and/or brain. Chapter 2 thus evaluates the relevance of the mentalism-behaviorism (MB) dichotomy in economics in light of recent debates and subsequent arguments in favor of mentalism. I argue that there are two problems with current conceptions of the MB dichotomy as it pertains to how economists and decision researchers interpret and gather evidence. First, it is unclear what the MB dichotomy pertains to or is about exactly—which is to say, economists and decision researchers may have different motivations for endorsing mentalism and/or for opposing behaviorism. Second, and more importantly, it is unclear how the MB dichotomy is supposed to improve or advance empirical research in economics and decision research—in particular, supporters of mentalism have the difficult task of clarifying what mentalism entails or consists in. In response to the first problem, I consider two common motivations for endorsing mentalism: one motivation appeals to the *choice-theoretic foundations* of economics; the other appeals to *scientific practice* in economics. In response to the second problem, I argue that the MB dichotomy likely won't advance or improve scientific practice in

contemporary economic settings because neither mentalism (nor behaviorism) are equipped to analyze and resolve explanatory problems that are unique to non-choice data, i.e. psychological and neuroscientific data. I conclude by discussing the limitations of functionalism, the mainstay of the mentalism defense book, and suggest alternative schemas to the MB dichotomy, some of which are employed in neighboring areas of the cognitive and behavioral sciences.

Chapter 3 considers whether, i.e. under what conditions, human persons behave like economic agents. In contrast to debates discussed in chapter 2, which take a predominantly *individualistic* approach to the analysis economic concepts and decision phenomena, this chapter demonstrates how external forces such as social institutions and informational structures both support and constrain individual behaviors. I argue that individualism is problematic as a basis for investigating social interaction. In so doing I examine the Don Ross' (2005, 2006) account of "multiple-selves" as a way of reconciling individuals' bounded rationality with their bounded individuality. Ross argues that individual persons are complex aggregations of selves, which arise in response to external pressures to regulate individual behaviors and enable the tracking of public norms and conventions. I thus investigate the different roles that selves play in Ross' broader philosophy of economics and I identify separate projects that arise therein. I distinguish three different roles for selves, which are *evolutionary*, *narrative*, and *economic*, and I argue that these roles contribute to two distinct, but overlapping, projects. I will argue that, while it is not problematic to conceive of selves according to their different roles, we should not presume that the functions or properties of selves in one role can serve the same purposes for different projects.

Having elaborated the importance of external forces for understanding humans' quasi-economic agency, Chapter 4 returns to the domain of individual decision-making—it asks: how do interdisciplinary approaches to decision research integrate psychological insights with economic methods? And, what are the conceptual and ontological challenges of such integration? Herein I critically examine how multiple-self models of intrapersonal and intertemporal choice have been integrated with dual-process and dual-system theories from social psychology and cognitive science. *Multiple-self models* of intrapersonal and intertemporal choice emerged in decision theory and game theory to help economists better understand the dynamics of internal conflict and to predict—and hopefully explain—choice anomalies and inconsistencies that arise over time. *Dual-process theories* of reasoning and judgment are another means of capturing internal conflict. It allows researchers to discern "higher" cognitive processing, which are associated with deliberative judgments and the ability to reason logically, from "lower", more primitive information processing, which is usually associated with affective states and visceral responses. I adopt the term 'multi-agent model' to denote models which conceive of multiple agents with contrasting psychological abilities. Such models seem to be growing in popularity given their purported

ability to predict and explain reasoning errors and decision anomalies due to internal conflict or lack of self-control. I analyze how multi-agent models conceive of and employ "selves" and "systems" for the purposes of representing intrapersonal and intraneural conflict. The chapter is structured according to three claims. The first and second claims establish that multi-agent models are conceptually as well as ontologically ambiguous. The third claim argues that such ambiguities can lead to problems in scientific understanding. The examination of multi-agent models is not only critical to understanding economists and psychologists jointly interpret and model self-control problems, but it further presents an important opportunity to study the effects of cross-disciplinary pollination of concepts and theories.

Chapter 5 builds on the arguments ventured in Chapter 4 and confronts the success story of behavioral economics. It investigates the role that dual process theory (DPT) has played in behavioral economics, and it questions what the descriptive limitations of psychologically dualistic models are. Cognitive scientists and philosophical psychologists alike have criticized the theoretical foundations of the standard view of dual process theory and have argued against the validity and relevance of evidence used to support it. Moreover, recent modifications of dual process theory in light of these criticisms have generated additional concerns regarding its applicability and irrefutability. I argue that this should raise concerns for behavioral economists who see dual process theory as providing psychologically realistic foundations for their models. In particular, it raises the possibility that dualistic models are not as descriptively accurate or reliable as behavioral economists presume them to be. In fact, the case can be made that the popularity of dual process theory in behavioral economics has less to do with the empirical success of dualistic models, and more to do with the convenience that the dualism narrative provides economists looking to sort out decision anomalies. I will argue that the growing number of criticisms against DPT leaves behavioral economists with something of a dilemma: either they stick to their purported ambitions to give a realistic description of human decision-making and modify their use of DPT, or they stick to DPT and modify their ambitions.

In Chapter 6 I offer concluding remarks and consider where one goes from here. First, Chapters 2 – 5 project two main approaches to reconciling the tension between agency and choice. One approach views individual persons as the primary objects of study for economics, and as such, psychology and neuroscience can help locate a more appropriate locus for the study of choice. The second approach views individual persons not as the primary object of study, (economic agents are the primary study, and they are ontologically distinct from persons). As such, choice should be construed as the outcome of external (market) pressures, which include important socio-cognitive supports. Hence, for each of these approaches, there are new pursuits and new philosophical questions to be considered.

Curriculum Vitae

EDUCATION

Erasmus University Rotterdam • Rotterdam, Netherlands • Thesis submitted Nov 19, 2018
PhD in Philosophy: Erasmus Institute for Philosophy & Economics (EIPE)
Title: *Agency & choice*
Supervised by: Prof.dr. Jack Vromen and Dr. Conrad Heilmann
Defense date: February 8, 2019

Erasmus University Rotterdam • Rotterdam, Netherlands • 2014
Research Master in Philosophy of Economics:
Erasmus Institute for Philosophy & Economics (EIPE)
Grade: 8.1

San Jose State University • San Jose, California, USA • 2011
MA in Philosophy: Department of Philosophy
Grade: 3.9

Humboldt State University • Arcata, California, USA • 2009
BA in Philosophy: Department of Philosophy
Grade: 3.6

AREAS OF SPECIALIZATION AND COMPETENCE

Specialization: Philosophy of science (economics, cognitive science), philosophy of mind, theories of bounded rationality, rational choice theory

Competence: Decision theory, game theory, ethical aspects of economics, logic, philosophy of language

PUBLICATIONS

Forthcoming	"From selves to systems: on the intrapersonal and intraneural dynamics of decision-making" (*Journal of Economic Methodology*)
2017	Grayot, J., (2017). The Quasi-Economic Agency of Human Selves. *Œconomia. History, Methodology, Philosophy,* (7-4), pp.481-511.

2016 Ross, D. & Grayot, J., (2016). Neural networks, real patterns, and the mathematics of constrained optimization: an interview with Don Ross. *Erasmus Journal for Philosophy and Economics*, 9(1), pp.142-155.

2014 Grayot, J., (2014). "(Mis)understanding adaptive preferences" - Research master thesis manuscript. *Erasmus University thesis repository*.

SELECTED TALKS & PRESENTATIONS

2017 "From selves to systems: on the refinement and integration of dualistic models of rational choice". *Interdisciplinary Perspectives on Behavioral Economics Workshop*, University of Helsinki, May 22-23

"From selves to systems: on the refinement and integration of dualistic models of rational choice". *5th Annual Nordic Network for Philosophy of Science*, University of Copenhagen, April 20-21

"Cognitivism in behavioral decision research: on the computational basis of multi-process accounts of cognition". *EIPE 20th Anniversary Conference*, Erasmus University Rotterdam, March 22-24

2016 "From selves to systems: implications for the refinement of economic rationality". *Young Scholars Initiative Plenary*, Central European University, Oct. 19-23

"The quasi-agency of human selves". *GRAT Graduate Conference in Philosophy of Science*, University of Antwerp, September 14

"Folk psychology and the human sociocognitive syndrome". *2nd Conference for the Int'l Association for Cognitive Semiotics*, Maria Curie-Sklodowska University, June 20-22

"Human selves as quasi-economic agents". *GREQAM International Conference of Economic Philosophy*, Aix-en-Provence, June 15-16

"Human selves as quasi-economic agents". *Graduate Conference in Theoretical Philosophy*, Twente University, April 20-22

2015 "What are we really talking about when we talk about folk psychology?". *Annual Conference for the Nederlandse Onderzoeksschool Wijsbegeerte*, Vrije University Amsterdam, Dec.12

"Folk psychology and the human sociocognitive syndrome". *Lunchtime Seminar Series*, Erasmus University Rotterdam, Dec. 2

"Minded agents, economic agents, and ordinary folk". *Conference 12th International Network for Economic Method*, University of Cape Town, South Africa, Nov. 18-23

"What are we really talking about when we talk about folk psychology?". *EIPE PhD Seminar Series*, Erasmus University Rotterdam, Oct. 15

"Negative adaptations: on the cognitive foundations of adaptive preference formation". *TilPS Graduate Student Conference, Philosophy of Science and Logic*, Tilburg University, Feb. 27

2014 "(Mis)understanding adaptive preferences: where description begins and implication ends". *EIPE PhD Seminar Series*, Erasmus University Rotterdam, March 17

TEACHING AND COURSE DEVELOPMENT

Part-time Lecturer • Rutgers University • August 16, 2017 - *Current*
Course instructor, "Introduction to formal reasoning and decision-making" (Phil 109 online). Terms completed: Fall 2017, Spring 2018, Fall 2018.

Docent • Erasmus University Rotterdam • Oct 1 – Dec 31, 2018
The position facilitates the development of "Philosophy of Science III", to be instructed by Dr. Tim de Mey of the Erasmus School of Philosophy.

Guest Lecturer • Erasmus University Rotterdam • Oct 1 – Dec 21, 2018
Philosophy of Science 1 (for Dr. Tim de Mey) and Philosophy of Science 2 (for Prof.dr. F.A. Muller) at the Erasmus School of Philosophy. Topics covered include "Values in Science" and "Theory".

Guest Lecturer • Vrije University Amsterdam • June 4 – 7, 2018
Two lectures and two workshops for the course "PPE in Practice" (PiP 2) at the John Stuart Mill College. Topics covered include: "Time Discounting and Climate Science".

Tutorial Instructor • Erasmus University Rotterdam • May 30 – June 13, 2018
Tutorial on "Philosophy of Economics" provided to the bachelor students at the Erasmus School of Economics. Instruction includes lecturing, exam preparations, and corrections. Previous tutorials given on, June 2 – June 16, 2017; May 19 – June 9, 2016; May 27 – June 17, 2015.

Workshop Leader • Erasmus University Rotterdam • Dec 14, 2014
Workshop on "nudge" policy and adaptive preference formation. OZSW winter school on "Philosophy, Policy, and Social Science" for research master students.

Non-tenured Lecturer • San Jose State University • Summer term 2011
Course instructor for Phil 57 "Introduction to logic and critical thinking" for California State University undergraduate degree.

Instructional Student Assistant • San Jose State University • Spring term 2011
Course instructor for Phil 57 "Introduction to logic and critical thinking" for California State University undergraduate degree.

TRAINING & RESEARCH COURSES

EGS3H Master Class on "The Brain" • May 11, 2017
Designed for PhD candidates in the Erasmus Graduate School of Social Sciences and Humanities (EGS3H); the course cuts across disciplines in which distinguished scholars present their research on the theme.

Young Scholars Initiative Plenary • Central European University • Oct. 19-23, 2016
In association with Institute for New Economic Thinking, the working group "Piecing together a paradigm" makes explicit the link between philosophy and economics and the relevance of philosophy for economics

EGS3H Master Class on "How to obtain an NWO Veni Grant" • Sept. 2016
Designed for PhD candidates in the Erasmus Graduate School of Social Sciences and Humanities (EGS3H); Courses give hands-on advice and guidance to post-doctoral researchers applying to the NWO Veni.

Course in Epistemic Game Theory 2016 • Maastricht University • June 5 – 19, 2016
Two-week intensive course on epistemic game theory at The EpiCenter; includes written exams, lectures on theoretical aspects of game theory, and exercise sessions.

EGS3H Master Class on "How to get your article published" • April 2016
Designed for PhD candidates in the Erasmus Graduate School of Social Sciences and Humanities (EGS3H); Courses cover techniques for planning, writing, and submitting papers for publication.

Winter School "Agents & Agency" • University of Groningen • Jan. 25-26, 2015
Two days of lectures pertaining to the formal models of and conceptual foundations behind agents and agency, with special emphasis on the ethics of agency and theories of agent psychology. Provided by OZSW.

Young Scholars Initiative workshop • University of Cape Town • Nov. 18, 2016
In association with the Institute for New Economic Thinking, the YSI pre-conference workshop on topics in philosophy of economics. Guest lectures by Julian Reiss and Till Grüne-Yanoff.

Workshop on Neurophilosophy with Patricia Churchland • Oct. 26, 2015
Participant in a special workshop with Patricia Churchland, including lecture and private Q & A session t Erasmus University Rotterdam, led by Dr. Tim de May and Maureen Sie.

Erasmus Philosophy Lecture 2015 with Alva Noë • May 13, 2015
Two private seminars with Alva Noë based on his works *Out of Our Heads* and *Varieties of Presence*. Courses were offered to PhDs as part of the fourth edition of the yearly Erasmus Philosophy Lecture.

Foundations of Cognitive Science: Mental Representation • Sept 21-23, 2015
Two days of lectures on the role of mental representation in human cognition at the Ruhr-Universität, Bochum, with special emphasis on action, intentionality, perception, and emotion.

Cursos de Verano: Summer School in Philosophy & Economics • July 6-8, 2015
Three days of lectures on micro-foundations in macro-economics at San Sebastian, Spain. Hosted by the Centre for Humanities Engaging Science and Society (CHESS) & International Network for Economic Method (INEM).

STUDENT SERVICE

Erasmus Journal of Philosophy & Economics • Sept. 2013 - *Current*
Copy-editor and guest referee. Over 50 articles edited and proofread.

PhD Council representative for Department of Philosophy • Sept. 2016 - 17
Representative for the Erasmus Graduate School of Social Sciences and Humanities (EGS3H) at Erasmus University Rotterdam

EIPE Research Assistant • Erasmus University Rotterdam • 2013 – 2014
Assistant for Dr. Conrad Heilmann, co-director of Erasmus Institute for Philosophy & Economics.

EIPE Student Representative • Erasmus University Rotterdam • 2013-2014
Organizer and liaison for research master student affairs within the department of theoretical philosophy.

Formal Ethics Conference • Erasmus University Rotterdam • May 30-31, 2014
Part of the organizing committee hosted by Erasmus Institute for Philosophy & Economics.

OZSW Conference • Erasmus University Rotterdam • Nov. 15-16, 2013
Part of the organizing committee hosted by Faculteit der Wijsbegeerte.

11th INEM Conference • Erasmus University Rotterdam • June 13-15, 2013
Part of the organizing committee hosted by Erasmus Institute for Philosophy & Economics.

Instructional Student Assistant • San Jose State University • 2010 – 2011
Teaching assistant for Dr. Karin Brown, Professor of philosophy. Courses assisted include: PHIL 010 “Introduction to Philosophy”; PHIL 070A “Ancient Philosophy”; PHIL 061 “Moral Issues”.

AWARDS AND DISTINCTIONS

Award for Multi-disciplinary excellence • 2016
Awarded for the most "creative and feasible" research proposal in the Dean's Masterclass on "the brain" at Erasmus University Rotterdam.

Non-EU student fee waiver • Erasmus University Rotterdam • 2012 - 2014
Scholarship awarded to non-resident students without external funding.

"Outstanding Graduate Student of the Year" Award • May 5th, 2011
Awarded to one MA student per year based on GPA, thesis, and/or involvement in course development at San Jose State University.

"Trial by Fire" Award for Graduate (ISA) Teaching • May 5th, 2011
Awarded to instructors who receive at least a mean score of 4 (out of 5) on cumulative student evaluations at San Jose State University.

REFERENCES

Prof.dr Jack Vromen
Faculty of Philosophy
Academic Director of EIPE
Erasmus University Rotterdam
vromen@esphil.edu.nl

Dr. Conrad Heilmann
Assoc. Professor of Philosophy
Co-director of EIPE
Erasmus University Rotterdam
heilmann@esphil.edu.nl

Dr. Elizabeth Camp
Rutgers University
Undergraduate Director
emc233@philosophy.rutgers.edu

Dr. Anand Vaidya
San Jose State University,
Professor of Philosophy
Anand.vaidya@sjsu.edu