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

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
Doctoral Committee:

Promoters:
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
# Table of Contents

**Preface** ........................................................................................................................................ viii

< preface removed from RePub submission >

1 **Introduction** ........................................................................................................................................ 1

1.1 Agency & Choice ................................................................................................................................. 1
1.2 Economics meets psychology and cognitive science ........................................................................... 3
1.3 Four questions about agency and choice ............................................................................................. 3
   1.3.1 On the curious role of mental states in economics ..................................................................... 4
   1.3.2 Individualism versus anti-individualism: an ontological debate ........................................... 7
   1.3.3 Mixing metaphors: dual-selves and dual-processes .................................................................. 8
   1.3.4 Why is two the magic number? .................................................................................................. 10
1.4 Outlook ............................................................................................................................................... 12
1.5 Bibliography ....................................................................................................................................... 12

2 **Two problems with the mentalism-behaviorism dichotomy in economics** ...................................... 17

< chapter removed under temporary embargo >

3 **The quasi-economic agency of human selves** ................................................................................... 18

3.1 Introduction .......................................................................................................................................... 18
3.2 Non-anthropocentric neoclassicism and multiple-selves ................................................................... 20
   3.2.1 Economic agency in an anti-individualistic economics ............................................................ 20
   3.2.2 Three interpretations of selves .................................................................................................. 23
3.3 Social-determination, black boxes, and the externality of intentions ................................................. 29
   3.3.1 Game-determination .................................................................................................................. 30
   3.3.2 Selves as black boxes ................................................................................................................ 33
   3.3.3 Externalizing intentionality ....................................................................................................... 35
3.4 Social selves versus sub-personal selves .............................................................................................. 37
   3.4.1 Against the view that selves are sub-personal .......................................................................... 37
   3.4.2 Neuroscientific control theory and participatory sense-making ............................................. 41
3.5 Concluding remarks .............................................................................................................................. 43
3.6 Bibliography ....................................................................................................................................... 44
Chapter 4: From selves to systems: on the intrapersonal and intraneural dynamics of decision making .................................................................48
  4.1 Introduction ..................................................................................48
  4.2 The emergence of multi-agent models: a brief overview ..................50
    4.2.1 From selves to systems and back ..............................................50
    4.2.2 Conceptual ambiguity surrounding selves and systems ..........53
  4.3 Agency and ontological ambiguity ..................................................55
    4.3.1 The uncertain agency of selves and systems ............................55
    4.3.2 Two types of ontological ambiguity .........................................56
  4.4 Three examples of multi-agent models in behavioral decision research 57
    4.4.1 A model of heuristic judgment (Kahneman & Frederick 2005) ....58
    4.4.2 The brain as hierarchical organization (Brocas & Carrillo 2008a) ....61
    4.4.3 Deliberative vs. affective systems (Loewenstein & O’Donoghue 2005) .....64
  4.5 Implications for scientific understanding .........................................66
    4.5.1 What does the model purport to explain? ...............................66
    4.5.2 How does the model achieve this goal? .....................................67
    4.5.3 Rebuttals and reconsiderations ..............................................69
  4.6 Concluding remarks ........................................................................71
  4.7 Bibliography ..................................................................................72

Chapter 5: Why behavioral economics needs to revise its faith in dual process theory 106

Chapter 6: Looking back and looking ahead ........................................79
  6.1 Looking back ..............................................................................79
  6.2 Looking ahead .............................................................................79
  6.3 Bibliography .................................................................................81

Samenvatting ......................................................................................83
Summary .............................................................................................87
Curriculum Vitae ................................................................................90
< section reserved for preface >
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. 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

---

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 unlikely 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 1970’s 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, 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.

---

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).

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


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.


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


Chapter 2
Two problems for the mentalism-behaviorism dichotomy in economics

< chapter removed under temporary embargo >
Chapter 3
The quasi-economic agency of human selves

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. *Economia. 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 con-
tinuously 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 indi-
vidual person that is bound by social and institutional constraints (2005). In this re-
gard, 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 sta-
bilization 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 convo-
luted, account of selves than others discussed in the economics literature.

In this paper, I investigate the roles that selves play within Ross’s anti-individu-
alistic 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 be-
lieved to be. To this end, I distinguish three different roles for selves—these are evolu-
tionary, narrative, and economic—and I argue that these roles contribute to two
distinct, but overlapping, projects in Ross’s broader philosophy of economics. 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 accord-
ance 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”, “nar-
rative”, and “economics”, sometimes these labels will coincide with alternative labels. For in-
stance, the evolutionary role is sometimes described as “biological”; the narrative role is some-
times described as “biographical”, and the economic role is sometimes described as “mathe-
matical”.

19
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

---
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. 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.

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.

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 enculturated 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 micro-economic 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. 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* 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* so that they become new agents who, still in S*, will play G*. (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 implies 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 neu-
roeconomics, a close examination of his (2006) and (2007b) articles on the evolu-
tionary 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 activ-
ity:

…human personalities—selves, that is—have been made phylogenetically possible
and normatively central through the environmental manipulations achieved collec-
tively 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 communi-
ties. 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 compen-
sating 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 achieve-
ment 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. 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 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. 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.

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 asssenting to an over-arching theory of mental

---

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).
representation (which would require linguists and cognitive scientists to figure out how people “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


Chapter 4
From selves to systems: On the intrapersonal and intraneural dynamics of decision making

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 Malecka 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. 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. 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. 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).

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).

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 \( u'' < 0 \). \( \theta_t \) 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 \( \theta_t \). The principal weighs the utility of all agents under a budget constraint that links lifetime consumption and lifetime labor

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

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 \( \theta_t \) 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:

\[
\{c_t(\theta_t), n_t(\theta_t)\}_{\theta_t \in \Theta}
\]

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 \( \theta_t \), 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 $\theta_t$, 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.\(^\text{10}\) 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 an 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 in 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:

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

where $h(W,\sigma) \ast [M(x^A, a) - M(x,a)]$ 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 indica-
tion 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 delibera-
tive 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:

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

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 Kahne-
man & Frederick (but like Brocas & Carrillo), Loewenstein & O’Donoghue portray
the interaction of systems by way of a principal-agent formalism, not by explicit de-
scriptions of either functional or physiological processes. The methodology for de-
picting 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
provide 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). 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 (striatal / 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


Chapter 5
Why behavioral economics needs to revise
its faith in dual process theories

< chapter removed under temporary embargo >
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


**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 economen 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
gedragseconomie 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.
< section reserved for curriculum vitae >