8-6-2007

Dual-Process Theories and the Rationality Debate: Contributions from Cognitive Neuroscience

Trevor Hannesson Kvaran

DUAL-PROCESS THEORIES AND THE RATIONALITY DEBATE:

CONTRIBUTIONS FROM COGNITIVE NEUROSCIENCE

by

TREVOR KVARAN

Under the direction of Eddy Nahmias and Andrea Scarantino

ABSTRACT

The past 40 years have seen an enormous amount of research aimed at investigating human reasoning and decision-making abilities. This research has led to an extended debate about the extent to which humans meet the standards of normative theories of rationality. Recently, it has been proposed that dual-process theories, which posit that there are two distinct types of cognitive systems, offer a way to resolve this debate over human rationality. I will propose that the two systems of dual-process theories are best understood as investigative kinds. I will then examine recent empirical research from the cognitive neuroscience of decision-making that lends empirical support to the theoretical claims of dual-process theorists. I will lastly argue that dual-process theories not only offer an explanation for much of the conflicting data seen in decision-making and reasoning research, but that they ultimately offer reason to be optimistic about the prospects of human rationality.

INDEX WORDS: Rationality, Neuroeconomics, Decision-Making, Investigative Kinds, Cognitive Neuroscience

DUAL-PROCESS THEORIES AND THE RATIONALITY DEBATE:
CONTRIBUTIONS FROM COGNITIVE NEUROSCIENCE

by

TREVOR KVARAN

A Thesis submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2007

DUAL-PROCESS THEORIES AND THE RATIONALITY DEBATE: CONTRIBUTIONS FROM COGNITIVE NEUROSCIENCE

by

TREVOR KVARAN

Major Professor:	Eddy Nahmias
	Andrea Scarantino
Committee:	Erin McClure

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2007

For my wife, Kristen,

and my parents,

Hannes and Pennie.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisors, Dr. Eddy Nahmias and Dr. Andrea Scarantino, both for their guidance and support throughout this project and also for allowing me the opportunity to be one of the first Brains and Behavior fellows from Georgia State University’s philosophy department. I would next like to thank Dr. Erin McClure, not only for agreeing to be a committee member for this thesis, but also for allowing me the opportunity to work in her lab and for putting up with endless questions about statistics and neuroimaging techniques. I would also like to thank my undergraduate professors, Dr. David Sherry and Michael Malone, for their encouragement early in my educational career. I would like to thank Bryan Miller for his helpful comments on this thesis and for navigating the uncharted waters of the Brains and Behavior program with me. Lastly, I would like to thank my wife, Kristen, and my parents, Hannes and Pennie, for all of their support over the course of the last two years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC: Anterior Cingulate Cortex

AI: Anterior Insula

dACC: Dorsal Anterior Cingulate Cortex

DLPFC: Dorsolateral Prefrontal Cortex

FMRI: functional magnetic resonance imaging

LPFC: Lateral Prefrontal Cortex

LTC: Lateral Temporal Cortex

MTL: Medial Temporal Lobe (MTL)

PPC: Posterior Parietal Cortex

rACC: Rostral Anterior Cingulate Cortex

VMPFC: Ventromedial Prefrontal Cortex

Chapter 1. Introduction

Imagine that you and a friend find a stack of one hundred $1 bills. Imagine further that you come up with a plan to split the money. The two of you agree that your friend will split the hundred dollars however he chooses between the two of you, but that you will be able to either accept or reject his offer. If you accept the offer, the money will be divided between the two of you however your friend has divvied it up, but if you reject the offer, you will both put the money back where you found it and both leave with nothing. Now imagine that the friend has just made his offer: $90 for him and $10 for you. Do you accept this offer and leave with $10 that you didn't have before, or reject the offer and leave with nothing? If you are like most people, you will reject the offer. And, according to some strict conceptions of economic rationality, you would be behaving irrationally when doing so—you would essentially be giving away $10.

Whether or to what extent humans can fulfill the rationality requirements proposed by various normative models of human reasoning and decision-making has been a matter of considerable controversy at least since the pioneering work of Daniel Kahneman and Amos Tversky (Tversky & Kahneman 1974). Many have argued that empirical studies have shown that humans are by and large irrational creatures, while others have mounted strong arguments which attempt to explain away the apparently threatening evidence by showing that human reasoning and decision-making is not too far from perfectly rational.

Much of this debate is a reaction to the large body of research that has followed in the wake of Kahneman and Tversky's initial studies, which has come to be collectively known as the *heuristics and biases approach*. In response to this research, which has been seen by some as having “bleak implications for human rationality” (Nisbett & Borgida 1975), evolutionary psychologists and philosophers have mounted their own empirical and conceptual projects, aimed at showing that humans are in fact very close to optimally rational. These two opposed views have generated an enormous amount of seemingly contradictory data. Recently, several researchers have posited that dual-process theories of reasoning and decision-making, which propose that there are two distinct types of systems underlying human cognition, may offer a way to make sense of these conflicting results (Evans & Over 1996, Stanovich 1999). Although dual-process theories (detailed in chapter 3) offer much promise for resolving many of the conflicts in the rationality debate, they currently suffer from a lack of clarity about what exactly dual-process theories are. Although some progress has been made (See Frankish 2004, Samuels forthcoming), very little work has been done explicating the criteria for membership of the two systems of cognitive processing that comprise dual-process theories. Additionally, it has been left unclear exactly how the two systems interact with each other to bring about behavior (though see Evans 2003 and Haidt 2001 for two attempts).

In Chapter 2 I will outline the primary positions within the rationality debate. Divergences in human performance from the solutions deemed optimal according to various normative theories have led to worries over the extent to which humans are rational. In section 2.1 I will present the normative/descriptive distinction that theorists and empirical researchers appeal to when determining whether humans are rational or not. In section 2.2 I will detail some of the research done in the heuristics and biases tradition, which has led many theorists to the pessimistic conclusion that humans are systematically irrational. These pessimistic conclusions have been challenged on several fronts. These challenges will be the focus of section 2.3. I will conclude Chapter 2 by suggesting that there is a middle ground between the two traditional positions in the debate, and that this position is supported by dual-process theories of reasoning and decision-making. Dual-process theories posit that humans have two distinct types of cognition, neutrally termed System 1 and System 2. System 1 tends to operate quickly, automatically, and developed early in our evolutionary history. System 2, by contrast, operates slowly, is consciously controlled, and developed relatively recently in our evolutionary history.

In Chapter 3 I will advance this response by examining how best to conceptualize the two systems of dual-process theories. I will first provide a fairly detailed explication of the two types of cognitive processing proposed by dual processes theories (sections 3.1 and 3.2). Then I will propose that these two systems should be considered *investigative kinds*, which are similar in many respects to *natural kinds*.

Following this conceptual examination, I will turn in Chapter 4 to recent evidence from various areas of cognitive neuroscience, particularly research done in the developing field of neuroeconomics. Some have argued that neuroimaging studies of reasoning and decision-making tasks have provided substantial evidence for dual-process theories. I will argue that a cognitive neuroscience approach to decision-making research can be highly fruitful and has much to contribute to the study of dual-process theories and more broadly to our understanding of human rationality.

In Chapter 5 I will conclude by pointing to some of the broad implications of a dual-process theory of decision-making, highlighting in particular the sense in which System 1 and System 2 are aimed at different and sometimes conflicting domains of rationality. Given that the two systems in many cases are in conflict with one another, I will argue that a new focus in the normative study of human rationality should be to determine when people should respond in accordance with system 1 processes and when people should respond in accordance with system 2. Correspondingly, an important descripitive project will become determining the extent to which people are able to select between the two systems.

Chapter 2: The Rationality Debate

Although it has been a crucial concept throughout the history of philosophy, psychology, economics, and several other fields,[1] the appropriate conception of rationality has been a notoriously controversial issue. However, as Edward Stein (1996) has pointed out, most researchers adopt what he calls the “Standard Picture” of rationality. According to the Standard Picture, to be rational is to reason and make decisions that accord with the rules of formal logic, probability theory, decision theory and other relevant formal systems. As Stein points out, “if the standard picture of reasoning is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with” (1996: 4). Although Stein is right that this project is by and large the default position within the debate over human rationality, as we will see in section 2.3 and later in Chapter 5, this conception of rationality is far from universally accepted.

Starting in the early 1970s, psychologists began to empirically investigate whether or to what extent human decision-making actually does accord with the standards of these normative systems. In doing so, they took themselves to be investigating the extent to which humans could be considered rational. The study of human reasoning and decision-making became a thriving area of research in psychology, and its implications were quickly recognized as a potential threat to traditional philosophical methodology. Philosopher Jonathan Cohen diagnosed the problem for philosophy when he wrote:

[1] See chapters 15-22 in Mele & Rawling (2004) for discussion of the role of rationality in specific domains, such as gender studies, science, economics, and the law.

> If psychologists were right, what they say would seriously discredit the claims of intuition to provide— other things being equal— dependable foundations for inductive reasoning in analytical philosophy….Since at many points analytical philosophy has to premise its accounts of reasoning on the data of human intuitions, its metaphilosophy can hardly afford to ignore this extensive literature. (1986: 149-150)

As Cohen sees it, the question of whether people can rightly be thought of as rational is a central problem for philosophy. Without rationality, philosophy as many in the analytic tradition practice it would have to be abandoned or largely reconsidered. Cohen and several others quickly mounted responses to this potential threat, and in large part the rationality debate has revolved around whether these defenses adequately secured the prospect of human rationality. These defenses have included both philosophical arguments and empirical research within evolutionary psychology.

2.1. The Normative/Descriptive/Prescriptive Distinction

The study of rationality is typically conceived in terms of three interrelated projects, each investigating decision-making at a different level. The *descriptive* project is concerned with empirically investigating the responses that humans actually give when reasoning or making decisions. The goal of the descriptive project is to create empirically adequate models which predict with a high degree of success the responses humans will give when presented with various reasoning and decision-making tasks. A descriptive model is therefore successful to the degree that it allows us to predict, given specific inputs and background conditions, how a person will reason or the choices they will make. Traditionally the descriptive project has involved the development of these models in terms of cognitive mechanisms based on evidence from behavioral responses, but as we will see in Chapter 4, cognitive neuroscience research is beginning to allow descriptive models to incorporate evidence about the processing done in the human brain (Montague 2006, Camerer, Lowenstein, and Prelac 2005, Chorvat and McCabe 2005).

The *normative* project, on the other hand, concerns itself with developing systems of reasoning that, if adhered to, will lead to optimal reasoning. Normative systems tell us how we ought to reason and the choices we should make if we hope to reason as well as possible. For example, many argue that expected utility and rational choice theory offer normative models for practical decision-making, and formal logic and probability theory do the same for reasoning and argumentation[2]. Much of the motivation behind the work done in the descriptive project has been to question whether people do, as a matter of fact, perform according to these normative models.

Although many theorists have claimed that empirical research has shown that human decision-making falls short of normative ideals, for reasons I will now address, it is highly controversial whether these shortcomings should be interpreted as proof of human irrationality. In addition to the normative ideals set forth by formal systems and the descriptive research of psychologists and neuroscientists, the extent to which humans can be considered rational also depends on the limits of human cognitive capabilities. As Keith Stanovich has pointed out:

> As interesting as such divergences between normative and descriptive models are, they cannot automatically be interpreted as instances of human irrationality. This is because judgments about the rationality of

[2] Briefly, according to expected utility theory, an agent's choice should be made by comparing the weighted averages of each potential outcome of a choice. Rational Choice Theory says that humans have a stable and complete set of preferences, and that humans should maximize their utility in light of these preferences.

> actions and beliefs must take into account the resource-limited nature of the human cognitive apparatus. (1999: 3)

Determining the upper limits of our cognitive capabilities and developing systems of reasoning and decision-making that optimize our reasoning given these limits is the work of the prescriptive project. The guiding assumption of the prescriptive project is that humans cannot be expected to live up to standards of rationality that are impossible for them to meet. As Stephen Stich has put it: "it seems simply perverse to judge that subjects are doing a bad job of reasoning because they are not using a strategy that requires a brain the size of a blimp" (1990: 27).

Philosophically, the prescriptive project is often seen as requiring the acceptance of an *ought-implies-can principle*, the basic idea being that we cannot be expected to do something that we cannot actually do. If this principle is accepted, then typically it will follow that the normative requirements placed on an agent will be highly constrained both by the cognitive abilities of that agent and by the environment the agent is in. As Samuels, Stich, and Faucher point out:

> Human beings do not have the same capacities as God or a Laplacian demon, and other (actual or possible) beings – e.g. great apes – may well have reasoning capacities that fall far short of those possessed by ordinary humans. In which case, if ought implies can, then there may be normative standards that one kind of being is obliged to satisfy where another is not. (2004: 42)

If the aims of the prescriptive project are taken seriously, then developing a thorough understanding of the cognitive abilities of humans takes on an added importance, for it not only provides us with purely descriptive information, but will come to have a significant impact upon the appropriate norms by which to judge the rationality of agents' reasoning and decision-making. Norms of decision-making and reasoning will become relativized or bounded by the limitations of the particular agent. If humans turn out to fall far short of the cognitive capacities required for the norms prescribed by probability theory and rational choice theory, then it may be that normative models of decision-making ultimately come to have very little relevance to most humans' everyday lives. If it turns out that we *are* able to achieve something like the norms of rational decision-making, then the normative models will be relevant to human life.

Reasoning and making decisions well is important because it leads to higher degrees of goal satisfaction (Stanovich 1999), and if it turns out that humans are capable of making optimal choices (or close to them), then the solutions uncovered through work in the normative project will be an important tool. The norms of decision-making will also become the prescripts for humans. If humans turn out to be considerably impoverished in their cognitive capabilities due to the limits of human psychology, then the normative theories may end up being less relevant to human decision-making. We would still need to devise methods to get our reasoning as close to normatively optimal as possible, but these could very well turn out to be so substantially removed from the normative solutions as to make the normative models unimportant.[3]

Even if the normative model ended up having very little in common with the prescriptive model best suited for humans, the question of optimal decision-making would still be an important question. It is certainly not out of the question that other types of systems, particularly artificial systems such as computers, might be able to meet

[3] In this sense even if our prescriptive models are significantly different from optimality, they will still need to take the normative project into account.

the standards of the normative models. Jonathan Baron (1985) makes this point when he writes:

> Although normative models may not provide] a good prescriptive model for ordinary people, it can be seen as a good prescriptive model for some sort of idealized creature who lacks some constraint that people really have.… We may thus think of normative models as prescriptive models for idealized creatures.… A good prescriptive model takes into account the very constraints that a normative model is free to ignore. (8-11)

These three distinct but interrelated projects illustrate the inherently interdisciplinary nature of the study of human rationality. The determination of whether humans are rational or not involves both empirical research of human cognition as well as theoretical and philosophical work aimed at determining both optimal reasoning and decision-making as well as what the implications of descriptive findings are for the norms which are ultimately prescribed for humans. In the following sections two approaches to the empirical study of reasoning and decision-making and the conclusions (one highly pessimistic of human rationality, the other highly optimistic) that have been made based on the findings will be addressed. These two positions have historically been opposed to one another. The rationality debate is by and large a debate about which of these two positions is correct. After detailing these two views, I will suggest that in fact these are not the only two positions available within the rationality debate, and that there is a third option which is neither as pessimistic nor as optimistic as the two traditional positions within the debate.

2.2. A Reason for Pessimism? The Heuristics and Biases Approach

Much of the debate surrounding human rationality has been a response to a body of research collectively known as the heuristics and biases program. Across a wide range of empirical studies, humans have been found to perform far below the normative standards of decision-making and reasoning (Kahneman, Slovic, & Tversky 1982, Kahneman &Tversky 2000, and Gilovich, Griffin, & Kahneman 2002). Based on this research, followers of the heuristics and biases tradition have argued that humans should be considered irrational, or at minimum that the assumption of rationality made by many philosophers and economists is dubious (Samuels & Stich 2004). After briefly illustrating the type of research that falls under the heuristics and biases heading, I will describe the theoretical claims offered in response to these findings.

One of the earliest and most studied problems in the reasoning and decision-making literature is Peter Wason's *selection task* (Wason 1966). In this famous task, participants are presented with four cards, two of which show a letter (one vowel and one consonant) and two of which show a number (one even number and one odd number). Participants are told that each card has a letter on one side and a number on the other. Participants are then told to indicate which cards they would need to flip over in order to test the rule: "if a card has a vowel on one side, then it has an odd number on the other side." Most participants correctly selected the vowel to turn over, but most also said that the odd number needed to be turned over, which is incorrect. Moreover, most participants failed to say that the even numbered card would need to be turned over. This result, which has been replicated numerous times, has been taken as evidence to support the claim that people systematically deviate from the principles of formal logic. As we will see in section 2.3 however, several alternative explanations of this divergence from the normative solution have been offered.

Within more explicit decision-making research, similar troubling results have been found. Take, for instance, Tversky and Kahneman's (1981) study in which participants were given one of two decision tasks. One group of participants was presented with the following problem:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate is as follows:
>
> If program A is adopted, 200 people will be saved.
>
> If program B is adopted, there is a 1/3 probability that 600 people will be saved and a 2/3 probability that no people will be saved.

A second group of participants was given a slightly different version that read:

> If program C is adopted, 400 people will die.
>
> If program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

When asked which plan to institute, a large majority in the first group chose program A, while participants in the second group heavily favored program D. This intriguing study is troubling for two reasons. The probabilities of the outcomes in the first two programs are the exact same as the probabilities of the outcomes in the second two programs. That is to say, the choice the first group is asked to make is exactly the same as the choice the second group is asked to make, and yet we see that participants respond significantly differently depending on how the question is framed. The task takes on an additionally problematic angle because within each group, the expected values of the two choices in terms of lives saved or lost are identical. If participants make decisions based on expected utility, then it would be expected that since there is no difference between the expected utility of the choices, responses would be split close to evenly between the two choices. The fact that in each group we instead see a significant majority favoring one choice over the other is cause to worry that people do not conform to the tenets of the theory of expected utility. Tversky and Kahneman argued that these results provided strong evidence that *systematically* people do not meet the normatively predicted solutions for these problems, and that the way in which choices are framed has a significant effect on the decisions people make. They famously concluded that people's decisions are guided not by formal principles like those of logic and decision-theory, but instead by heuristics, simple rules which rely on contextual clues and allow for fast, efficient, but often sub-optimal choices.

These two examples illustrate the general project and typical findings of heuristics and biases research. In a number of different tasks, and in comparison to many normative models, from formal logic to decision theory, humans appear to underperform relative to the normatively predicted solutions. Based on these kinds of results, several increasingly strong conclusions have been drawn by proponents of this research. Samuels and Stich (2004) have identified these claims, from least to most controversial, as:

> (1) People's intuitive judgments on a large number of reasoning and decision-making problems regularly deviate from appropriate norms of rationality.

> (2) Many of the instances in which our judgments and decisions deviate from appropriate norms of rationality are explained by the fact that, in making these judgments and decisions, people rely on heuristics…"which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors.
>
> (3) The *only* cognitive tools that are available to untutored people are normatively problematic heuristics. (2004: 285-286)

As Samuels and Stich point out, while the first two of these conclusions seem at least plausible interpretations of the empirical evidence, there is little reason to believe 3 is true. The heuristics and biases program has certainly uncovered many examples in which humans deviate from normative conceptions of rationality, and positing heuristics as cognitive mechanisms that humans use to make fast decisions is a plausible, although certainly not uncontroversial explanation of why humans perform poorly on these tasks. But, even given these two conclusions, it does not follow that the only thing we have to rely on when making decisions or reasoning are normatively problematic heuristics. These pessimistic claims represent one end of the rationality debate. In the next section, several responses to the heuristics and biases findings will be addressed.

2.3. Panglossian Responses: Philosophical and Empirical Defenses of Human Rationality

Evolutionary psychologists and several philosophers have been eager to defend an optimistic perspective of human reasoning and decision-making abilities. Highlighting this optimism, Stanovich (1999) labeled this position in the debate 'Panglossian.' Two different Panglossian approaches have been taken. The first relies upon philosophical argument to defend the claim that humans are very close to perfectly rational (Cohen

1981; Dennett 1987). Although this version of the Panglossian position is extremely interesting, it has been widely discussed and definitively criticized elsewhere, and is considerably outside the focus of this thesis. As such, it will not be considered in any detail here[4]. Instead, I will focus on a second and more problematic challenge to the pessimistic claims of the heuristics and biases tradition.

The second and perhaps more significant approach to endorsing a Panglossian view of human decision-making has been taken by evolutionary psychologists. As Samuels and Stich have illustrated in several articles (Samuels and Stich 2004, Samuels, Stich, & Bishop 2000), evolutionary psychologists have responded to the findings of the heuristics and biases tradition by mounting a challenge to both the norms that heuristics and biases researchers appeal to and their empirical work. Broadly, evolutionary psychologists have argued that in comparing human performance to the normatively prescribed responses, the heuristics and biases tradition has been applying confused conceptions of probability theory and decision-theory. Moreover, they argue that when you apply the correct norms, people no longer show the dramatic divergences from the normative solutions that the heuristics and biases tradition has reported. Evolutionary psychologists have developed an empirical research program aimed at showing how, by taking into consideration our evolutionary history, it is possible to illustrate both how good humans actually are at reasoning and also why humans perform poorly on specific tasks such as those used in the heuristics and biases studies.

[4] For what I take to be definitive criticisms of the philosophical Panglossianism of Dennett and Cohen, see Stich 1990, Stanovich 1999, Stein 1996, and Thagard & Nisbett 1983.

Evolutionary psychologists begin their research with several guiding assumptions. First, our reasoning and decision-making abilities would have evolved as adaptations to problems faced by our evolutionary ancestors. If this is the case, then humans should perform significantly better on problems presented in a way that our ancestors were likely to have dealt with than on ones which they would not have been presented with. Second, these adaptations would develop as modules in the mind that are designed by evolution to deal with domain specific problems (Samuels & Stich 2004: 288).

Starting from these assumptions, evolutionary psychologists, most prominently Gerd Gigerenzer, Leda Cosmides and John Tooby have argued that people are in fact quite good at meeting many of the normative standards of probability theory, decision theory and formal logic provided the information is presented in a format similar to one with which our ancestors would have been presented. Discussing probability, Cosmides and Tooby write: “Some of our inductive reasoning mechanisms do embody aspects of a calculus of probability, *but they are designed to take frequency information as input and produce frequencies as outputs*” (italics mine 1996: 3). Gigerenzer (1994) has gone even farther, arguing that the heuristics and biases tradition has proved nothing about human rationality because they work from a confused notion of probability. Gigerenzer has defended a ‘frequentist’ probability theory, according to which probability statements are literally meaningless without some sort of reference class, meaning that probabilities cannot be applied to single events. Almost all reasoning studies which test probability performed in the heuristics and biases tradition ask probability questions about single events, which Gigerenzer argues makes them ultimately unthreatening to conceptions of human rationality because they are misapplying the norm they are interested in studying. When combined, the points made by Cosmides and Tooby and Gigerenzer practically guarantee that humans will meet the standards of rationality, for the appropriate *normative* understanding of probability theory is in terms of frequencies, and humans have evolved to be quite good at dealing with probabilities when presented as frequencies (Samuels, Stich, and Bishop 2002). According to evolutionary psychology, the cases that have been presented by researchers in the heuristics and biases tradition have produced virtually meaningless results because their research rests on the wrong normative theory of probability. Remember that according to the Standard Picture of rationality, whether or not agents are rational is a matter of whether their reasoning or decision-making conforms to the standards of various normative theories. If the heuristics and biases researchers have systematically studied the wrong normative systems, then they ultimately have said nothing about human rationality.

To show empirically that people do in fact reason well when both inputs and outputs are presented in terms of frequencies, Cosmides and Tooby (1996) reformulated a task that had been widely used in decision-making research--the Harvard Medical School Problem. The Harvard Medical School Problem was designed to test the Bayesian reasoning abilities of faculty and graduate students at Harvard Medical School. The original formulation ran as follows:

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person’s symptoms or signs? (Casscells, Schoenberger, and Grayboys 1978, 999)

The correct Bayesian response to this problem is 2%, but even when presented to faculty and students of Harvard, only 18% of participants gave the right answer, with 45% saying that the correct answer was 95%.

Cosmides and Tooby reformulated this question in terms of frequencies. Their new version read:

> 1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.
>
> Imagine that we have assembled a random sample of 1000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people. Given the information above:
>
> On average, how many people who test positive for the disease will *actually* have the disease? ____ out of _____ (1996, 24)

Unlike in the original study, and consistent with their ‘frequentist hypothesis,’ Cosmides and Tooby found that when presented in the new formulation, more than three quarters of participants gave the right answer. This study, along with several similar studies in which studies from the heuristics and biases tradition were rewritten in terms of frequencies (Fiedler 1988, Hertwig and Gigerenzer 994) were seen by many as a strong challenge to the pessimistic claims made by heuristics and biases researchers.

A second example of the evolutionary approach has been the attempts to explain responses to various versions of the Wason selection task detailed in 2.2. The Wason selection task has been shown to be highly context-dependent, but exactly what aspects of context play a role in success at the selection task has been a matter of considerable confusion. For instance, in a well know reformulation of the selection task, Griggs and Cox (1982) presented participants with four cards, two of which were labeled with drinks (one beer, one a non-alcoholic beverage) and two cards with ages (one over 21 and one under 21). Participants were told that each card had a drink on one side and the age of the person drinking on the other side. Participants were then asked to indicate which cards must be checked to determine if people were obeying the law that: 'if a person is drinking beer, then he or she must be over 21.'

Unlike in the original formulation, in which only 25% of people selected the two cards necessary to test the rule, 75% chose the correct cards in this formulation. Cosmides and Tooby (1992) explained the differences found between the two formulations by positing that animals would have evolved a mechanism designed to detect cheaters over the course of our evolutionary history because this would significantly increase fitness. This would explain why people are good at solving the Wason selection task when the context is related to cheating but bad at it when the task is provided in abstract terms. Responses to the Wason selection task can be predicted with a high degree of success based on whether the context would activate the hypothetical 'cheater detection' module that Cosmides and Tooby posit.

Evidence from evolutionary psychology seems to indicate that our reasoning may be much better than the heuristics and biases literature has led us to believe. As with the heuristics and biases tradition, Samuels and Stich (2004) have identified three increasingly optimistic claims made by evolutionary psychologists:

4) There are many reasoning and decision-making problems on which people's intuitive judgments do not deviate from appropriate norms of rationality (295).

5) Many of the instances in which our judgments and decisions accord with appropriate norms of rationality are to be explained by the fact that, in making these judgments, we rely on mental modules that were designed by natural selection to do a good job at nondemonstrative reasoning when provided with the sort of input that was common in the environment in which our hominid ancestors evolved (295).

6) Most or all of our reasoning and decision making is subserved by normatively unproblematic "elegant machines" designed by natural selection, and thus any concerns about systematic irrationality are unfounded (295).

These three claims illustrate just how different the Panglossian position is from the heuristics and biases position. Where heuristics and biases researchers conceive of human reasoning and decision-making as relying on the 'shoddy software' of normatively problematic simple heuristics, evolutionary psychologists hold that most of cognition is served well by elegant machines designed to work optimally in almost any circumstance we encounter. Even in cases where the heuristics and biases approach suggest that humans are rational, it is quite different from the rationality the evolutionary psychology program imagines. In one case humans are deemed rational because even though there is a significant gap between the descriptive abilities of humans and the normatively suggested solutions, the impoverished cognitive processing of humans is such that they are doing the best they can. From the Panglossian perspective, humans are also doing the best they can, but 'the best they can' is quite close to normatively optimal, once we have applied the appropriate normative system. How to resolve the differences between these two positions has been one of the primary focuses of the contemporary rationality debate. In the next section and the remaining chapters the focus of the thesis will be on a recent proposal for resolving these differences.

2.4. A Middle Ground?

Recently, philosophers (Samuels & Stich 2004, Samuels forthcoming, Pollock 1991), psychologists (Evans 2003, Stanovich 1999, 2004, Kahneman and Frederick 2002), and neuroscientists (Montague 2006) have proposed that the extreme positions of the heuristics and biases tradition on the one hand and the Panglossian approach on the other can be reconciled with each other through a third, more moderate, position. Defenders of this position recognize both that there are many cases in which people systematically reason in ways that diverge from normatively predicted responses, but also that there are many cases in which people's responses track the normative solutions quite closely. In terms of the three claims that Samuels and Stich (2004) identify for each position, the moderate position accepts both the pessimistic claim (1), which says that on a large number of tasks, people deviate from the rationally suggested response and the optimistic claim (4) which says that there are a large number of situations in which people do behave rationally. Certainly given the immense amount of reasoning and decision-making that people do in everyday life, there is room for each of these claims to be true.   Claims (2) and (5) are also both potentially true statements about the mechanisms underlying reasoning and decision-making, and there is no *prima facie* reason that these claims are in conflict. Claim (2) proposed that many cases in which we deviate from the rationally suggested response we are relying upon simple heuristics, whereas claim (5) proposed that our rational reasoning and decision-making is underscored by evolutionarily selected modules. Unlike claims (1), (2), (4), and (5) however, the moderate takes both claims (3) and (6) to be moments of "rhetorical flourish" (Samuels, Stich & Bishop 2002) and more importantly holds that both are most likely false. Nothing in the evidence from either side appears to suggest either that (3), which claims that all we have to rely on in decision-making and reasoning are poorly performing heuristics, or (6) which claims that all or most decision-making and reasoning is supported by "normatively unproblematic" modules are true and when taken as a whole, the evidence from each side of the debate seems to falsify the theoretical claim of the other.

To explain the entire set of data from research on reasoning and decision-making, this moderate position proposes a dual-process theory in which there are two distinct types of cognition. These two types of processes will be explained in considerable detail in the following chapter. After outlining the dual-process theory, I will argue that the two types of cognition are particularly sensitive to two different conceptions of rationality, one evolutionary and the other the normative rationality suggested by the Standard Picture. By recognizing that there are two systems each operating according to a different conception of rationality, we can make progress in studying human rationality, and additionally, develop new paths that both normative and descriptive research in rationality should focus on.

Chapter 3: The Dual-Process Theory

Dual-process theories have recently become popular explanations of many aspects of cognition. According to dual-process theories, there are two distinct types of systems that underlie our cognitive processing (Chaiken & Trope 1999). One system typically operates quickly, is non-conscious, can operate in parallel with other processes, is relatively undemanding of cognitive resources, and is often thought to be associated with the emotions and intuition. Many believe that this evolutionarily older system is shared with other animal species. Conversely, the second system is slow at processing information, under volitional control, is serial in its information processing, is relatively demanding of cognitive resources, and is often associated with reason or rationality. This system is typically thought to have evolved more recently and to be an exclusively human property. Although these two systems have been labeled in many ways depending on the domain of interest and the aspects of the systems that particular theorists want to highlight, to stay as neutral as possible with regard to the two processes, I will adopt Stanovich's (1999) terminology of calling the two processes simply *System 1* and *System 2*.[5]

3.1. System 1

As Stanovich (2004) has pointed out, even the relatively neutral System 1 terminology brings with it a serious problem. Specifically, it implies that there is a single System 1. In fact, this is almost certainly not the case. There are most likely multiple

[5] For a thorough listing of the various theorists and the labels they have given to the two processes, see Stanovich (2004: 35).

systems that meet the criteria for being considered a System 1 process (exactly what these criteria are, and how System 1 processes should be categorized will be addressed shortly). Stanovich writes: “Using a term…which implies a single cognitive system…is really a misnomer. In actuality, the term used should be plural because it refers to a (probably large) set of systems” (2004: 37). Stanovich describes System 1 processes as an “autonomous set of systems” (2004: 37) that share many properties with classic Fodorian modules (Fodor 1983), but diverge in important ways and should not be thought of simply as modules.

Like modules, System 1 processes are typically fast, automatic, and “mandatory” in the sense that “their operation is obligatory when triggered by relevant stimuli; central systems cannot make [System 1] processes refrain from triggering.… Central processes can, however, override the output of [System 1] in determining a response” (2004: 39). System 1 processes operate quickly and place little strain on cognitive resources because they have developed through evolutionary history to deal with specific problems common in a natural environment, are triggered by a minimal and specific set of stimuli, and produce stereotyped responses that do not require conscious control. Because of this, however, System 1 processes are susceptible to error when they are triggered in a situation that does not match the environment or problem type in which the process was designed to operate efficiently.

This failing of System 1 has been offered as an explanation of the data of both the heuristics and biases approach and the evolutionary psychology approach to reasoning and decision-making. As Samuels, Stich, and Bishop (2002) have pointed out, evolutionary psychologists have focused their research on testing reasoning and decision-making abilities with an eye to the way these problems would have been presented to our evolutionary predecessors. In doing so they have found that in these types of situations, humans (and other animals) are quite good "intuitive statisticians" (Cosmides and Tooby 1996) and utility maximizers (Glimcher 2003). In contrast, heuristics and biases research has tended to rely on problems that are considerably withdrawn from the format in which our predecessors would have received probability information (Kahneman and Tversky 1973). We would expect, given the properties associated with System 1, to find differences between evolutionarily common and uncommon reasoning and decision-making problems.

Unlike Fodorian modules, an important characteristic of System 1 is that they are not a clearly specified set of systems. According to Fodor, modules are exclusive to the five senses and language. System 1 processes are not restricted to these systems. They include many higher cognitive processes, such as reasoning and decision-making, and with practice, education, and training, processes that once required System 2 processing (which will be explained in the following section) can be integrated with and eventually become part of System 1 processes. This will become an important aspect of the dual-process theory in relation to rationality, for it will allow for the hope of an ameliorative project aimed at improving human reasoning and decision-making so that it more often approximates normative rationality.

Although there is a sense in which System 1 processes are deeply "unintelligent" in that they are highly inflexible, responding to simple triggering stimuli regardless of context, and running to completion of their process even in cases where inhibiting this response might be advantageous, this simple processing is also their greatest virtue. Fodor’s (1983) comments regarding the similar lack of intelligence of modules capture this point: “What you save by this sort of stupidity is not having to make up your mind, and making up your mind takes time” (64). Because humans cognitive resources are constrained in many ways, the “fast and frugal” style of System 1 processes are crucially important to an organisms survival.

A final component of System 1, which Stanovich (1999) in particular has stressed is the lack of individual differences in System 1 processing. Evidence supporting this claim has come from a series of studies in which performance on reasoning and decision-making tasks has been compared with standard measures of analytical intelligence (i.e. IQ, SAT scores) (Stanovich and West 1998, Capon *et al.* 2003). These studies have found that in reasoning and decision tasks that emphasize frequency and contextual evidence, such as those favored by evolutionary psychologists, there is no correlation between performance on the task and analytic intelligence. However, on tasks of the type favored by heuristics and biases researchers, there are significant correlations found between performance on the task and analytic intelligence. These findings have been interpreted as offering evidence for the lack of individual differences in System 1 processing. When tasks are contextual and provided in a frequentist format, for example, System 1 processes are likely to arrive at the normatively correct answer, and it is on these types of problems that no significant individual differences are found. On tasks that require System 2 analytic capabilities, large individual differences have been found.

To summarize, System 1 characterizes kinds of cognitive processes that are fast, automatic, unconscious, mandatory, use few cognitive resources, show few individual differences and developed relatively early in our evolutionary lineage. System 1 processing is highly sensitive to context, and produces near-optimal responses in situations that our evolutionary predecessors would have encountered. As has been briefly indicated already, the properties of System 2 stand in stark contrast to the properties that have been identified with System 1. The next section will focus on detailing the primary components of System 2 processes.

3.2. System 2

As with System 1 processes, it is likely that System 2 is not one particular system, but rather several systems that process information in a particular way. Whether System 2 is one system or many is a considerable controversy within the study of dual-process theories. Evans (2002), for instance, has argued that System 2 is necessarily a single central processing system because it is only in relation to this system that the multiple System 1 processes can be meaningfully grouped. He writes: "All that really links dual-process theories together is the nature of System 2 and the way in which implicit and autonomous processes (of whatever kind) appear to compete with it for control of our behavior" (205). Although there may be reasons for thinking that as a matter of fact humans have only a single System 2 process, there is certainly no reason that necessarily humans must have a singular System 2 process. Like System 1, it will be most useful from the standpoint of making scientific progress to think of System 2 as a particular type of cognition, with several characteristic component parts. It is an empirical question whether there are one or many System 2 processes, but it is certainly not foundational to the concept of a System 2 process that humans only have one.

System 2 processes are characterized by their slow, serial information processing. As opposed to System 1 processes that can quickly run many tasks in parallel, System 2 processing is limited by being able to engage only one task at a time. System 2 processing is not only significantly slower than its evolutionarily older counterpart, but requires significantly more cognitive resources as well. These resources include working memory capacity, time, and attention among many others. Taking these problematic aspects into consideration, one might initially wonder whether System 2 processes should be considered an unfortunate hindrance in our cognitive make-up. However, System 2 processes allow for many of our most important capabilities, such as abstract reasoning and volitional control and turn out to be of particular importance when considering the impact of dual-process theories for the rationality debate.

Among the primary advantages of System 2 processing is that it "allows us to sustain the powerful context-free mechanisms of logical thought, inference, abstraction, planning, decision-making, and cognitive control" (Stanovich 2004: 47). System 2 processes allow us to contemplate the future and to think counterfactually and abstractly about concepts free of their natural context. Additionally, and perhaps most characteristic of System 2 processing, it allows for conscious control. Contrary to System 1 processes, which respond blindly to specific stimuli and produce a stereotyped response that once begun must follow through to completion, the conscious control of

System 2 allows for flexibility and creativity in the type of response generated. As our world becomes less and less like the environment that our evolutionary predecessors engaged with, this flexibility becomes all the more important.

A last component of System 2 processing, and perhaps the most important with regard to the stance toward human rationality which I adopt here, is that of its ability to "override" System 1 processes. John Pollock's (1991) work has been especially important in emphasizing this feature of System 2. For Pollock, this override component is crucial for tempering the stereotyped System1 processes when they are activated in an environment or situation in which they will produce negative results. As an example, Pollock points to the System 1 process that predicts the trajectory of an object in motion. This System 1 process produces quick, and for the most part, reliable judgments about trajectory. However, there are situations and environments in which the System 1 processes will lead to substantially *unreliable* judgments of trajectories.[6] In these cases, System 2 processes can override the System 1 judgments. Exactly how this override ability of System 2 occurs is poorly understood, and an important question to resolve within the study of dual-process theories. As we will see in Chapter 4 however, neuroscientific research is beginning to provide us with information about this override function of System 2.

The override problem points to a larger gap in the understanding of dual-process theories, that of how the two systems interact with each other. Although some

[6] For a familiar example of this phenomenon, one might think back to the "moon bounce" balls available in many supermarket vending machines that many kids play with. Because of the imperceptible spin and the special properties of the material the balls are made out of, when bounced, the balls would seem to dart and bounce in completely unexpected directions.

philosophical work has been done on this problem (Scaife manuscript) and several theorists have posited plausible theoretical models (Haidt 2001, Evans 2003) the specifics of how these systems interact remains in many respects an open question. As we shall see in the next chapter, one of the most promising aspects of taking a cognitive neuroscience approach, and in particular a neuroeconomic approach, to the study of reasoning and decision-making is that it may provide information about the interactions between the two systems in a way that purely behavioral research cannot.

To reiterate the key components of System 2 processes, they are typified by slow, serial, context-independent processing and have developed relatively recently in our evolutionary history. Additionally, they allow for conscious control, provide substantial amounts of flexibility in their operation, but also creating significant demands upon cognitive resources, particularly working memory. A last component common to System 2 processes is the ability to override the rote, and sometimes unreliable, responses generated by System 1 processes. Although this override component has not been thoroughly detailed in the literature, I will propose that a cognitive neuroscience approach will shed new light on these interactions.

3.3. System 1 and System 2 as Investigative Kinds

Now that the two systems comprised within dual-process theories have been established, it is important understand the most efficient way to study them. This question has recently been addressed by Richard Samuels (forthcoming), who has defended a view of System 1 and System 2 as cognitive kinds. On Samuels' view, System 1 and System 2 are distinct cognitive kinds because they provide a useful way to systematically and reliably decompose cognition. By and large I agree with Samuels analysis of System 1 and System 2, but will adopt the language of investigative kinds (Brigandt 2003) as a more useful label for the two systems, in particular because it emphasizes their roles in scientific investigation. Although this is not the place for a thorough defense of System 1 and System 2 as investigative kinds, I will present in some detail what I take investigative kinds to be, and why it is beneficial to think of System 1 and System 2 as investigative kinds.

Although sections 3.1 and 3.2 identify many important properties, individual researchers have tended to focus on a single property that divides processes into either System 1 or System 2 processes. For any distinction that has been made to distinguish the two processes, one researcher or another has focused on it as the single criterion for classification as either a System 1 or System 2 process. Although classifying the two systems according to this single criterion method has its advantages, chief among them its theoretical economy, it suffers from a serious problem. If dual-process theories merely offer the claim that cognitive processes can be divided according to some criteria (fast/slow, conscious/unconscious, automatic/controlled), then the theory becomes an uninteresting one because the claim is much too broad to lead to useful generalizations in scientific investigation. It looks like the theory, conceived in this way, is probably true, but trivially so. The fact that some cognitive processes are faster than others, or that we have more or less control over some processes has been well established in cognitive and social psychology, and if this is all that the dual-process theory is illustrating, then it is hard to see what is interesting about it, and even harder to see how it would provide any usefulness in scientific investigation.

However, dual-process theories do say something interesting about cognition. Properly conceived, it is not the fact that cognitive processes vary according to some property or other, but that there is a distinct set of properties characteristic of System 1 and System 2 processes (see Fig. 1 at end of chapter), and that these properties tend to cluster together (Samuels forthcoming). That is to say, cognitive processes that exhibit one property thought to be characteristic of System 1 (i.e. fast processing) also tend to have several other properties associated with System 1 processing (i.e. automaticity, early evolutionary development, and undemanding of cognitive resources). One of the primary advantages of cluster concepts is that they allow for borderline cases, and dispense with the need for a rigid set of necessary and sufficient conditions. Although System 1 properties tend to cluster together, and likewise for System 2 properties, individual System 1 and System 2 processes need not, and likely will not, contain all of the properties associated with the system.

If this is the proper way to conceive of System 1 and System 2, then it alleviates one criticism that has recently been mounted against dual-process theories (Evans 2006). This criticism has focused on recent empirical evidence that shows that there are some processes that are both fast, automatic, and cognitively cheap, but almost certainly evolutionarily recent in human development (Girelli et al. 2000). It has been argued that these types of cases, in which processes exhibit many of the properties of System 1, but have some overlap with System 2 properties, put pressure on dual-process theories. But, if the cluster concept view is correct, then there are no necessary or sufficient properties for category membership; all that is required is that the system exhibits a sufficiently large number of the properties that characterize it.

This formulation of the System 1 and System 2 concepts leads to the question of whether they should be considered as natural kinds. Certainly, the conception of the two systems as relatively stable property clusters is in the spirit of many contemporary explanations of natural kind concepts (Boyd 1988, 1999, Kornblith 1993, 2002). Recently however, Paul Griffiths (2004) has argued that it may be appropriate to cease using the label of natural kind as extensively as it currently gets used, due in large part to its historical baggage. As Griffiths points out, “the actual name ‘natural kind’ has been rendered increasingly inappropriate by what Boyd has termed the ‘enthusiasm’ for the concept to which it refers” (2004: 906). Emphasizing this enthusiasm, Griffiths points out that the *insights* into scientific practice garnered from the natural kind model have been usefully and justifiable applied to areas outside the natural sciences (particularly to psychology), but that *the term itself* is problematic outside of the natural sciences. As he puts it: “Given the historical baggage that attaches to the term ‘natural kind’ and its literal inappropriateness to many of the scientific categories to which it is now applied, it seems preferable to introduce an alternative” (ibid.) The alternative which Griffiths favors, and which I will adopt here, is that of an investigative kind (Brigandt 2003).

Ingo Brigandt (2003) introduced the term investigative kind. Describing investigative kinds, Brigandt writes:

> An investigative kind is a group of things that are presumed to belong together due to some underlying mechanism or a structural property. The idea that these entities belong to a kind might be due to some interesting similarities. Instead, an investigative kind is specified by some non-trivial underlying feature or process that is presumed to account for the observed similarities. An investigative kind concept thus originates when a certain pattern among a class of objects is observed and it is assumed to be founded on some theoretically important, but yet unknown relevant mechanism that generates this pattern. (2003: 1309)

Following from this passage, investigate kinds are typified by their role in scientific enquiry. Investigative kinds group together sets in such a way that meaningful inductive generalizations can be made based on them. Brigandt highlights criteria that investigative kinds share. These criteria are:

1. They pick out a set that is presumably related based on an underlying structure or mechanism that the members of the set share.

2. They originate when a pattern is recognized among a class of objects that is assumed to be based in some common mechanism or underlying structure.

3. They are associated with the empirical search for the mechanism underlying the kind.

4. They will be guided by hypotheses about the underlying mechanism.

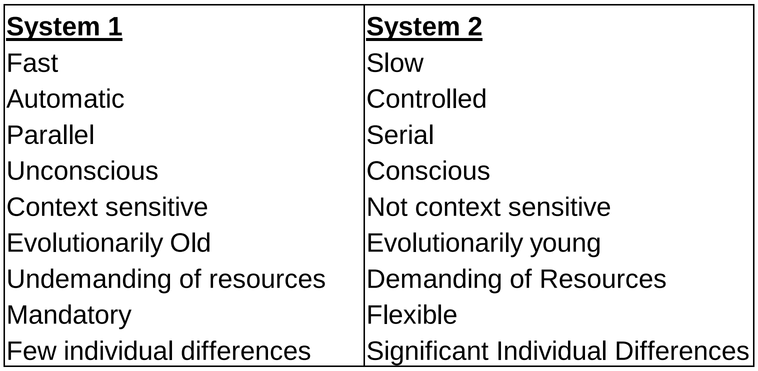
5. They will likely undergo revision based on findings from scientific investigation. This revision may take the shape of mild adjustments to the concept, the splitting of the concept into two distinct concepts, or even elimination of the concept completely.

Criteria (1) and (2) illustrate the focus of investigative kind concepts on underlying structure and mechanisms that tie together the members of a given investigative kind. For System 1 and System 2 processes, this underlying mechanism takes the form of the property clusters addressed earlier. Criteria (3) and (4) and (5) highlight the interplay between theses concepts and the scientific process. Investigative kinds both shape and are shaped by the empirical research into the concept. Initial hypotheses about the investigative kind concept in question leads to particular research questions and methodological approaches. In turn, the outcomes of these investigations provide increasingly rich information about the concept, leading to revision of the initial criteria for category membership.

This process has been carried out over the history of the study of dual-process theories. Dual-process theories were initially posited to explain the perplexing patterns of behavior found in reasoning and decision-making research. They were proposed to explain the seemingly conflicting findings of heuristics and biases and evolution psychology researchers. These initial (and quite simplistic compared with current dual-process theories) theories led to specific programs of research (individual difference studies, studies of base-rate neglect) that generated more data that were used to refine and reformulate the system 1 and System 2 concepts. In this back and forth interplay between the concept and the empirical data generated from and about it, significant forward progress is made. In the case of dual-process theories this progress has manifested itself in an increasingly meaningful way of partitioning cognitive processes, and a deeper and richer understanding of many of the most puzzling aspects of particular domains of human cognition, including those of interest for present purposes: reasoning and decision-making.

Recently, a new approach has been taken to studying dual-process theories of reasoning and decision-making. This new cognitive neuroscience approach is already uncovering new information about the nature of dual-process theories and in particular the way in which system 1 and system 2 interact with each other. This approach, which is the focus of chapter 4, stands to add much to the study of both dual-process theories in general and to their role in reasoning and decision-making tasks in particular.

Table 1: Characteristic Properties of System 1 and System 2



System 1	**System 2**
Fast	Slow
Automatic	Controlled
Parallel	Serial
Unconscious	Conscious
Context sensitive	Not context sensitive
Evolutionarily Old	Evolutionarily young
Undemanding of resources	Demanding of Resources
Mandatory	Flexible
Few individual differences	Significant Individual Differences

Chapter 4: Evidence From the Cognitive Neuroscience of Judgment and Decision-Making

Recently, several studies have applied a cognitive neuroscience approach to studying human reasoning and decision-making. This approach merges traditional behavioral tasks, such as those discussed in chapter 2, with modern neuroscientific methodologies such as single-cell recordings and neuroimaging. Much of this research has come to be known as neuroeconomics. Although still a field in its infancy, neuroeconomics—the merging of methodologies and research questions from neuroscience, economics, and cognitive psychology—has already made significant contributions to the understanding of human decision-making. By bringing methodologies from the neurosciences to bear on these questions, there is potential to more thoroughly understand how decisions are made, and also to begin comparing information about human neurobiology with the hypothetical cognitive mechanisms posited in much of cognitive science. It is this last aspect that has been of most usefulness in relation to dual-process theories.

Part of the conception of System 1 and System 2 processes as investigative kinds is the notion that all instances of the kind are connected by some underlying structure or mechanism. Although there is much that can be learned from purely behavioral studies, they are also limited in many well-known ways. Chief among their limitations is that they give almost no direct evidence about the mechanisms underlying the domain in question. Daniel Gilbert (1999) humorously illustrates this point with an analogy between the psychological researcher and a good detective. He writes:

> Although a talented detective may be able to rule out one or more architectures, no detective can rule *in* just one…. When inferences about architecture are informed only by knowledge of inputs (tap tap) and outputs (spurt spurt) then the conceptual Erector Set of gates, sensors, and other optional accessories affords the creative tinder an endless number of ways to link the tapping to the spurting. (8)

It is in improving this shortcoming of behavioral methodologies that cognitive neuroscience is most beneficial. Although it brings with it a host of its own methodological worries,[7] these are not insurmountable, and the benefits of a cognitive neuroscience approach allow researchers to ask questions and investigate their area of interest at a level that would never be possible through traditional behavioral studies. The ability to open up the 'black box' of the mind for the first time, and to examine the physical processes occurring in the brain allow for significantly stronger inferences about the mechanisms underlying reasoning and decision-making.

Cognitive neuroscience also allows for decision-making research to occur at several levels. Studies using neuroimaging technologies such as functional Magnetic Resonance Imaging (fMRI) have captured much of the interest in the area, but there is also significant work being done both at the molecular level (Kosfeld *et al*. 2005) and at the single neuron level (Glimcher 2003). By operating at multiple levels, and using multiple methodologies, the cognitive neuroscience approach has quickly developed an impressive amount of information about the workings of the brain during decision-making and reasoning tasks.

[7] See Hardcastle & Stewart (2002) for a particularly thoughtful critique of cognitive neuroscientific methodologies. Among the criticisms of fMRI are that it lacks sufficient spatial and temporal resolution, that it measures neural activity indirectly through blood flow, and that appropriate parameters for data analysis have yet to be developed and widely accepted.

Although the field is still far too young to draw conclusive arguments one way or the other about whether humans are rational or not, there is already ample evidence from which we can begin to see how major problems within the rationality debate may be resolved. The remainder of this chapter will first discuss the brain regions that have been proposed as neural correlates of System 1 and System 2 processes. I will then review several recent studies from the cognitive neuroscience of reasoning and decision-making. These studies lend additional support to the dual-process theory of decision-making, and have proved especially useful in improving our understanding of the way that System 1 and System 2 interact in the brain.

Matthew Lieberman (2007) has proposed that the neuroscience of higher-level cognition (particularly social cognition such as decision-making) can be fruitfully studied by adopting a dual-process theory. He focuses largely on the automatic and controlled features of System 1 and System 2 respectively, and has suggested several brain areas that may underlie System 1 and System 2 processing. There is often a danger in cognitive neuroscience research that the ‘how’ question which motivated the research (i.e. how do people make decisions) will be replaced with a potentially less useful ‘where’ question (which areas of the brain are active when people make decisions). Both because of this concern, and because of issues of space within this thesis, I will briefly lay out some of the primary brain areas which have been implicated with regard to System 1 and System 2, and then turn more thoroughly to how knowledge of these areas can is useful in furthering the understanding of dual-process theories and the neural-basis of decision-making and reasoning.

The brain systems that Lieberman (2007) nominates as potentially underlying System 1 processes include the amygdala, basal ganglia, lateral temporal cortex (LTC), ventromedial cortex (VMPFC), and dorsal anterior cingulate cortex (dACC). In keeping with the criteria for System 1 processes, each of these brain systems has been implicated in automatic, non-conscious, and relatively fast aspects of cognitive processing. Additionally, each of these brain systems developed relatively early in our evolutionary history.

In contrast to these brain systems, Lieberman proposes several others that may underlie System 2 processing. These include the rostral anterior cingulated cortex (rACC), lateral prefrontal cortex (LPFC), posterior parietal cortex (PPC), hippocampus and medial temporal lobe (MTL). Lieberman emphasizes the LPFC in particular, writing that it is "the heart of [System 1], as it is involved in numerous higher cognitive processes that are experienced as intentional and effortful, including working memory, implementation of top-down goals and plans, episodic retrieval, inhibition, and self-control" (296)[8]. Although much of the work that Lieberman relies upon when categorizing brain systems relevant to System 1 and System 2 comes from studies of specific brain areas, recent research has focused on the relationships between System 1 and System 2 brain areas during different types of cognitive and social tasks. This research is quickly expanding the understanding of how the two systems relate and interact with each other.

[8] Although this is not the place to review the large literature on each of the brain areas implicated, interested readers can find references in Lieberman (2007) to primary research on each brain area.

One method for investigating the neural basis of dual-process theories is to have participants use tasks that would use characteristically System 1 or System 2 processing. If dual-process theories are correct, and if there are distinct brain systems that underlie System 1 and System 2 processes, then there should be significantly different brain areas activated during tasks that heavily favor System 1 or System 2 processing. In fact, in studies of many areas of cognitive and social neuroscience, including categorization, prejudice, self-knowledge, and personality, it was found that when the tasks strongly favored System 1 processing, the brain areas noted by Lieberman (2007) as relevant for System 1 were significantly active, whereas when the tasks favored System 2 processing, the areas noted by Lieberman as relevant for System 2 were significantly active (Rauch et al. 1995, Cunningham et al. 2003, Lieberman & Eisenberger 2004, Eisenberger et al. 2005). Although there is certainly much more research that needs to be done, and in many ways the study of the neuroscientific basis of dual-process theories is still in its infancy, these studies, and many others like them give good evidence in favor of dual-process theories of cognition.

As briefly mentioned above, one way in which neuroscientific evidence can advance our understanding of dual-process theories is by allowing insight into the interactions between System 1 and System 2. It is thought that one function of System 2 is to override System 1 in specific situations in which its fast, automatic, and non-conscious processing might be disadvantageous. To study this theory, cognitive neuroscientists have begun to investigate aspects of cognition where it is thought that an override of automatic processing is likely to occur. One of the most promising areas for this approach is in the regulation and suppression of emotions. It has been found in multiple studies that passively viewing emotionally salient images correlated significantly with amygdala and VMPFC activation (both regions thought to underlie System 1 processing), whereas when participants were told to consciously try to regulate their emotional response to the images, there was a correlation with both decreased activation of the amygdala and VMPFC and significant activation of the LPFC (Ochsner et al. 2002, Gross 1998). Additionally, it was found that during these emotion regulation tasks, the larger the LPFC response was during reappraisal, the smaller the amygdala and VMPFC responses were during reappraisal. As Lieberman (2007) puts it:

> …some participants activated the LPFC a lot and others just a little. The more any participant activated LPFC, the greater the reduction in [System 1] activity. This analysis suggests that [System 2] is not merely "speaking more loudly" than [System 1]. Rather, it appears that, under certain circumstances, while [System 2] is speaking up, it is also taking away [System 1's] "microphone," so that its volume is diminished. (303)

The findings of these studies, as well as several from the neuroeconomic literature which will be discussed shortly, have provided significant evidence in favor of dual-process theories of cognition. As illustrated so far, there are relatively stable sets of brain regions that are highly correlated with System 1 processing and System 2 processing, as well as some evidence that shows neurally how System 2 might override System 1 in certain situations. Although this research provides a strong foundation for a general understanding of the neural basis of dual-process models of cognition, it hasn't looked specifically at the areas of cognition that are of interest in this thesis, reasoning and decision-making. Recently, researchers have begun to study the neural underpinnings of reasoning and decision-making. This research has been collectively termed neuroeconomics, and in only a short period of time, it has already begun to provide an understanding of what is happening in the brain when people make decisions. The remainder of this chapter will focus on research from this exciting new field.

Providing some of the best recent evidence for a dual-process theory of decision-making is Sanfey *et al.*'s (2003) article "The Neural Basis of Economic Decision-Making in the Ultimatum Game." Sanfey *et al.*'s experiment makes use of a well-known game from behavioral economics-- the ultimatum game. In the ultimatum game, subjects are placed in pairs and each is given a role in the game. One is the proposer and the other is the responder. The proposer is given $10 to divide up between herself and the receiver however she sees fit. After the proposer has divided the money, the offer is then told to the responder, who can either accept what the proposer has given him, or reject the offer, in which case neither player gets any money. This game is of particular interest because it offers, at least according to many theories of economic rationality, very clear predictions about how rational agents should play the game. According to expected utility, the proposer should always give as little as possible, because that maximizes the money that she keeps for herself, and the responder should accept any offer, no matter how small, because it is always better to have some money than to have no money, which is what he will receive if he rejects the offer. In multiple studies it has been found that people often behave irrationally by rejecting low offers (those in which the receiver is offered 3 or less of the 10 dollars).

In Sanfey *et al.*'s experiment, participants played the ultimatum game as the responder while in an fMRI scanner. The proposer was played by a computer, which played a specific pattern of offers. The participants were deceived into believing that they were actually playing against another human except in control cases, in which they were told that they were playing a computer. To simplify the game, only the following divisions of the $10 were used during the game: {5,5}, {7,3}, {8,2}, and {9,1}.

The results of the brain scans showed that three areas of the brain showed significant activation during games: the anterior insula (AI), the dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC). As noted previously, the DLPFC and the ACC have been implicated in System 2 processing, while the AI is known to be associated with disgust, pain, and negative emotions in general (Derbyshire et al. 1997, Calder et al. 2001). The study found that during low offers, activity in the AI was increased. When the activity in the insula became more active than the activity in the DLPFC and ACC, the player tended to reject the offer. When the activity of the DLPFC and the ACC were higher than that of the AI, the participant tended to accept the offer. When the participants were aware that they were playing against a computer, regardless of whether the offer was unfair or not, AI was not activated to the degree that it was activated during interpersonal play, the assumption being that the players did not have an emotional response (or at least not one of corresponding intensity) to a low offer from a computer.

This study provides a fascinating first glance at multiple systems competing in the brain to produce decision-making behavior. From these results, it appears that when presented with a low offer, there are multiple brain areas, one of which (DLPFC) has already been proposed as a neural correlate of System 2 processes, while the other (AI) is known to be implicated in negative emotions and pain. By measuring the relative activity of these two areas, a high degree of predictive power about a participant's choice is gained. This suggests that counter to traditional models, which suggests that there is just one system of cognitive processing occurring during decision-making, that by embracing dual-process theories, and investigating them through cognitive neuroscience methods, we can gain a deeper insight into how decisions are made.

McClure et al. (2004) reported further evidence in support of a dual-process theory of decision-making. They were interested in how to explain the phenomena of inequity in time discounting. According to traditional models of economic rationality (Koopmans 1960), agents should treat equal delays in time equally (e.g. the difference in utility between receiving a reward today instead of 24 hours later should be equal to the difference in utility between receiving a reward in one year as opposed to one year and one day). However, when tested, people offered $10 now or $11 tomorrow tend to take the immediate option, while when offered $10 in a year or $11 in a year and a day people tent to take the $11.

McClure et al. (2004) theorized that inequity in time discounting could be explained in terms of the interactions between multiple neural systems. Explaining this they write:

> Short-run impatience is driven by the limbic system, which responds preferentially to immediate rewards and is less sensitive to the value of future rewards, whereas long-run patience is mediated by the lateral prefrontal cortex and associated structures, which are able to evaluate trade-offs between abstract rewards, including rewards in the more distant future. (2004:504)

This hypothesis was taken to be plausible based on previous findings in several areas. First, discrepancies in time discounting appear to be a uniquely human characteristic. In almost all other species, including our closest primate relatives, delay of gratification is virtually undocumented (Kagel *et al.* 1995). Given that one of the primary neuroanatomical differences between humans and all other species is the size of prefrontal cortex, they hypothesized that gratification delay might have its neural substrates broadly in this area. Additionally, evidence from patients with brain damage has suggested that damage to the prefrontal cortex often leads to a diminished ability for long term planning, as well as a tendency toward preferences for immediate rewards (Damasio 1994).

On these grounds, McClure et al formulated three empirically testable hypotheses.

1) When given a choice that includes an option for an immediate reward, limbic structures will be preferentially engaged relative to choices that do not include an immediate reward.

2) Areas in prefrontal cortex will exhibit similar activity across all choice conditions relative to when the participant is at rest.

3) When a participant chooses a delayed reward option, there will be greater activity in lateral prefrontal areas that in limbic areas.

To test these hypotheses, participants were given a series of choices between smaller offers received earlier (i.e. $5 immediately following completion of the task) or larger but delayed offers ($40 given in six weeks) while in an MRI scanner. McClure et al found that all three of their hypotheses were validated. Areas traditionally associated with the limbic system (ventral striatum, medial orbitofrontal cortex, and medial prefrontal cortex) were significantly more activated when immediate rewards were offered than in trials in which immediate rewards were not offered. They also found that across all decisions, there was significant activation of several areas of lateral prefrontal cortex, supporting their second hypothesis. Most interestingly from a dual-process perspective, they found that when the areas found to be significant in the prefrontal cortex were significantly more active limbic structures, participants were significantly more likely to choose the delayed choice over the immediate choice.

These two studies as well as several others in related areas such as moral judgment (Greene et al. 2001, 2004) and reasoning (Goel & Dolan 2003) have all found that when the activity of the brain is examined, it appears to operate as if there are two types of processes which interact, and based on this interaction, behavior is highly predictable. The findings from these studies are highly suggestive of a biologically realistic dual-process theory, which McClure et al point to when they write: "Human behavior is often governed by a competition between lower level, automatic processes that may reflect evolutionary adaptations to particular environments, and the more recently evolved, uniquely human capacity for abstract, domain general reasoning and future planning" (2004 506). This characterizes the dual-process theory well, and the behavioral data gathered in conjunction with the imaging data makes the case even stronger. One of the hallmarks of System 1 processing is that it is relatively fast compared to System 2. McClure *et al*. (2004) found that when participants chose the immediate reward, they were significantly faster than when they chose the delayed, but larger reward. Additionally, in many of these studies it has been pointed out that the faster, evolutionarily older system also has a strong affective or emotional component related to it, which although not addressed in detail in the earlier discussion of System 1 processing is often thought to be included among the properties typical of the process (Sanfey *et al*. 2006).

Although these studies are only an initial step toward understanding the biological mechanisms underlying reasoning and decision-making, they suggest that a cognitive neuroscience approach will be a highly fruitful one to take. Not only does it help to confirm many of the theoretical claims made about dual-process theories, but it also allows a way to begin making progress into understanding (1) the underlying mechanisms of the two systems and also (2) the interactions between System 1 and System 2 processes. These have traditionally been two of the most difficult to investigate aspects of dual-process theories, and also two of the most important. Particularly when trying to draw conclusions about human rationality, whether or in which circumstances humans are able to override the older, more hardwired System 1 processes with the more flexible System 2 is crucial. In the next chapter, I will conclude by returning to the question underlying this thesis: are humans rational? Like many who endorse a dual-process view of reasoning and decision-making, I think the answer is neither as pessimistic as some have claimed nor as optimistic as others have asserted.

Chapter 5. Conclusion

If it is the case that human reasoning and decision-making are best understood within a dual-process framework as I have suggested throughout this thesis, then what does this mean for the study of human rationality? I think it signals a shift in the way that we traditionally conceive of the most important questions within the study of rationality. In this final chapter, I will suggest where I think the study of rationality should focus at both the normative and the descriptive level. Fully drawing out the implications of a dual-process theory of reasoning and decision-making for rationality would be the work of a much larger and more ambitious project, which I hope to undertake in the future. But, in this chapter I will sketch the direction that I think fruitful projects within the rationality debate should head.

As discussed in Chapter 2, heuristics and biases researchers have taken a fairly pessimistic view of human rationality based on many studies that appear to show that people do not perform according to the prescriptions of various normative theories. Conversely, evolutionary psychologists have argued that humans *are* by and large rational, and have produced their own research to defend this position. I believe that dual-process models give us an explanation of why we find these apparently conflicting results.

As Stanovich (2004) has suggested, one fallout of the System 1 and System 2 processes is that they accord well with two different conceptions of rationality. The evolutionarily older, fast, and automatic System 1 is highly attuned to evolutionary rationality, in which to be rational is to reason or choose so as to maximize fitness. Over the course of human evolutionary history, System 1 has developed to be highly successful at leading to those choices that maximize the fitness of the genes of an individual. System 2, on the other hand, allows for the maximization of the type of normative rationality that has been discussed throughout this thesis, the rationality that Stein (1996) calls the Standard Picture. In most cases, both of these systems of rationality propose the same choice. That is to say, most of the time the choice that is optimal according to the Standard Picture is also optimal according to evolutionary rationality. But, there is nonetheless a significant amount of decision-making and reasoning in which the two systems of rationality do not suggest the same choice.

Evolutionary psychologists have tended to focus their research on studies in which the reasoning or decision-making tasks are presented in a way that would make use of evolutionarily older, System 1 cognitive processes. Additionally they have focused on cases in which rationality in both the evolutionary and the Standard Picture sense prescribes the same choice. On the other hand, heuristics and biases researchers have focused on decision and reasoning tasks in which System 1 processes lead to sub-optimal choices, or on cases in which the evolutionary and Standard Picture systems of rationality suggest different outcomes. Because of these different focuses, the two groups came to radically different notions of human rational capacity.

Approaching human rationality from a dual-process perspective opens up new questions at both the descriptive and the normative/prescriptive level. Descriptively, the questions of when and how humans are able to override System 1 processes with System 2 control becomes crucial. Additionally, the question of whether people can become better at overriding System 1 through education and cognitive training becomes crucial. In answering these two questions, I believe that a cognitive neuroscience approach will become particularly useful. It offers us the best current methodology by which to investigate and understand the unique interactions between the multiple systems in the brain that lead to decision-making and reasoning. By better understanding the biological and neural properties of each system, we will be better able to answer the descriptive questions which, if dual-process theories are correct, are most important.

At the normative level new questions for research are introduced as well. Primarily, a new prescriptive question that arises is when one should make choices in accord with System 1 and when one should override System 1 choices with System 2 so as to accord with the Standard Picture of rationality. Certainly it is not the case that, even if it were possible, one would want to always override System 1 processes. After all, there are many, many cases in which the fast and automatic decision-making of System 1 is far superior to the slow, and consciously controlled System 2. When one inadvertently steps into the street in the path of a speeding car, it is certainly better to react quickly, relying on hardwired System 1 processes than to contemplate the costs and benefits of returning to the sidewalk as opposed to getting hit by a car.

The normative question becomes when ought one rely on System 1 and when ought one rely on System 2. By developing research into this question, human decision-making stands to be significantly improved. Much of the problem within the rationality debate has been that since both ends of the debate assumed only one type of cognitive processing, they did not see a large opportunity for improving human rational capacities.

Dual-process theories suggest that there may be room for significantly improving our rational abilities. In order to develop this potential however, it will require research into the questions suggested here at both the descriptive and prescriptive level. Moreover, both of these new programs will be highly interrelated. As more descriptive information is uncovered about when and how people *can* override System 1 processes with System 2 processes, the normative and prescriptive conclusions about when one ought to override System 1 may change. Moreover, the question of whether humans descriptively are rational will become a matter of whether, given the capacities of System 1 and System 2, they actually do override System 1 when it is prescribed that they ought to.

Throughout this thesis I have proposed that the traditional rationality debate, with the pessimistic claims of heuristics and biases researchers representing one end of the debate and the Panglossian claims of evolutionary psychologists representing the other, has missed a crucial moderate position which becomes viable if dual-process theories of cognition are possible. I have suggested that there is good reason to believe that dual-process models are a useful distinction within the study of reasoning and decision-making, especially when conceived of as investigative kind cluster concepts. Moreover, I have suggested that there is now growing evidence about the neural underpinnings of dual-process theories, and that this evidence is opening up a new understanding of the way that the two systems interact to produce behavior. In this last chapter I have briefly sketched where I think rationality research, both descriptive and normative, should focus its attention in the future.

Within the context of the rationality debate, dual-process theories are an exciting and I believe important contribution. After more than 40 years of productive and extremely important research from both ends of the rationality debate, we are currently at a point where a new and radically different approach can be taken. By merging methodologies ranging from philosophy to psychology and from economics to neuroscience, a wealth of new information is being gathered and our knowledge is expanding at an incredible rate. It appears that if this interdisciplinary approach continues on its current path, a significantly improved understanding of what it means for humans to be rational is almost on the horizon.

References

Baron, J (1985). Rationality and intelligence. Cambridge, England: Cambridge University Press.

Boyd, R. (1988). How to be a moral realist, in Sayre-McCord, G. (ed.), *Essays on Moral Realism*. NY, Cornell Univesity Press, 181-228.

Boyd, R. (1999). Homeostasis, Species, and Higher Taxa. In Robert A. Wilson (ed.), *Species: New Interdisciplinary Essays*. Cambridge, MA: MIT Press 141-185.

Brigandt, I. (2003). Species pluralism does not imply species eliminativism. *Philosophy of Science* 70, 1305–1316.

Calder, A.J., Lawrence, A.D., and Young, A.W. (2001) Neuropsychology of Fear and Loathing, Nature Reviews Neuroscience, 2(5), 352-363.

Camerer, C. Lowenstein, G., Prelec, D. (2005). Neuroeconomics: How Neuroscience Can Inform Economics. Journal of Economic Literature 34 (1): 9–64.

Capon, A., Handley, S.J., & Dennis, I. (2003). Working memory and reasoning. *Thinking and Reasoning*, 9, 203-244.

Casscells, W., Schoenberger, A. and Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. New England Journal of Medicine, 299, 999-1000.

Chaiken, S. & Trope, Y. (1999). (Eds.) *Dual-process theories in social psychology*. New York: Guilford Press.

Chorvat, T. and McCabe, K. (2005). "Neuroeconomics and Rationality" . Chicago-Kent Law Review, Vol. 80, p. 101.

Cohen, L. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4, 317-370.

Cohen, L.J. (1986) *The Dialogue of Reason.* Oxford: Oxford University Press.

Cosmides, L. and Tooby, J. (1992). Cognitive adaptations for social exchange. In Barko, Cosmides, and Tooby (1992) 163-228.

Cosmides, L., & Tooby, J. (1996) Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1), 1-73.

Cunningham, W.A., Johnson, M.K., Gatenby, J.C., Gore, J.C., & Banaji, M.R. (2003). Neural components of social evaluation. *Journal of Personality and Social Psychology*, 85, 639-649.

Damasio, A. (1994). *Descartes' Error: Emotion, Rationality and the Human Brain*, Grosset Books: New York, Putnam.

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

Derbyshire, SWG, Jones AKP, Gyulai F, Clark S, Townsend D, Firestone L. Pain processing during three levels of noxious stimulation produces differential patterns of central activity. *Pain* 1997; 73: 431-445.

Eisenberger, N.I., Lieberman, M.D., & Satpute, A.B. (2005). Personality from a controlled processing perspective: an fMRI study of neuroticism, extraversion, and self-consciousness. *Cognitive, Affective, and Behavioral Neuroscience*, 5, 169-181.

Evans, J. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128, 978-996.

Evans, J. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Science*, 7(10), 454-459.

Evans, J. (2006). Dual system theories of cognition: Some issues. Presented at the 2006 Annual Meeting of Cognitive Science Society, Vancouver.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. Psychological Reseach, 50, 123-129.

Fodor, J. (1983) *The Modularity of Mind*. Cambridge: MIT Press.

Frankish, K. (2004) *Mind and Supermind*, Cambridge: Cambridge University Press.

Gigerenzer, G. (1994) Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright and P. Ayton, eds., *Subjective Probability*. New York: John Wiley

Gilbert, D. T. (1999). What the mind's not. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology*. New York: Guilford.

Gilovich, T., Griffin, D., & Kahneman, D. (eds.) (2002). *Heuristics & Biases.* New York, NY: Cambridge University Press.

Girelli, L., Lucangeli, D., & Butterworth, B. (2000). The Development of automaticity in accessing number magnitude. Journal of Experimental Child Psychology, 19, 104-122

Glimcher, P. (2003). *Decisions Uncertainty, and the Brain: the Science of Neuroeconomics.* Cambridge, MA: MIT Press.

Goel, V. and Dolan, R. (2003). Explaining Modulation of Reasoning by Belief. *Cognition*, 87(1), B11-B22.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., & Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral Judgment. Science, Vol. 293, Sept. 14, 2001, 2105-2108.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron, Vol. 44, 389-400.

Griffiths, P. E. (2004). Emotions as Natural Kinds and Normative Kinds. *Philosophy of Science 71 (5 Supplement: Proceedings of the 2002 Biennial Meeting of the PSA).*901-911.

Griggs, R. and Cox, J. (1982). The elusive thematic materials effect in Wason's selection task. *British Journal of Psychology*, 73, 407-420.

Gross, J.J. (1998). Antecedent and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. Journal of Personality and Social Psychology, 74, 227-237.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionalist approach to moral judgment. Psychological Review, 108(4), 814-834.

Hardcastle, V.G., Stewart, C.M. (2002). What do brain data really show? Philosophy of Science, 69, S72-S82.

Hertwig, R. and Gigerenzer, G. (1994). The chain of reasoning in the conjunction task. Unpublished manuscript.

Kagel, J., Battalio, R., & Green, L. (1995). *Economic Choice Theory: An Experimental Analysis of Animal Behavior.* Cambridge: Cambridge University Press.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman

(Eds.), *Heuristics and Biases: The psychology of intuitive judgment*. Cambridge, Cambridge University Press, 49-81.

Kahneman, D. and Tversky, A. (1973) On the psychology of prediction. *Psychological Review*, 80, 237-251.

Kahneman, D., & Tversky, A. (2000). *Choices, Values, and Frames.* Cambridge: Cambridge University Press.

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

Koopmans, T. J. (1960). Stationary ordinal utility and impatience. *Econometrica* 28, 287–309.

Kornblith, H. (1993). *Inductive Inference and Its Natural Ground* Cambridge: MIT Press

Kornblith, H. (2002). *Knowledge and its Place in Nature*. NY, Clarendon Press.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U. & Fehr, E. Oxytocin Increases Trust in Humans. Nature 435, 673–676, 2005.

LeDoux, J.E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York: Simon & Schuster.

Lieberman, M.D., & Eisenberger, N.I. (2004). Conflict and habit: A social cognitive neuroscience approach to self. In A. Tesser, J.V. Wood, & D.A. Stapel (Eds.), *On building, defending and regulating the self: A psychological perspective* (pp. 77-102). New York: Psychology Press.

Lieberman, M. D. (2007). The X- and C-systems: The neural basis of automatic and controlled social cognition. To appear in E. Harmon-Jones & P. Winkelman (Eds.), *Fundamentals of Social Neuroscience* (pp. 290-315). New York: Guilford.

McClure, S.M., Laibson, D.I., Loewenstein, G., Cohen, J.D. (2004) Separate neural systems value immediate and delayed monetary rewards. *Science*, 306: 503-7.

Montague, R. (2006). *Why Choose This Book?* New York: Dutton Press

Nisbett, R. and Borgida, E. (1975). Attribution and the social psychology of prediction. *Journal of Personality and Social Psychology*, 32, 932-943.

Ochsner, K.N., Bunge, S.A.,Gross, J.J., & Gabrieli, J.D. (2002). Rethinking feelings: an fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neurosciences*, 14, 1215-1229.

Pollock, J.L. (1991). OSCAR: A geral theory of rationality. In J. Cummins & J.L. Pollock (Eds.), Philosophy and AI: Essays at the interface, Cambridge: MIT Press,189-213.

Rauch, S.L., Savage, C.R., Brown, H.D., Curran, T., Alpert, N.M., Kendrick, A. (1995). A PET investigation of implicit and explicit sequence learning. *Human Brain Mapping*, 3, 271-286.

Samuels, R. (forthcoming) The magical number two, plus or minus: Some comments on dual-processing theories of cognition.

Samuels, R. and Stich S. (2003) "Rationality and Psychology". In A. Mele (ed.) *The Oxford Handbook of Rationality*. New York: OUP, 279-300.

Samuels R., S. Stich and M. Bishop (2002) Ending the Rationality Wars: How To Make Disputes About Human Rationality Disappear. In R. Elio (ed.) *Common Sense, Reasoning and Rationality*. New York: Oxford University Press, 236-268.

Samuels R., S. Stich, and L. Faucher (2004) "Reason and Rationality" I. Niiniluoto, M. Sintonen, & J. Wolenski *Handbook of Epistemology*. Dordrecht: Kluwer. Pp. 131-179.

Sanfey, A.G., Loewenstein, G., Cohen, J.D. & McClure, S.M. (2006). Neuroeconomics: Cross-currents in research on decision-making. Trends in Cognitive Sciences 10, 108-116.

Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., Cohen, J. The neural basis of economic decision-making in the ultimatum game. Science, 300: 1258-1261, June 2003.

Scaife, R. Dual-process and Cognitive Checking. (unpublished manuscript).

Stanovich, K. (1999) *Who is Rational?* New Jersey: Lawrence Erlbaum Associates.

Stanovich, K. (2004) *The Robot's Rebellion.* Chicago: University of Chicago Press.

Stanovich, K. & West, R. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, 127, 161-188

Stein, E. (1996). *Without Good Reason*. Oxford: Clarendon Press.

Stich, S. (1990). *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Stich, S. & Nisbett, R. (1980) Justification and the psychology of human reasoning. *Philosophy of Science*, 47, 188-202.

Thagard, P. & Nisbett, R. (1983) Rationality and charity. *Philosophy of Science*, 50, 250-267.

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

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.

Wason, P., Shapiro, D. (1966). Reasoning. In *New Horizons in Psychology*. Penguin, Hammondsworth, UK.