

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2014

On the Plurality of Thought: Beyond Dual-System Views

Tilli, Cecilia

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01w6634376g>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

ON THE PLURALITY OF THOUGHT:
BEYOND DUAL-SYSTEM VIEWS

Cecilia Inés Tilli

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PHILOSOPHY

Adviser: Sarah-Jane Leslie

April 2014

© Copyright by Cecilia Inés Tilli, 2014.

All rights reserved.

Abstract

The contrast between two types of thinking, one fast and automatic, the other slow and deliberate, is present in many philosophical and psychological accounts of the mind. It is also present in the common sense distinction between *reason* and *intuition*. Recent psychological data has provided support for this *duality of thought*. Dual-process views have been proposed for memory, reasoning, judgment, and decision-making. Researchers have sought to unify these views under a general dual-system theory: one system underlies the fast, automatic, low-effort, high-capacity processes; the other system underlies the slow, deliberate, high-effort, low-capacity processes.

In the dissertation, I show that there is no non-ad hoc way of keeping the distinction between the two systems. First, I consider the relation between systems and processes, and their role in modeling the cognitive architecture of the mind. I claim that we need to abandon the idea of a single, unified dual-system view and distinguish a duality at the micro-architectural level (a duality of processes) from a duality at the macro-architectural level (a duality of systems or duality of the mind). I show that a distinction at the micro-architectural level can explain (to a certain degree) the clustering of properties described in the dual-process literature. However, it cannot support a general dual-system view, that is, a macro-architectural duality of the mind. Appeals to underlying cognitive architecture, the use of working memory, and levels of analysis are all useful to explain some of the data and make sense of the complexity of processes involved in “higher-level” cognition. However, none of them can provide a robust account of the dual-system view. Instead, I propose a hybrid model of the mind, one that takes into account micro-architectural features and includes a

third type of process, which is crucial for conflict resolution and cognitive control. To conclude, I discuss an application of the dual-system view to moral judgment and decision making. I show that, unless we have independent means of establishing the reliability of the two types of processes, there is no reason to assume, on neural data alone, that slow deliberation will be more reliable than moral intuition.

Table of Contents

Abstract	iii
Acknowledgments	vi
Introduction	1
Chapter One	8
Chapter Two	38
Chapter Three	71
Chapter Four	104
Chapter Five	139
Conclusion	167
References	171

Acknowledgements

My experiences at Princeton have been invaluable and among the most stimulating of my life. Two reasons stand out: one was the opportunity to study and research some of the most interesting, challenging topics I have encountered in my academic life; the other was the chance to meet and get to know amazing people, some of whom have remained close friends.

My path into the cognitive sciences would not have been the same without the guidance of my doctoral adviser, Sarah-Jane Leslie. I would like to express my heartfelt appreciation for her advice, supervision, patience, and support, especially at times when I was failing the most.

I owe a very important debt to Gilbert Harman. From the beginning of my doctoral studies, he offered his support in all kinds of ways and showed me that interdisciplinary interests were valuable and worth pursuing.

I would like to express my gratitude to all the faculty at Princeton's Department of Philosophy for teaching me how to do philosophy in a wonderful collegiate atmosphere. In particular, I am very grateful to John Hawthorne, who gave insightful comments and suggestions on an earlier draft of the dissertation. I would also like to thank Mark Johnston for encouraging me to complete this project in various ways, and offer a special thanks to Adam Elga and Samuel Wang (faculty member of the Princeton Neuroscience Institute), for participating in the examination committee. I would also like to express my gratitude to the Department's staff, a special *family* on whom I could always rely for help, conversation (and sweets!) when needed.

I would like to express my heartfelt appreciation to all the friends and family members who provided company and support during the challenging time of writing this dissertation: Anna, Chiara, Dalila, Dora, Giorgia, Hugo, Katy, Laura, Leonie, Marilé, Meghan, Naima, Natalia, and Peque. A special thanks goes to Todd, a constant companion during the tribulations of graduate life and beyond.

I owe immense gratitude to three dear friends. My best friend, Paula, has been present since the beginning of my doctoral studies, sharing all the excitement, happiness, and challenges of graduate life. One could not ask for a more constant, caring, and supportive friend. My friends Giorgos and Santiago provided invaluable support during the most trying, difficult times of my life. Despite being an engineer, Santiago was happy to give me feedback on the dissertation whenever I needed it. He also made countless sacrifices – including a nine hour trip to Boston in the middle of a blizzard – to help me through the difficulties of graduate life. Giorgos called me from the other side of the world for months to make sure I was making good progress on the dissertation. He also helped me during one of the saddest, most baffling times of my life. They say you know your friends in the bad times. They have certainly proven to be true friends.

I would like to express my deepest, heartfelt appreciation to my mother, who is a constant reminder of what the strength of a woman can accomplish. She is also an excellent listener and a steady source of irony and laughter, a key ingredient in this unusual life I lead. Last, but not least, my father showed me the importance of a life devoted to knowledge and self-discovery. He remains the single most important influence in my education. Together, my mother and father taught me the value and significance of an educated mind. For this reason, I dedicate the dissertation to them.

To my mother and to the memory of my father

Introduction

The contrast between two types of thinking, one fast and automatic, the other slow and deliberate, is present in many philosophical and psychological accounts of the human mind. It is also present in the common sense distinction between *reason* and *intuition*. In recent decades a wealth of research in several areas of psychology has provided empirical and theoretical support for this *duality of thought*. Among other phenomena, dual-process views have been proposed for learning, memory, judgment, decision making, and moral, social and deductive reasoning.

Some researchers have sought to unify these views under a general dual-system theory. According to this theory, there are two general systems of cognition. One system (System 1) underlies fast, automatic, low-effort, high-capacity processes (type-1 processes); the other system (System 2) underlies slow, deliberate, high-effort, low-capacity processes (type-2 processes). The list of properties most commonly attributed to each type of process is the following:

Type-1 Processes	Type-2 Processes
Fast/rapid	Slow
Automatic	Controlled
Low effort	High effort
Unconscious	Conscious
Implicit	Explicit
Domain specific	Domain general
High capacity	Low capacity

The dual-system view claims that most (if not all) dual-process views can be subsumed under a general dual-system framework. This means that the duality of processes for memory, reasoning, judgment, decision-making, and other cognitive functions is the result of a general duality of the mind: two systems that work alongside in the brain and that sometimes compete for control of behavior. This view makes some strong claims about the cognitive architecture of the mind: first, it claims that there are two main types of cognitive processes and that they manifest a cluster of opposite properties; second, it states that the two types of processes manifest opposite characteristics because they are due to two different systems (and not just two modes of operation of the same system); finally, it claims that these two systems operate independently, sometimes cooperatively and other times competing for cognitive control.

Though initially popular, the general dual-system view has faced considerable scrutiny in the last decade. The main challenges involve (i) explaining the clustering of properties that characterizes each type of process, (ii) finding a unified and coherent explanation of the difference between the two systems, (iii) showing that the empirical results that support different dual-process views also support a general dual-system view (i.e., showing that there is a general consistency and unity in the data).

Several accounts have been proposed. Some authors explain the differences between the two systems by distinguishing their processes' speed, content, and controllability (Kahneman & Frederick, 2002), other by proposing a distinction in terms of levels of processing (Frankish, 2009), or different involvement of working memory (Evans, 2009). In this dissertation I will show that none of these explanations is sufficient to account for the high variability of properties described in the literature. Further, I will argue that the reason why these (and other accounts) fail to explain the

available data is that there isn't a unified and consistent opposition of two cognitive types but of several types of processes. My position explains not only why the explanations proposed so far are inadequate but also why it has been so difficult to align all the different dual-process views into a unique dual-system theory: no single distinction can explain all the relevant data.

In the dissertation I show that there is no non-ad hoc way of keeping the distinction between the two systems. Appeals to underlying cognitive architecture, the use of working memory, and levels of analysis are all useful to explain some of the data and make sense of the complexity of processes involved in "higher" cognition. However, none of them provides a satisfactory answer to the question: what distinguishes the two systems? My own view is that there is no privileged distinction, no two systems, but a set of properties that are cognitively significant because they identify important differences in how we learn, process, and use information. And even though these differences do not always align, there are some types of processes that are particularly important because they embody the prototypical construal of the contrast between *intuition* and *reasoning*. I will also show that these two 'preferred' types of processes (i.e., slow, effortful, conscious thinking and fast, effortless, unconscious intuition) many times owe the clustering of properties to different underlying cognitive micro-architectures (that is, different computational principles). However, there is no strict correspondence between underlying micro-architecture and phenomenological properties, and this is why there are processes that share properties with the two preferred types. Also, I will claim that at the level of macro-architecture (the general cognitive architecture of the mind) there is another important type of process (type-3 processes) which is involved in conflict resolution and cognitive control. Only recently have these processes been studied independently; before, they were either confused with type-2 processes or simply ignored.

All these factors render the *duality of the mind* hypothesis highly implausible: there are more than two types of processes and more than two kinds of systems under any plausible interpretation. However, dual-process and dual-system views have played a crucial role in advancing our understanding of the complexity of the cognitive architecture of the mind.

The transition from a single, unified view of the mind to a dual one is the result of an important realization: the plurality and complexity of cognitive processes, even at the level of “higher” cognition. This has implications for psychology and philosophy. Psychologically it’s important to realize that processes can be automatic or controlled, conscious or unconscious, implicit or explicit, etc., and that these properties do not always align.

Philosophically, understanding the complexity of the mind and deflating the single mind view has important consequences for discussion on human rationality. Indeed, a *single* mind suggests unison, accord. A *dual* (or plural) mind suggests options, alternatives, and hence potential conflict. The inherent conflict in the complex human mind constitutes an important realization for cognitive science, one that has important consequences for any theory of the mind, particularly normative theories.

Chapter One provides an overview of dual-process theories of deductive reasoning, judgment, decision making, memory, learning, and social cognition. I survey some of the key experiments that supported the development of these views. I draw on research by psychologists, neurophysiologists, economists, and philosophers. I also explain why the use of heuristics leads to systematic errors in human reasoning. Finally, I introduce the generic dual-system view and

summarize some of the most popular versions of the theory. The goal of this chapter is to show the breadth and apparent homogeneity of the experimental results.

In **Chapter Two** I address theoretical and methodological issues regarding the dual-system view. First I make some conceptual remarks about the relation between systems and processes, and their role in modeling the cognitive architecture of the mind. I also distinguish between two types of cognitive architectures, micro and macro-architectures. Then I assess Samuels' (2009b) proposal of understanding dual-system views in terms of a distinction between cognitive kinds: two types of systems (S1 and S2) that underlie two types of processes (type-1 and type-2). Against Samuels, I claim that we need to abandon the idea of a single, unified dual-system view. Instead we should distinguish a duality at the micro-architecture level (a duality of processes) from a duality at the level of macro-architecture (a duality of systems or a duality of the mind).

In **Chapter Three** I introduce the debate about the micro-architecture of the mind. I describe the two micro-architectures—classical and connectionist—and summarize their strengths and shortcomings. Against the original debate, I claim that there are no formal results that can decide between the two architectures and that both architectures are able to model all processes. Instead, I motivate an empirically oriented approach. I suggest that there is enough empirical information from human cognitive tasks as well as artificial intelligence to determine the choice of architecture. Next, I show that a distinction at the micro-architectural level can explain (to a certain degree) the clustering of properties described in the dual-process literature. In particular, it can explain why some processes that are associative are also fast, automatic, effortless, unconscious, and implicit. The same explanation can be provided for processes that are rule-based, slow, deliberate,

effortful, conscious, and explicit. Indeed, one should expect certain properties to cluster in the case of associative processes and the opposite for classical processes. To support this claim, I work through some of the main results in the empirical literature and I show why they are compatible with a *hybrid* architecture, partly classical and partly connectionist. Further, I show that this is not only an empirically adequate solution but it is also well supported by theoretical considerations.

In **Chapter Four** I return to the question of the plausibility of a dual-system view. Having shown that the differences between the two types of processes can be explained in terms of different underlying micro-architectures, I consider whether this distinction can support a dual-system view. I argue that, though robust, this distinction cannot support a general dual-system view, that is, a macro-architectural *duality of the mind*. There are three main reasons. First, the distinction at the micro-architectural level only explains the clustering of properties to a certain degree. Not all type-1 processes can be modeled by a connectionist architecture, and the same can be said for type-2 processes. Secondly, there are independent reasons for separating the properties of a system from the properties of their processes, and there are also reasons for preferring hybrid systems, rather than solely classical or connectionist ones. Finally, there are considerations relating to the macro-architecture of the mind that support a plural rather than a dual-system view. In particular, there are strong reasons to introduce a third type of process related to executive control and conflict resolution. To conclude, I argue that despite its many shortcomings, the dual-system view delivers important insights about the nature of the mind and the complexity of higher cognitive processes, and for this reason it is both philosophically and psychologically significant.

In **Chapter Five**, I discuss an application of the dual-system view to philosophical debate. I consider a very influential set of experiments by Joshua Greene on the psychology of moral judgment and decision making (Greene et al., 2001, 2004, 2008, 2009). First, I describe the experiments and summarize the empirical results. I then criticize some of the authors' choices of methodological design. Finally, I critically assess the implications that these results have for moral philosophy. For example, Greene (2003, 2007) claims that the exercise of cognitive control tends to favor consequentialist judgments, and that this fact shows that consequentialist judgments (and the mechanisms generating them) are more accurate. I disagree. It is far from clear that the exercise of cognitive control favors consequentialism and, even if so, whether this shows the superiority of the consequentialist response. Finally, the authors assume that slow, deliberate, and conscious reasoning is generally more accurate than fast, automatic, effortless intuitive thought. But research shows that there are some reasoning and decision making tasks for which intuitive responses are more reliable. Before we assume that 'rule-based reasoning' is better at solving moral conflict than intuition we need to find an independent way of assessing the correctness of both judgments (assuming there is such a method). Until then, there is no reason to assume, like Greene, that rule-based, conscious, deliberate reasoning will provide more reliable answers in moral judgment and decision making. And the same can be said for all those tasks for which we do not yet have independent means of establishing truth conditions or levels of accuracy.

Chapter One

I. Dual-Process Views

The contrast between two types of thinking, one fast, automatic, and effortless, another slow, deliberate, and effortful is not new in psychology. Over a century ago, William James distinguished between associative thought, which is merely reproductive, and true reasoning, which can help us in novel situations. He also remarked that, “the more of the details of our daily life we can hand over to the effortless custody of automatism, the more our higher powers of mind will be set free for their own proper work” (James, 1890/1950, p. 122). Similar distinctions had already been proposed by philosophers.¹ The everyday distinction between “reason” and “intuition”² also reflects this perceived duality of thought.

During the last four decades there has been a generalized proliferation of dual-process theories in the main areas of cognitive psychology, including memory, learning, reasoning, judgment and decision making, social and moral cognition. In some areas (like social cognition) dual-process theories have become the standard view. Though most dual-process views were developed independently and do not share all of their characteristics, all the theories claim that there are two different processes underlying higher cognitive capacities. According to these theories, one type of

¹ See, e.g., Aristotle (2000), Hume (1739/1978), and Leibniz (1765/1981).

² This notion refers to a cognitive capacity and hence is different from the notion of epistemic intuition: “a label for any immediate (or not explicitly inferential) assessment of any claim of interest to epistemologists” (Nagel, 2007, p. 793). These notions are related of course since epistemic intuitions might be the result of intuitive thinking (i.e., fast and automatic thinking).

process (type-1) is fast, automatic, effortless, and non-conscious, while the other (type-2) is slow, deliberate, effortful, and conscious.

The different dual-process views have been developed independently. Some theories have been reformulated several times and, in some fields, more than one theory has been proposed. This makes the literature on dual-process views too vast and complex to be summarized here.³ Instead I will concentrate on the latest formulations of the leading theories in the following areas: reasoning, judgment and decision making, memory and learning, and social cognition.⁴ I will also review, whenever relevant, the main experimental studies that led to the formulation of the different dual-process views.

Reasoning

Evans' (1984, 1989) heuristic-analytic theory of reasoning was inspired by the results of two sets of experiments. The first was conducted by Wason and Evans (1975)⁵ and focused on the discrepancies between participants' reports and their choices on the Wason selection task. The second was conducted by Evans, Barston, and Pollard (1983) and focused on the influence of belief in the assessment of logical validity.

The Wason selection task is one of the most studied tasks in the psychology of reasoning. It was first introduced by Wason (1966) and has been tested under many different formulations.⁶ In the abstract formulation, participants are asked to solve the following problem:

3 For a review, see Evans (2008), Frankish and Evans (2009), and Frankish (2010).

4 In the last chapter of the dissertation I will assess dual-process views in moral cognition.

5 According to Evans (2006), the terms "dual-process", "type-1", and "type-2" were first mentioned in this article.

6 For a relevant review, see Oaksford and Chater (1994).

There are four cards below. Each of them has a letter on one side and a number on the other side. Two of these cards are shown with the letter side up, and two with the number side up.

A D 4 7

Indicate which of these cards you have to turn over in order to determine whether the following claim is true: *If there is an A on one side of a card, then there is a 4 on the other side.*

Though the logically correct answer is to turn the A and the 7 cards, participants consistently fail to choose this solution (they usually choose to turn the A and the 4 cards or only the A card). The task has been tested hundreds of times and on average only 10-20% of people solves it correctly (Evans, 2003).

In Wason and Evans (1975), the authors recorded responses to both a positive and a negative version of the task. The negative version of the task was exactly the same as the one described above, with the exception that the rule contained a negation on the second part: “If there is an A on one side of a card, then there will **not** be a 4 on the other side” (p. 143). As predicted by the researchers, participants also chose to turn the A and 4 cards⁷, that is, the cards with the letter and number mentioned in the rule. As a result, Wason and Evans (1975) hypothesized that there is a *matching bias* at the root of participants’ responses. Instead of following logical rules, participants

⁷ In this case, the normatively correct response.

match their response to the cards mentioned in the rule, as if the only relevant cards were those explicitly mentioned in the rule.

In Evans, Barston, and Pollard (1983) participants were asked to evaluate the logical validity of a group of arguments, some of which were valid and some invalid. The task was further complicated by the fact that half of the valid arguments had an unbelievable conclusion, based on common sense prior beliefs. These arguments were referred to as ‘conflict arguments’. Though the believability of the conclusion has no influence on the validity of the argument, the experimenters noticed that it creates a cognitive conflict for the participants. A similar observation was made when the argument was invalid but it had a believable conclusion: participants found it harder to see the invalidity of the argument compared to cases where the argument was invalid and the conclusion unbelievable. The following are some examples of conflict and non-conflict arguments taken from Evans et al. (1983, p. 297):

(1) *Valid argument, believable conclusion:*

No police dogs are vicious.

Some highly trained dogs are vicious.

Therefore, some highly trained dogs are not police dogs.

(2) *Valid argument, unbelievable conclusion:*

No nutritional things are inexpensive.

Some vitamin tablets are inexpensive.

Therefore, some vitamin tablets are not nutritional.

(3) *Invalid argument, believable conclusion:*

No addictive things are inexpensive.

Some cigarettes are inexpensive.

Therefore, some addictive things are not cigarettes.

(4) *Invalid argument, unbelievable conclusion:*

No millionaires are hard workers.

Some rich people are hard workers.

Therefore, some millionaires are not rich people.

The fact that participants failed to correctly assess logical validity when the believability of the conclusion conflicted with the validity of the argument shows, according to the authors, that there is a conflict between participants' common sense beliefs and their assessments of logical validity. This phenomenon, the influence (and predominance) of previous knowledge on assessments of logical validity, was labelled *belief bias*.

Evans (1984, 1989) developed a dual-process theory of reasoning in order to account for the biases documented in these experiments. In the original formulation of the theory reasoning is due to the interaction of two types of processes: heuristic processes generate specific representations of the input while analytic processes derive judgments or inferences from these

representations. In this formulation the two processes work sequentially to produce the corresponding mental model:

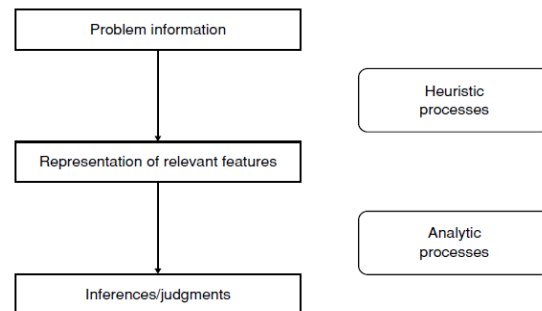


Figure 1. Evans (1984, 1989) heuristic-analytic reasoning model. A representation of the theory in its original formulation. (Taken from Evans, 2006, p. 379)

According to Evans (2006), “biases were accounted for by the proposal that logically relevant information might be omitted or logically irrelevant information included at the heuristic stage. Since analytic reasoning could be applied only to these heuristically formed representations, biases could result” (p. 378).

Based on subsequent experiments, Evans presented a reformulation of the theory in which heuristic processes cue context-dependent representations which generate default responses (Evans, Over, & Handley, 2003). By contrast, analytic processes only sometimes intervene and replace the heuristic system’s default response with another response. This means that the interaction between the two types of processes is sometimes sequential and other times competitive, depending on whether the analytic processes override the default response generated by the heuristic processes. Finally, since the analytic system works on the context-dependent representations generated by the heuristic system, it is still susceptible to bias, even when it inhibits the heuristic response.

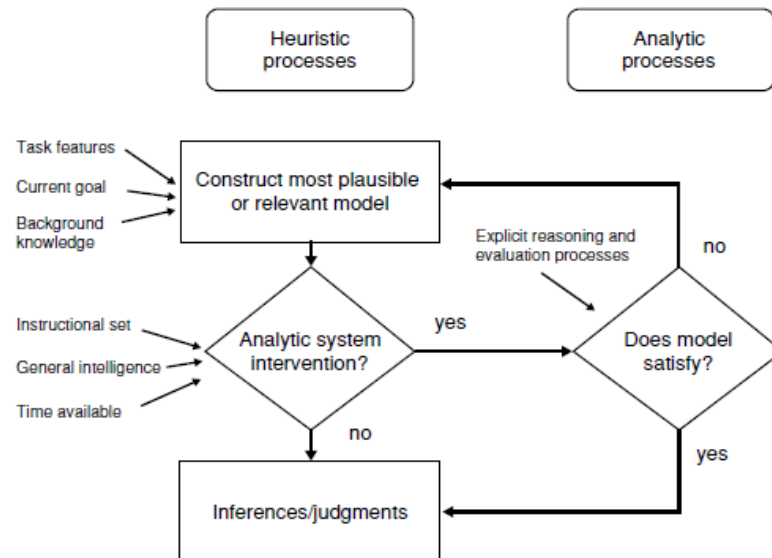


Figure 2. Evans' (2006) heuristic-analytic theory. The revised model shows the interdependence (and sometimes competition) between heuristic and analytic processes. (Taken from Evans, 2006, p. 381)

The revised theory is consistent with recent experimental results showing that working memory capacity and time pressure influence the ability of participants to solve problems correctly (i.e., to ignore the default mental models cued by the heuristic processes)⁸. In addition, the revised theory allows analytic processes to be responsible for a number of reasoning biases.⁹

⁸ I will discuss this issue in more detail below (see Roberts & Newton, 2002; Stanovich, 1999; Evans & Curtis-Holmes, 2005). Evans (2006) also claims that dispositional attitudes influence the performance of the analytic processes (p. 383).

⁹ As Evans (2006) points out, "a key feature of hypothetical thinking is the tendency of the analytic system to hold on to representations that are merely good enough, leading to such characteristic features as the endorsement of fallacious inferences in deductive reasoning and apparent confirmation biases in hypothesis testing" (p. 383).

Judgment and decision making

Since the 1970's, Kahneman, Tversky, and other researchers¹⁰ have consistently shown that judgment and decision making under uncertainty is often due to the automatic use of heuristics, and not the deliberate use of formal rules. In this context, a heuristic is a problem solving strategy that consists in evaluating an attribute that is easier to measure than the target attribute (Tversky & Kahneman, 1974). The replacing (heuristic) attribute is usually a *natural assessment*: a property of the stimulus that is automatically evaluated when it's first processed by the individual (e.g., size, distance, similarity, familiarity, affective valence). Because these characteristics are readily available to the individual, they are much easier to evaluate than the actual target and, consequently, they are ideal candidates to work as heuristic attributes (Tversky & Kahneman, 1983).

Similarly to Evans' view, Tversky and Kahneman (1974) claimed that there is a conflict between judgments and decisions made according to formal rules (such as probability laws) and those relying on heuristics. Though the use of heuristics does not necessarily lead to mistakes, under many conditions it can certainly bias judgment and decision making. Hence the name "heuristics and biases literature" for the research documenting related phenomena.

In what follows, I will describe some of the most influential experiments in this line of research – which now spans over four decades and has been very prolific. In particular, I will focus on a series of experiments that show how people systematically violate probability laws.

1. In Kahneman and Tversky (1973), a group of participants was told that:

¹⁰ See Tversky and Kahneman (1974), Kahneman, Slovic, & Tversky (1982), Gilovich and Griffin (2002). Some of these experiments will be discussed in detail below.

A panel of psychologists has interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100. (p. 241)

Another group of participants was given the same paragraph with the base rate information reversed (70 engineers and 30 lawyers). Participants were then presented with the following two descriptions:

Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles. (p. 241)

Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues. (p. 242)

While Dick's description was created to be neutral as to whether he is a lawyer or an engineer, Jack's description was designed to be very representative of engineers – i.e., his description is close to the stereotype of an engineer. When participants were asked to estimate the likelihood that Jack was an engineer, the estimates were quite high in both groups, regardless of

base rate information. When participants had to assess the likelihood of Dick being an engineer, again participants ignored base rate information. The median estimate was 50% for both groups. This clearly violates probability laws: since the description does not provide any information about the likelihood of the person being an engineer or a lawyer, participants should assess this likelihood based on base rate information (which is 70/30) only. Instead, participants calculated the degree of similarity between Dick and the stereotypical lawyer, the degree of similarity between Dick and the stereotypical engineer, and concluded that both are more or less the same and hence the likelihood of Dick being a lawyer or an engineer is roughly the same.

2. In another experiment, Tversky and Kahneman (1982a) gave participants the following scenario:

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city in which the accident occurred. You are given the following facts: 85 percent of the cabs in the city are Green and 15 percent are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each of the two colors 80 percent of the time. What is the probability that the cab involved in the accident was Blue? (pp. 156-7)

According to probability laws, in order to correctly assess the probability of error in a 'Blue cab' judgment, one needs to take into account prior probabilities (such as the likelihood of seeing a Blue cab). Given that 85% of cabs are Green and 15% Blue, this is the total spectrum of possibilities:

- An 85% chance of seeing a Green cab, out of which 20% will be seen as Blue and 80% as Green.
- A 15% chance of seeing a Blue cab, out of which 80% will be seen as Blue and 20% as Green.

In sum, there is a 17% chance that the witness saw a Green cab as Blue, and a 12% chance that she saw a Blue cab as Blue. So, one should only be 41% confident on the witness' assessment that the Blue cab was actually Blue. However, most participants ignore base rate information and report a certainty considerably above 41% (usually at around 80% since the problem states that the witness is 80% reliable when it comes to identifying cab colors).

3. In an experiment by Tversky and Kahneman (1983) a group of undergraduates from UBC was asked to estimate how many words of the form '----ing' they would expect to find in a 2000 words text. Another group of undergraduates was asked to estimate how many words of the form '-----n-' they would expect to find in a 2000 words text. Though the second type of word includes the first, the median estimate for the first group was 13.4 words while the median estimate for the second group was 4.7 words. Similar results were recorded when comparing words of the form '----ly' and of the form '----l-' (with median estimates of 8.8 and 4.4 respectively).

4. In one of the literature's most famous experiments—known as 'Linda the bank teller'—participants were given a detailed description of a person named Linda and they were asked to judge the likelihood of a number of related propositions (e.g., how likely it was that Linda had a certain profession or played a certain sport). Here is the description provided:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations (Tversky & Kahneman, 1982b, p. 92).

Participants were then asked to rank the likelihood of several propositions, including the following three:

- Linda is active in the feminist movement. (F)
- Linda is a bank teller. (T)
- Linda is a bank teller and is active in the feminist movement. (T&F)

According to probability theory, the probability of a conjunction can never be higher than the probability of either conjunct. So, the probability of T&F cannot be higher than that of T (or F). Nevertheless, experimenters found that most participants did not adhere to this rule. Rankings were of the form $F > T \& F > T$ for 85% of participants.¹¹ Further, in 89% of the cases participants ranked the conjunction as more likely than the proposition that expressed the unrepresentative feature (T). There was also a high correlation (.98) between mean rankings of likelihood and mean rankings of representativeness.¹²

Later, Tversky and Kahneman (1983) performed a similar experiment with doctors and medical students in an effort to determine whether the results of the original experiment were due to participants' inexperience with probabilistic problems, or to a disposition shared by novices and experts alike. The results were clear: the *expert* participants made the same kind of probabilistic

11 Participants were undergraduate students at the University of British Columbia.

12 Similar results were obtained for the following experimental setting (Tversky & Kahneman, 1982b, p. 92):

Bill is 34 years old. He is intelligent but unimaginative, compulsive, but generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.

- Bill is an accountant. (A)
- Bill plays jazz for a hobby. (J)
- Bill is an accountant who plays jazz for a hobby. (A&J)

Rankings were of the form $A > A \& J > J$ in 87% of participants. In 92% of the cases, participants ranked A&J as more likely than J. Finally, the correlation between mean rankings of likelihood and mean rankings of representativeness was .96.

mistake. When confronted with the mistake, most of them did not attempt to justify it but were instead perplexed that they would have made such an “elementary error” (p. 302).

5. Casscells, Schoenberger, and Grayboys (1978) presented the following problem to a group of faculty, medical students, and staff at Harvard Medical School:

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?
_____ % (p. 999)

A disquieting 45% of respondents completely ignored base rate information (such as the prevalence of the disease) and reported a 95% as their answer (quite far from the Bayesian response, 2%). Only 18% of the respondents gave an answer close to the correct one.

These and other experiments show that people’s probability assessments can be subject to different biases, such as judging a conjunction to be more probable than one of its conjuncts (conjunction fallacy) or ignoring base rate probabilities (base rate neglect). Further, we have seen that people’s assessments of logical validity are subject to belief¹³ and confirmation¹⁴ biases. These are not isolated results, the heuristics and biases literature has shown that people are predisposed to many other biases when making judgments and decisions under uncertainty.¹⁵

13 Letting common held beliefs influence validity assessments.

14 Looking for signs of confirmation (instead of refutation) when testing a hypothesis.

15 Some of the heuristics and biases documented in the literature are: hindsight bias, overconfidence effect, non-regressive predictions, my-side bias, inaccurate co-variation estimation, inappropriate anchoring, illusory correlation, belief perseverance, preference reversals, outcome bias, commission bias, failure of inconsistency detection, failure to generate alternative explanations, unrealistic optimism, etc. For a thorough review, see Gilovich, Griffin, and Kahneman (2002).

Though I have only focused on a handful of experiments, related research has shown remarkable consistency across results. Further, researchers in the heuristics and biases tradition have noticed that when participants provide an answer that doesn't conform to normative accounts of rationality it is usually the result of a fast, automatic, and intuitive kind of reasoning process. By contrast, normatively correct responses seem to be the consequence of participants' conscious and effortful application of rules.

A simple experiment by Frederick (2005) has shown this phenomenon strikingly. Undergraduates at several institutions were given the following problem:

A bat and a ball cost \$1.10 in total, the bat costs \$1 more than the ball. How much does the ball cost? (p. 26)

Most students report an initial tendency to answer "10 cents".¹⁶ This answer just *springs to mind*: it is fast, automatic, and it is not clear where it comes from. On the other hand, when they take a moment to reflect on the problem, they realize that it is not the correct answer.¹⁷ In order to get the right answer they need to engage in conscious algebraic reasoning. This is a very simple yet compelling example of how two different reasoning processes seem to coexist in the mind. It also shows how the two processes have very different phenomenologies.¹⁸

16 In verbal reports, participants told the experiments that they first considered the answer "10 cents" and then changed it to "5 cents".

17 Almost all students that do not answer "10 cents" end up providing the correct response, i.e., "5 cents" (Frederick, 2005, p. 27). However, many students do not override the *intuitive* response. In his Nobel Prize Lecture, Kahneman reports that 50% of students at Princeton University and 56% of students at the University of Michigan gave the wrong answer (Kahneman, 2003, p. 1450).

18 I'm sure that if the students were told about their mistake, they would probably show the same kind of reaction that expert participants showed when told about their probability mistakes (i.e., wonder how they could have made such an elementary mistake). In fact, most of the experiments in the dual system literature are such that people can understand and see the

In sum, these experimental results (and many more) seem to describe the same kind of duality of thought. In order to account for these similarities, several researchers¹⁹ have postulated the existence of two systems of reasoning working in parallel. One system is responsible for the automatic, rapid, effortless, associative, and unconscious processes, while the other system underlies the deliberate, slow, effortful, rule-based, conscious processes. I will describe this dual-system view in more detail below.

Memory and learning

Initially memory and learning were studied independently of each other in cognitive psychology. But since the 1950's psychologists have sought to provide models that simultaneously account for the learning process and the memories formed as a result.²⁰ Hence, advances in the understanding of how we store different types of knowledge have led to insights into how we learn that knowledge. One major advance in the area was the realization that there is not a unique memory system in the brain: a long list of experiments in psychology and neuroscience has shown that there are several memory systems in the brain (see Schacter & Tulving, 1994; Eichenbaum & Cohen, 2001).

correctness of the normative answer once it is explained to them. Rather than incomprehension, the problem seems to be that participants do not engage the correct reasoning mechanism when trying to solve these puzzles.

19 Evans (1989), Sloman (1996), Stanovich (1999), Smith and DeCoster (2000), Kahneman and Frederick (2002).

20 The connection between the two is clear in the study of patients with localized brain damage. The inability to learn a specific type of knowledge is correlated with damage to areas that store that type of knowledge, but it does not imply the inability to recall similar types of memories if they were formed before the damage occurred. This lends support to the claim that when a memory system is impaired, the corresponding learning mechanism is also impaired.

One important distinction is that between procedural and declarative memory, which roughly corresponds to the distinction between *knowing how* (procedural memory) and *knowing that* (declarative memory). Procedural memory includes skills, habits and behaviors, while declarative memory is divided into semantic memory (memories of particular facts, e.g., that Paris is the capital of France) and episodic memory (memories of particular events, e.g., a specific summer trip to Paris). Emotional memory constitutes a third system which stores associations between different stimuli and positive or negative emotional valence (i.e., reward versus aversion emotional states). The basis for the distinction between these three memory systems is dissociation in processing and brain localization.²¹ From different experimental results, it has been concluded that the amygdala is associated with emotional memory, the hippocampus with declarative memory, and the striatum with procedural memory.

Researchers have also distinguished between long-term memory (which includes all of the above) and short-term memory, which is a limited time storage that keeps information active before it is permanently stored. Finally, short-term memory (or parts of it, at least) has been re-described as working memory: a capacity to store information that is needed to process concurrent tasks.²²

The insight that we store different types of memories in different systems led to the supposition that learning involves different types of processes too. As a result, new distinctions were proposed. For example, the distinction between implicit and explicit memory systems is founded

21 These two factors have been tested in humans and rats: damage to the brain area associated with each memory system results in difficulties in learning/recall of that particular type of memory and not the others. For a review, see Eichenbaum and Bunsey (1995).

22 This view considers working memory as more related to attentional processes than to memory formation processes. For an overview of theories of memory, see Schacter and Scarry (2000).

on learning studies by Arthur Reber. Reber (1965, 1967) developed a dual-process view of learning that distinguishes implicit and explicit learning processes.²³ According to this view, implicit learning occurs unconsciously and without a deliberate attempt to learn. Further, implicit learning produces abstract or structural knowledge about the stimulus.²⁴ By contrast, explicit learning takes place when we make a conscious effort to learn some piece of knowledge, whether it is a fact, a rule, or a skill.²⁵

At the time the commonly held view was that learning (particularly for complex knowledge) required conscious effort. Reber's research focused on processes leading to implicit learning of complex stimuli, and the nature and extent of the resulting knowledge. Two of his most famous experiments (Reber 1965, 1966, 1967) were designed to study the implicit learning of grammatical rules and probabilistic dependencies.

Grammar Learning

In the first experiment, Reber (1965, 1967) used a synthetic grammar to produce a number of basic sentences.

23 This is one of the earliest dual-process views in contemporary psychology.

24 Reber (1989, p. 219): "Implicit knowledge results from the induction of an abstract representation of the structure that the stimulus environment displays, and this knowledge is acquired in the absence of conscious, reflective strategies to learn."

25 Examples would be consciously learning the capitals of Europe, the conjugation rules of verbs in a foreign language, or how to tie a special kind of knot.

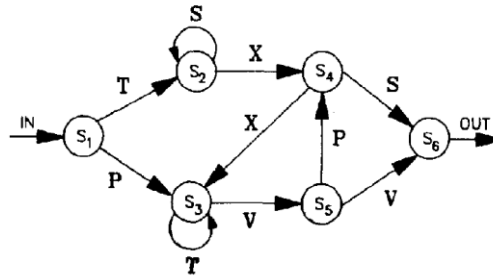


Figure 3. Synthetic Grammar. A sample synthetic grammar used in Reber's (1965, 1966) experiments. (Taken from Reber, 1967, p. 856)

During the acquisition phase participants are exposed to some of these sentences but not to the rules of the grammar. Later, during the testing phase, participants are exposed both to some of the basic sentences as well as ungrammatical sentences. For each of these sentences, participants are asked to assess whether it conforms to the rules of the grammar (rules to which they have not been explicitly exposed). Results in many replications of this type of study have shown that participants respond well above chance, showing implicit learning (learning without conscious effort) of the rules of the grammar. The fact that participants are unable to state the rules of the grammar provides further evidence that knowledge of the rules is implicit.²⁶

Probability Learning

In the second experiment (Reber, 1966), participants were exposed to a rapid sequence of events with an underlying stochastic structure. In the testing phase, participants were asked to make predictions about the occurrence of future events. The results showed that their performance was

²⁶ In much the same way that native speakers of a language can tell whether a sentence is grammatical but cannot explicitly state the rule that makes it grammatical.

well above chance. Reber (1966, 1989) hypothesized that this is due to the acquisition of tacit knowledge about the structure of the environment. That is, in the acquisition phase, participants become increasingly sensitive to the stochastic structure of the environment so that they learn to anticipate events with high accuracy. This, in turn, leads to an improvement in their assessment of the probability of a certain event, though it is not in itself an effect of learning probabilities but rather of exploiting environmental structure.

What passes in the literature for probability learning is actually a much more subtle process in which participants learn implicitly about the stochastic structure of an event sequence to which they have been exposed. In the course of making predictions, they mimic its structure and thereby generate a sequence of responses, one by-product of which is an approximate matching of the probabilities of the events—that is, *probability learning*. (Reber, 1989, p. 220)

Reber's research has played a key role in understanding the connection between different types of learning and different types of knowledge. In particular, it has been instrumental in explaining why people can solve tasks for which they have no apparent knowledge or instruction. A further conclusion derived from this line of research is that implicit knowledge can influence reasoning, judgment, and decision making without conscious awareness (Reber, 1989, 1993).²⁷

Briefly, another relevant dual-process view in the field of memory is that developed by Reyna and Brainerd (1995). The authors propose a distinction between verbatim and gist memories. Verbatim memories are precise and quantitative representations of the original event which quickly

²⁷ I will explain this in more detail below, in the section on social cognition.

decay in time. Gist memories are less precise (fuzzy) traces of experience which carry rich semantic content that is related to the individual's knowledge, culture, and level of development. Both types of memories can be used in judgment and decision making, producing very disparate results (Reyna & Brainerd, 1995).

Social Cognition

Since the 1980's psychologists have used dual-process theories to explain several phenomena in social psychology. These dual-process models of social cognition share an assumption with the results that I summarized in the previous section: that implicit knowledge can (and usually does) influence explicit judgments and decisions. Though most of the research was developed in isolation from studies in cognitive psychology, the conclusions are strikingly similar: people have implicit and explicit social attitudes and they are not always consistent. Though people verbalize and consciously hold certain social attitudes, their judgments and behaviors are sometimes bent by inconsistent implicit attitudes, of which they are not aware.

One of the most influential theories in the field was developed by Bargh (1989, 2006), who claimed that there are automatic and controlled processes involved in social judgment and behavior. Stereotypes associated with a stimulus are automatically invoked and compete with a person's explicitly stated social beliefs, often leading to a dissociation between verbalized attitudes and actual behavior. According to Bargh (1999, 2005), it is very difficult to override the automatic processes and this is the reason for the pervasive influence of stereotypes in behavior.

Support for this view often comes from *priming* studies. In Bargh, Chen, and Burrows (1996) a group of participants was primed with words usually associated with the elderly (e.g.,

Florida, old, lonely, grey, retired) while the other was primed with neutral words. The researchers noticed that participants in the first group walked more slowly after the task. In another experiment by Bargh et al. (1996), a group of undergraduate students was primed with impolite words (e.g., *aggressively, bold, rude, bother, disturb*) while another group was primed with neutral words. The authors noticed a higher tendency to interrupt the experimenter in the first group.²⁸

This dual-process model has been applied to other areas of social cognition, including persuasion and attitude change (Chen & Chaiken, 1999), prejudice and stereotyping (Devine, 1989), and attitudes and evaluations (Wilson, Lindsey, & Schooler, 2000).

II. The Transition from Dual-Process to Dual-System Views

So far I have shown that dual-process views have been proposed for memory, learning, judgment, decision making, reasoning, and social cognition. These views usually contrast a fast, automatic, effortless, and unconscious type of process with a slow, deliberate, effortful, and conscious type of process.

Given the breadth and consistency of the empirical results, a number researchers²⁹ sought to unify the particular dual-process theories into a general dual-system view of cognition. The general dual-system view seeks to account for all (or at least most) dual-process views and make stronger, more general claims about the cognitive architecture of the mind. According to this view, there are two cognitive systems: one system underlies fast, automatic, effortless, and unconscious processes, while the other underlies slow, deliberate, high effort, conscious processes.

²⁸ For a review of similar experiments in the social cognition literature, see Smith and Collins (2009).

²⁹ Among others, Evans (2003, 2009), Stanovich (1999, 2004), Frankish (2004), Carruthers (2006), Sloman (1996).

The two systems were given different names by different authors: implicit and explicit (Reber, 1993; Evans & Over, 1996), experiential and rational (Epstein, 1994), heuristic and systematic (Chaiken, 1980), heuristic and analytic (Evans, 2006), associative and rule-based (Sloman, 1996), reflexive and reflective (Lieberman, 2003). Later, following Stanovich (1999), researchers started using the labels ‘System 1’ and ‘System 2’ to refer to the two systems. I will follow this practice and use the labels ‘S1’ and ‘S2’ for short.

A brief overview of the most significant and recent dual-system views follows.

Kahneman’s view

According to Kahneman and Frederick (2002), System 1 uses heuristics while System 2 relies on rule-based reasoning. Further, S1 is automatically engaged and passes its response to S2, which sometimes intervenes and overrides S1’s response, replacing it with its own. Hence, the interaction between the two systems is *default-interventionist*:³⁰ S1 acts by default and S2 intervenes only when needed.

According to the authors, what differentiates the two systems is: (i) their speed (slow vs. fast), (ii) their controllability (automatic vs. deliberate or controlled), and (iii) the content on which they act (affective, concrete, specific, prototypic, and causal propensities for S1; neutral, statistical, abstract, and set theoretic for S2).³¹ They also claim that the capacity of both systems is equal and that cognitive processes can “migrate” from one system to the other as expertise and skill

³⁰ This expression was coined by Evans (2003).

³¹ Kahneman and Frederick (2002): “We use *systems* as a label for collections of processes that are distinguished by their speed, controllability, and the contents on which they operate” (p. 51).

are acquired (for example, recognizing chess positions can change from being an S2 process to being an S1 process).

In recent work, Frederick (2005, p. 26) described the two types of processes in this way:

- Type-1: executed quickly with little conscious deliberation; occurs spontaneously and does not require or consume much attention; it is unaffected by intellect, alertness, motivation or the difficulty of the problem being solved.
- Type-2: executed slowly and more reflectively; the kind of mental operations that require effort, motivation, concentration, and the execution of learned rules.

In 2011, Kahneman published *Thinking, Fast and Slow*, a very influential book on the dual-system view in which he claims that the central difference among the two systems resides in the amount of effort, attention, and control that the two types of processes require.³²

System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration. (Kahneman, 2011, p. 20)

For Kahneman (2011), the central contrast is between fast, automatic, unconscious, low-effort processes, and slow, deliberate, conscious, effortful processes. Here is a list of examples of both types of processes:³³

³² “The highly diverse operations of System 2 have one feature in common: they require attention and are disrupted when attention is drawn away.” (Kahneman, 2011, p. 22)

³³ Regarding the processes listed for System 2, Kahneman (2011) says: “In all these situations you must pay attention, and you will perform less well, or not at all, if you are not ready or if your attention is directed inappropriately” (p. 23).

SYSTEM 1	SYSTEM 2
➤ Detect that one object is more distant than another. ➤ Orient to the source of a sudden sound. ➤ Complete the phrase “bread and...” ➤ Make a “disgust face” when shown a horrible picture. ➤ Detect hostility in a voice. ➤ Answer to $2 + 2 = ?$ ➤ Read words on large billboards. ➤ Drive a car on an empty road. ➤ Find a strong move in chess (if you are a chess master). ➤ Understand simple sentences. ➤ Recognize that a “meek and tidy soul with a passion for detail” resembles an occupational stereotype.	➤ Brace for the starter gun in a race. ➤ Focus attention on the clowns in the circus. ➤ Focus on the voice of a particular person in a crowded and noisy room. ➤ Look for a woman with white hair. ➤ Search memory to identify a surprising sound. ➤ Maintain a faster walking speed than is natural for you. ➤ Monitor the appropriateness of your behavior in a social situation. ➤ Count the occurrences of the letter a in a page of text. ➤ Tell someone your phone number. ➤ Park in a narrow space (for most people except garage attendants). ➤ Compare two washing machines for overall value. ➤ Fill out a tax form. ➤ Check the validity of a complex logical argument.

Table 1. Examples of processes produced by each system (Kahneman, 2011, p. 22).

According to Kahneman (2011), the dual-system view explains why people sometimes make very simple (and surprising) mistakes when making probability judgments, assessing the validity of an argument, or even solving a simple algebraic puzzle: they engage in the *wrong type* of reasoning. They use the fast, automatic, and consciously opaque system when they should engage System 2 instead. And even if they are not sure of how they reached this answer, they can certainly *feel its pull*.³⁴ Further, Kahneman claims that S1 gets automatically engaged in any reasoning task, and that it is S2's general function to oversee, authorize, and (sometimes) override S1's response. This means that depending on the task and other factors (e.g., general intelligence, time constraints, and motivation), S2 will sometimes prevent S1's answer from being verbalized and will instead provide its own response.³⁵

Stanovich's view

According to Stanovich (1999, 2004), System 1 is not a unique system but a set of independent sub-systems, The Autonomous Set of Systems (TASS). Many of the systems in TASS are modular,³⁶ that is, informationally encapsulated and cognitively impenetrable. These systems encode most of our pragmatic and common sense knowledge and help us solve tasks that have a strong contextual component. In Stanovich (1999), System 2 was hypothesized to be a general

34 Frederick (2005); Kahneman and Frederick (2002).

35 See, e.g., the 'bat and ball' problem described above.

36 Here 'module' needs to be taken in a similar sense to that proposed by Fodor, but definitely not in the exact same way. According to Fodor (1983), modules are domain specific innately-specified processing systems of the mind. The key feature of a module is that it is informationally encapsulated: it neither has access to other modules' information nor it provides information to them. It is also domain-specific: it deals with a reduced and particular kind of information. Finally, according to Fodor, the results of the computations of a module are non-conceptual outputs. This is because Fodor only takes input and output systems as modular and not the general thinking capacity. Psychologists have modified this notion to include modules that are involved in reasoning processes. Evolutionary psychologists have gone further and extended the notion to the whole of the mind's processes. This is known as the Massive Modularity Thesis. For reasons of space, I will not elaborate on this last issue.

domain thinking capacity that intervenes when the task is logical or abstract. Later Stanovich (2009) divided S2 into two different sub-systems: the algorithmic mind and the reflective mind. Stanovich's proposal is supported by studies on individual differences: depending on measures of general intelligence and burden on working memory, different systems will get engaged in solving a particular task.

Evans' view

Evans (2009, 2010) has the most recent and sophisticated view. According to Evans (2009), we should distinguish two *minds*, rather than two general cognitive systems. A mind is “a high-level cognitive system capable of representing the external world and acting upon it in order to serve the goals of the mechanism” (p. 35). These processes give rise to higher cognitive functions, including learning, memory, reasoning and decision making. Up until recently, the standard psychological view (based on the folk psychology view) claimed that there was only one mind in control of all higher cognitive functions. This control is exerted consciously and is based on conscious beliefs and desires. Drawing on empirical results from dual-process views, Evans (2009, 2010) opposes the common-sense view and claims that there isn't only one, but **two** minds in charge of higher cognition: learning, reasoning, judgment, and decision making are all dually controlled.

In Evans' (2010) view, one mind (the old mind) underlies all the fast, automatic, and low effort processes. It is composed of several systems and its processes are similar to those in non-human animals. It is called “the old mind” because it involves older brain areas (areas that developed earlier in evolution) and produces those cognitive processes that are also seen in non-human animals. By contrast, “the new mind” extends over the most recently developed areas, in particular,

the prefrontal cortex and all the processes sub-served by this area. Evans' proposal is the most sophisticated because it attempts to combine data supporting most dual-process theories (and not just those from judgment and decision making). It also uses facts about our evolutionary history to explain why there are two minds and how they compete –and sometimes collaborate– to produce the experience of a unified mind.

III. Summary on dual-system views

According to the general dual-system view, cognition is the result of the interaction of two systems, System 1 and System 2. System 1 is responsible for the fast, automatic, and effortless cognitive processes (type-1) while System 2 is in charge of slow, deliberate, and effortful processes (type-2).

It has also been claimed that while S1 is an evolutionarily old system common to humans and other animals, S2 is characteristically human – it is the system that allows for abstract and hypothetical thinking (Evans, 2003; Carruthers, 2006). Some also conjecture that S1 is constituted by a set of autonomous modules that are mainly domain specific (Stanovich, 1999; Bolender, 2001; Carruthers, 2006) and its performance is independent of individual differences in general intelligence (Stanovich, 1999). By contrast, S2 is alleged to be a more recently developed system for abstract reasoning, a domain general system, whose capacity is correlated with general intelligence and working memory span. In some cases the two types of processes are thought to be complementary (e.g., implicit and explicit learning), sometimes competitive (e.g., logical vs.

heuristic reasoning). Other times, factors such as cognitive ability, thinking styles, and environmental cues may result in the two processes operating alternatively.

TYPE-1	TYPE-2
Fast	Slow
Automatic	Controlled
Low effort	High effort
Unconscious	Conscious
Implicit	Explicit
Intuitive	Analytic
Evolutionarily old	Evolutionarily recent
Shared with animals	Uniquely human
Independent of general intelligence	Dependent on general intelligence
Domain specific	Domain general
Contextualized	Abstract
Parallel	Serial
Associative	Rule-based

Table 2. Type-1 and type-2 processes. A list of the properties most commonly attributed to each kind of process.

However, it is not as easy as it seems to unify all the various empirical results into a single dual-system view. First, many of the dual-process views add some additional properties that do not cluster well with the other properties in the list. Even worse, sometimes one researcher says that S1

has a certain property but another researcher says the opposite (either that it doesn't have it or that S2 has that property). If we take into account the full list of characteristics associated with each type of process, it is particularly difficult to come up with a single, consistent dual-system view. There is no unique, defining list of features that characterizes each type of process and that can support a distinction at the system level.

Secondly, there is a concern about the robustness of the dual-system proposal. Even if all of one system's processes shared a similar list of properties, the source of that similarity would remain unknown: it would be an open question whether each system is a collection of processes with similar properties or whether there is some underlying reason for the clustering. Indeed, any arbitrary list of features would divide cognitive processes in two: processes that have those features and processes that don't. Hence, we could conclude from *any* arbitrary list of processes that there are two cognitive *systems* underlying such distinction. For the theory to have some theoretical import, we need to show that the dual-system view explains in some important way the clustering of properties seen in the literature. We would also need to show that this explanation is stronger than competing explanations – for example, the view that there is only one system with two modes of processing.³⁷ Another competing view is the one that claims that there are many systems, some which produce type-1 processes, some that produce type-2 processes, and some that produce other types of processes that have not been considered in the literature so far.

³⁷ The single system view claims that cognition is due to a general system which can give rise to type-1 and type-2 processes, depending on certain factors (for example, motivation, time constraints, etc.). Gigerenzer (1996) argues against the dual-system view.

The question of what makes S1 a *different* system than S2 needs to be given a proper response, an explanation that goes beyond stating a list of properties for each type of process. Unfortunately, most dual-system views rely on this type of argument to ‘show’ the existence of the two systems. In Chapter Two, I will explain why we need a different approach to address the question of the duality of cognition. I will show that current dual-system views cannot satisfactorily account for the set of empirical results described above and that a different view of the cognitive architecture of the mind is needed, one that includes more than just two systems and two types of processes.

Chapter Two

I. The Dual-System View

The studies described in Chapter One show some remarkable consistencies. They show that higher-level cognition (e.g., reasoning, judgment, decision making) can sometimes be the result of fast, automatic, unconscious processes, and other times of slow, deliberate, conscious processes. These processes can feel effortless or difficult, intuitive or analytic, concrete or abstract. Indeed, cognitive processes contrast in all kinds of ways: implicit vs. explicit, associative vs. rule-based, parallel vs. serial, old vs. new, animal-like vs. uniquely human, domain-specific vs. domain-general.³⁸

Based on the extensive research on dual-process theories, some researchers have claimed that there is a fundamental *duality* to the mind and that this duality is due to the existence of two cognitive systems (Kahneman and Frederick, 2002; Bolender, 2001; De Neys, 2006; Evans, 2006; Sloman, 1996; Smith & DeCoster, 2000). The standard dual-system view claims that one of the systems (S1) employs automatic, fast, effortless, associative, parallel, and consciously opaque processes, while the other system (S2) underlies deliberate, slow, effortful, rule-based, sequential, consciously transparent processes. Further, S1 is supposed to be an evolutionarily old system common to humans and other animals. In contrast, S2 is thought to be uniquely human, it is the system that allows for abstract and hypothetical thinking, and is considered to be a recent

³⁸ Please refer to Table 2 in Chapter One.

development in the human evolutionary time scale. Some (Bolender, 2001; Stanovich, 1999) also conjecture that S1 is constituted by a set of autonomous modules that are mostly domain-specific, whose competence is independent of individual differences in general intelligence. S2, on the other hand, is correlated with general intelligence and working memory span.

There are many different versions of the dual-system view. Some proposals emphasize the two systems' different evolutionary histories (Stanovich, 1999, 2004),³⁹ brain localization (Goel, 2003, 2007, et al. 2000; Lieberman, 2000, 2003, 2007),⁴⁰ type of representation, domain specificity vs. domain generality (Carruthers, 2006),⁴¹ processing modes (Sloman, 1996), controllability and use of working memory (Evans, 2009). But none of these accounts has been successful in providing an empirically adequate dual-system view.

Indeed, the question remains: is there a unifying principle underlying these cognitive dualities? Is it possible to formulate a general theory capable of explaining all the different dual-process views? Some authors (Keren and Schul, 2009; Samuels, 2009b) have pointed out that there are several issues that need to be addressed for a dual-system view of cognition to become a robust scientific theory.

The first problem that all these theories face is the diversity of the views proposed. As mentioned, different theories attribute different properties to each system and offer diverse

39 Stanovich (1999, 2004, 2009) claims that the differences between the two systems can be explained by their different evolutionary history. He also claims that accuracy in solving formal problems (such as mathematical, logical, and probabilistic puzzles) is related to the use of System 2 and that its capacity is correlated with general intelligence and the capacity or span of working memory. I believe this last point is partly based on a circular argument since standard intelligence tests usually measure performance on the kind of rule-based thinking that is attributed to System 2.

40 These researchers propose that the two systems operate through different neural pathways in the brain. However, identification of the brain areas that are the domain of the different systems processes has been difficult so far.

41 That is, whether a cognitive process can take several types of inputs or only a specific type.

explanations for the distinction. There is very little consensus in the literature regarding what distinguishes the two systems. Hence, for the dual-system view to remain a plausible hypothesis of mental architecture, there has to be a unification of at least a critical number of theories or, if this is not possible, a process of ruling out the non-viable hypotheses.

Secondly, assuming that it is possible to formulate a single, unified account of the duality of the mind, it remains to be shown that this particular theory is in agreement with the available data. That is, a unified dual-system view must also account for the empirical results described in Chapter One (i.e., dual-process views of memory, learning, reasoning, decision making, etc.).

Lastly, even if the proposed theory were able to incorporate most individual dual-process views and explain the majority of the empirical data, it should also suggest falsifiable conditions (i.e., conditions under which the theory is shown to be wrong). The dual-system view is an empirical theory of psychology, not a mere theoretical construct;⁴² one should be able to formulate conditions that would falsify it.⁴³

Unfortunately, formal discussions of meta-theoretical aspects of dual-system views are rather scarce. Among the few contributions, Samuels (2009b) stands out as the main exponent of this type of work.⁴⁴ Samuels argues that it is possible to formulate an empirically falsifiable dual-system view. In fact, he believes that it is possible to address the first concern and provide a unified dual-system view. He does not believe that we have addressed the second concern (i.e., provided

42 “Dual-process and dual-system theories are empirical theories of human psychology” (Frankish, 2010, p. 914).

43 Keren and Schul (2009) think that this is impossible and that researchers should dismiss dual-system theories altogether.

44 There have been other fruitful discussions in philosophy regarding dual-system views. For example, Dennett (1978), Frankish (2004), Carruthers (2006), and Gendler (2008) have developed dual-process views of belief. Frankish and Carruthers have also proposed dual-system views of cognition.

an account of how this unified dual-system theory explains the different empirical results) nor the third (i.e., established the theory's falsifying conditions).

In what follows, I will critically assess Samuels' (2009b) defense of the project of developing a dual-system view of cognition. I will argue that it is not possible to formulate such a view: no dual-system view can successfully account for the empirical results described in the previous chapter, at least not a substantial part of them. I will claim that we can explain certain dualities in the mind but not in terms of one unified dual-system view. I surmise that in order to make progress on the second issue (linking the empirical results to a theory of cognitive architecture) we need to abandon the idea of a single, unified dual-system view. Instead, we will have to distinguish a duality at the micro-architectural level (a duality of processes) from a duality at the level of macro-architecture (a duality of systems or a duality of the mind).

Before I move on to assessing Samuels' (2009b) proposal, there are a few conceptual remarks to be made about the relation between systems and processes, and their role in modeling the cognitive architecture of the mind.⁴⁵

II. Systems and Processes

The label 'system' is used widely and liberally in cognitive science, often incorrectly and without consideration for usage. In particular, it has not been formally established in which ways a system differs from a mechanism, a process, a structure, and other related entities. In their paper "What are the memory systems of 1994?" Schacter and Tulving (1994) address this issue and seek to

⁴⁵ As Samuels (2009b) points out, dual-process theories "are in the business of providing (at least partial) answers to questions about what cognitive systems there are, and what properties they possess" (p. 129).

distinguish *memory systems* from *memory processes* (e.g., encoding, rehearsal processes) and *types of memory* (e.g., verbal, semantic memories).⁴⁶ The authors state that “a memory system is defined in terms of its brain mechanisms, the kind of information it processes, and the principles of its operation” (Schacter & Tulving, 1994, p. 13). They later add that “the properties of any system include rules of operation, kind of information, and neural substrates” (ibid., p. 16).

Similarly, Keren and Schul (2009) claim that “a process is conceptually a different unit of analysis than a system” and that a system is “an entity that transforms one state (an input) into another state (an output)” (these transformations determine the way in which the task is performed). They also state that “processes are more specific than systems, and yet, abstractly, the same process can run on different informational contents in different systems. Thus, processes and systems are not the same and must be distinguished theoretically” (p. 534).

It seems that both approaches are committed to these claims:

- A system is defined by its neural substrates, its informational content, and its rules of operation. Hence, a system is an independent neural mechanism that transforms inputs into outputs by using specific rules of operation (or “computational principles”⁴⁷).
- Processes and systems are different entities. A process usually describes the rules of operation of a system. Though *process* is a more specific notion, it is also more generally applied (e.g., the same process can be employed by different systems).

⁴⁶ Schacter and Tulving (1994) also think that the distinction between implicit and explicit memory does not refer to two different systems but to two ways of encoding information.

⁴⁷ The terminology used by Samuels (2009b).

In sum, a system is individuated/identified by its processes, its informational content, and its neural substrates.⁴⁸ In general, there are a number of ways to elucidate whether a process belongs to one system or another. Two of them are based on neuro-physiological data. For example, if area A is damaged and, as a consequence, process B is damaged but not process C, then B (but not C) is part of the system neutrally located in area A. Similarly, if area A lights up (for example, in fMRI⁴⁹) when we perform a task that is solved by engaging process B, then process B belongs to the system neutrally located in A. A more “cognitive” way of distinguishing between processes and their corresponding systems is by modeling performance and human errors.

III. Micro and Macro-Architectures

Further complexity arises from positing dual-system views as theories of cognitive architecture, a notion that can also be interpreted in different ways. On the one hand, a cognitive architecture can be understood as a global theory of cognition that integrates the models of several cognitive systems. In this case, a model describes the workings of a particular system and a cognitive architecture provides a general, unified explanation of the workings of all systems. It also provides a way of connecting the different models, showing their underlying similarities. Samuels (2009b) refers to this notion of cognitive architecture when he points out that “a central explanatory

48 The way to identify the neural substrates of a system is usually through dissociation studies: if damage to area A does not entail damage to a certain process but damage to area B does (and viceversa), then the two areas are considered part of different systems. For example, Broca’s and Wernicke’s areas are both involved in language performance but they are considered different systems because damage to one does not correlate with problems in performance for the other’s associated processes.

49 Functional magnetic resonance imaging.

objective of much cognitive science is to provide an account of our mental architecture: to characterize the various systems and structures from which the human mind is composed” (p. 129).

Under this interpretation, a cognitive architecture includes different cognitive systems and tries to make sense of their relations and interdependencies. The purpose is to provide a global account of cognition and not a specific model for one particular system. This kind of global architecture has also been developed in the context of Artificial Intelligence (AI).⁵⁰ Their purpose is similar: to encompass different approaches in the diverse subfields of AI. When referring to this class of architecture, I will be using the term “macro-architecture”.

But there is another use for the term “cognitive architecture” which is more specific and deals with a different issue: that of the nature of the algorithms used by specific cognitive processes when these are understood as computational. Regarding this matter, two main types of cognitive architectures have been proposed: classical and connectionist. I will refer to these as “micro-architectures”. In Chapter Three, I will address the merits of a hybrid micro-architectural theory, especially its capacity to support dual-process views.

IV. Samuels’ Proposal

Samuels (2009b) distinguishes two general frameworks concerning the relationship between the dual-process theories proposed thus far: we can think of each dual-process view as an independent scientific hypothesis about a particular cognitive faculty, or we can consider the

⁵⁰ Examples of these are production systems (e.g., Newell’s GPS, SOAR) and representational systems (e.g., Rumelhart and Ortony’s Schemata, Minsky’s Frames).

different theories as particular instances of a more general dual account of cognition. The second option – the only one that merits consideration in Samuels' opinion – claims that there is some generic dual-system view that can accommodate all (or at least most of) the individual experimental results. This view, according to Samuels (2009b), is committed to the following two theses (p. 132):

- Dual-Cluster Thesis: Cognitive processes tend to exhibit either the S1 or the S2 property clusters.⁵¹
- Dual-System Thesis: there is a division in our cognitive architecture – a division between cognitive systems – that explains this clustering effect.⁵²

Though the two claims can be independently endorsed, according to Samuels “it is the endorsement of the latter [dual-system thesis] that gives dual process theory much of its import and substance” (p. 132). I concur with Samuels to the extent that the dual-system thesis – along with claims about the macro-architecture of the mind – has given impetus to the dual-process and dual-system literature. But I also think that conflating the two claims has caused much of the confusion in the literature.

Further, I do not think this is the only acceptable conclusion to be drawn from the Dual-Cluster Thesis. We need to carefully assess the link between processes and systems and the dependence of the dual-system thesis on the dual-cluster thesis. It seems to me that the Dual-

51 The clusters are: associative, heuristic, parallel, automatic, unconscious, low demand, relatively fast, contextualized, evolutionary old, and conserved across species (S1), and rule-based, analytic, serial, controlled, conscious, high demand, relatively slow, decontextualized, evolutionary new, and unique to humans (S2). (Samuels, 2009b, p. 131)

52 “The existence of cognitive processes that exhibit distinct property clusters enables us to identify a bipartite division in the natural kinds that underwrite cognition”, where the “natural kinds” being divided are types of cognitive systems. (Samuels, 2009b, p. 131)

Cluster Thesis could just as well lead to a similar view that explains the clustering effect by a division in micro-architecture, that is, a division in the computational principles implemented by each process. Consequently, there are three independent theses that need to be proven to support a dual-system view:

- Dual-Cluster Thesis: cognitive processes tend to exhibit either type-1 or type-2 property clusters.
- Dual-Process Thesis: there is a division in our cognitive micro-architecture – a division between processes – that explains this clustering effect.
- Dual-System Thesis: there is a division in our cognitive macro-architecture – a division between cognitive systems – that explains the division between processes.

Even though Samuels tends to use the terms ‘processes’ and ‘systems’ interchangeably,⁵³ one should be precise in their usage. Much of the general confusion concerning the dual-system view is due to the loose and equivocal use of the terms ‘process’ and ‘system’.⁵⁴ For expository reasons I will address these problems in Chapter Four, after I have assessed the independent merits of the Dual-Process and Dual-Cluster theses.

⁵³ “Positing two processes or systems with broadly similar characteristics.” (Samuels, 2009b, p. 129)

⁵⁴ Use of the term “mechanism”, which Samuels seems to equate to “system”, is equally imprecise: “Thus according to dual-process theorists, not only do cognitive processes exhibit disjoint clusters of properties, but this putative fact also provides reason to posit a bipartite division between cognitive mechanisms.” (Samuels, 2009b, p. 131)

V. A type or a token view?

According to Samuels, there is a further distinction to be made—that between a *type* and a *token* version of the generic dual-process view.⁵⁵ The token version claims that there are, in total, two *particular* (token) cognitive systems in the mind. The type version claims that there are two *kinds* (types) of cognitive systems. Hence there can be more than one particular S1 or S2 in the mind. The contrast is between having only two systems per mind, or having many systems, some of which are S1 and some of which are S2.

It is worth noting that Samuels is defending a *natural kind* view: according to this view, there are two types of natural cognitive processes. In his paper “Delusion as a Natural Kind” Samuels claims that “a central aspect of any natural kind is that its members share many non-accidentally related—though logically unconnected—scientifically important properties (or relations).” (Samuels, 2009a, p. 58) According to this view,⁵⁶ natural kinds are determined by a cluster of properties that co-vary reliably (though not invariably) and which possess three characteristics: discreteness, homogeneity, and mind-independence:

1. Discreteness. Given that natural kinds are *kinds*, they must be reasonably discrete classes of entities that can be demarcated from other phenomena.
2. Homogeneity. A second characteristic of natural kinds is that they need to be fairly homogeneous.

⁵⁵ Even though Samuels aims to analyze the dual-system view, his claims are mainly based on the differences between the two types of processes, not systems. For this reason, I will concentrate on the dual-process view, rather than the dual-system view. In Chapter Four, I will address the issue of whether a dual-cluster and dual-process view is sufficient to support a dual-system view.

⁵⁶ A view that relies on Boyd’s (1991) account of contingently clustering families of properties.

3. Mind-independence. At the very heart of the natural kind concept is the idea that some kinds are real kinds in nature, where ‘real’ does not merely mean that they exist, but that they are not ideal –i.e., that their existence is in some appropriate sense mind independent. (ibid., pp. 53-54)

VI. The token theory

Samuels (2009b) lists several factors that undermine the plausibility of the token theory. First, if we take the token version to claim that there are *exactly* two particular systems, the thesis is implausible given the complexity of human cognition:

Cognitive scientists quite generally, and dual-process theorists in particular, routinely advocate a strategy of functional decomposition which conceives of relatively complex systems as hierarchically decomposable into relatively simple component sub-systems. On any remotely plausible version of this story, there are going to be *far more* than two systems. (Samuels, 2009b, p. 133)

Alternatively, the thesis could be taken to claim that, at some level of analysis, there are two particular systems. Again, Samuels (2009b) contends that even at any particular level, there are “likely to be a great many systems for a wide range of different mental processes, including perception, memory, reasoning, emotion, language and no doubt many others” (p. 135).

Samuels then considers the plausibility of there being just two systems of reasoning, given that many researchers defending dual-system views have concentrated on this faculty. Samuels' response in this case is against the possibility of isolating *reasoning* from other cognitive capacities.

He claims that it would be quite difficult to identify two particular systems of reasoning – two systems that are uniquely and completely in charge of reasoning – given the reliance of reasoning on memory, learning, and even perceptual processes.

Lastly, it is possible to understand the dual-system view as a *partially token* theory: under this interpretation one of the systems is unique while the other is a collection of systems. This interpretation was in fact defended by Stanovich (1999, 2004) and Evans (2006). Stanovich (1999) originally distinguished between a set of autonomous systems (System 1 or TASS) and a general domain thinking capacity (System 2). Evans (2006) distinguished between an analytic reasoning system and several heuristic input systems. Finally, Fodor's (1983) distinction between modules and a general thinking capacity can be interpreted as analogous to the partially token dual-system view. Fodor claims that there is only one token of S2 (the general thinking capacity) and several tokens of S1 (the modules).

According to Samuels, there is a powerful reason to oppose the partially token view, namely that the type of processes that S2 is in charge of show the kind of heterogeneity that gave us reason not to postulate a unique pair of systems. It should be observed that many of the researchers who first defended Fodor's view have now moved away from it. According to Samuels (2009b), type-2 processes “differ in precisely those respects that should lead us to conclude that they involve different cognitive systems. Specifically, they differ in ways that are typically taken as relevant for the individuation of cognitive mechanisms” (pp. 136-137). That is, they differ in their functional and computational properties. He then concludes that a token version of the dual-system thesis is a highly implausible one.

I agree with Samuels' analysis of the token version of the theory. It is not possible to defend a token dual-system view since a system, however liberally construed, cannot account for the complexity of all higher cognitive processes (let alone all cognitive processes). In fact, I would argue that Samuels should have emphasized this point more strongly. Most of the problems faced by dual-system views stem from the lack of a consistent and robust interpretation of the notion of 'system' and the ensuing inconsistency in its application to different mental phenomena.

VII. The type theory

Samuels dismisses the dual-system *token* view on the basis of the implausibility of there being only two systems (under any plausible interpretation of the term 'system'). Hence the only interesting formulation of the dual-system view is the type version, the one that posits two types of cognitive systems.⁵⁷ Samuels believes that this is a view with many merits, but not without problems. He points out the following issues.

The Specification Problem: there are many different characterizations of the property clusters associated with each system, but clearly not all of them can be correct.

In order to address this problem, there are two steps that need to be taken. First, refine the properties that are part of each cluster (for example, there are many different notions of control

⁵⁷ Later he mentions that the type view is "the claim that our minds are composed of two different kinds of cognitive mechanisms" (Samuels, 2009b, p. 137). As I will show later, this lax use of the words systems, mechanisms, processes is at the basis of most of the confusion and the problems faced by dual-system and dual-process literature.

and automaticity⁵⁸). Second, one should specify which properties should remain in the list and which should be eliminated. For example, the distinction in terms of evolutionary history is very controversial since some of the processes that belong to System 1 are not old on an evolutionary time scale (and similar issues arise with the distinction between serial and parallel processes). Researchers need to address these problems or remove these properties from the final list of defining characteristics.

Samuels believes that neither issue (named the “conceptual” and the “inclusion” problems) represents a major obstacle for the type theory. He claims that the lack of clarity about property clusters arises “because it is quite possible for distinct, individual mechanisms to differ in *some* respects even if they are members of the same broad class” (ibid., p. 136). Further, he believes that this is simply evidence of a theory in development and that the kind of refinement needed is common for scientific theories that are in early stages of development.

I believe, however, that Samuels misses the central problem of the type thesis. Granted, all scientific theories need revision and the process of consolidating a hypothesis is a long and arduous one, especially when the hypothesis is general and concerns many different, complex phenomena. However, at the core of the type thesis is the claim that there are two types of cognitive processes whose differentiation is in terms of a list of clustering properties. If the two lists cannot be specified because of the absence of two sufficiently coherent clusters of properties, then what is left of the theory? Of course, if one could identify the underlying mechanism that gives rise to the clustering, then one would have reason to adopt the dual-system view. The problem is that the direction of

⁵⁸ I will discuss this issue in more detail below.

inference is the opposite: we claim that there are two systems because we infer that there have to be two (distinct) underlying structures responsible for the two clusters of properties reported in the literature.

Samuels defends this claim by making an analogy between a disease (e.g., influenza) and its symptoms (high temperature, cough, fever, etc.) and a system (e.g., S1) and its properties (fast, automatic, unconscious, etc.). While the presence of the disease is generally correlated with the presence of the symptoms, the connection is not always valid in both directions. One can have the disease and not all of the symptoms, and one can have most or all of the symptoms and not have the disease. Similarly, a cognitive process can be type-1 without manifesting all the characteristic properties, or it can have several of the characteristics of the first cluster without being type-1.

My contention is that this is not an appropriate analogy. In the case of a disease there is in fact an underlying natural kind that supports the individuation and that explains why not all the symptoms are always present. In the case of the dual-system view, the issue at stake is whether we can posit such underlying natural kinds. As of now, the main reason for doing so is the existence of processes possessing two distinct clusters of properties. Of course, if there were in fact two underlying mechanisms or structures giving rise to these clusters, then the vagueness of the two clusters would not constitute a problem (since there would be in fact two underlying natural kinds). We would not require the cluster of properties to explain the difference between the two systems, as we would have another kind of explanation available. As in the case of influenza, what explains the difference between that disease and another is not the presence of the symptoms but the presence of a virus (with variable symptoms).

Of course, someone might argue that preceding our knowledge of viruses and bacteria (the underlying causes), there was the identification of a cluster of symptoms, and that it was on the basis of that clustering that we posited the existence of the underlying disease. My response is two-fold: first, a considerable problem of the dual-system view is that the clusters themselves are not as consistent as it was once thought (a crucial problem with the view which I will address in more detail in the next section). Following the analogy, if a set of symptoms appeared sometimes wholly, sometimes partly, sometimes only one at a time, it would be very difficult to establish how many distinct diseases are responsible for the set of phenomena. Secondly, in the case of diseases, we already have a natural kind (*disease*) for which we have developed criteria for its specification (e.g., we would not consider the following symptoms as indicating the presence of a disease: waking up an hour earlier for work, dressing in bright colors, and starting to drink coffee with lots of milk). The more we know about the natural kind, the easier it is to specify particular types of it.⁵⁹

Unfortunately, little is known about cognitive processes and even less about how to distinguish different kinds. We lack knowledge about the kind of properties that would distinguish among cognitive processes. For example, is *speed* one of them? What about *content*? At this point, given the general conceptual confusion involving notions like *system*, *process*, *structure*, and *mechanism*, it is quite difficult to determine which properties would be helpful in establishing differences among them.

59 If we think back to the time when there was no knowledge of diseases, then it would have made sense to doubt whether the symptoms of a cold were due to an underlying disease or something different. In fact, this is the problem that we currently face with mental diseases, it is not clear what set of symptoms allow/support the inference to an underlying disease. Before a set of phenomena can be classified as a sub-kind, we need to identify what are good signs of divisions in the larger kind (disease). In the case of cognitive processes, it is difficult to tell which properties can discriminate between types.

The foregoing should not be taken to mean that I do not agree with the general strategy of inferring underlying natural kinds from the clustering of properties. But I do think that the clustering needs to be robust: there has to be a sufficient number of cases that present the same set of properties consistently. There cannot be a continuum of cases that present many variations of such properties. As Samuels (2009a) points out, “given that natural kinds are kinds, they must be reasonably discrete classes of entities that can be demarcated from other phenomena” (p. 52) and “they need to be fairly homogeneous kinds” (p. 53).

Ideally one would need to provide an explanation for the clustering. And this seems to be the most important factor.⁶⁰ But, if such an explanation is not available, then a robust phenomenon of clustering is critical. If not, what is left to support the supposition of an underlying kind?

In sum, there is a general tendency to move back and forth between (i) assuming that there are two underlying kinds that manifest different clusters of properties and (ii) inferring that there are two underlying kinds on the basis of the presence of the two clusters of properties. This confusion becomes apparent when Samuels addresses the next problem faced by the type theory.

The Crossover Problem: it seems that there are several cognitive processes (e.g., language) that exhibit characteristics from both clusters of properties. If this is correct for a considerable number of processes, then one cannot infer two cognitive kinds, since properties for processes do not cluster in the clear-cut way proposed by the type view.

⁶⁰ After all, we are willing to accept any set of symptoms as signs of a new disease as long as an explanation of the mechanism that gives rise to the cluster of symptoms is available.

Samuels believes that crossovers do not represent a real problem as long as they are not “too numerous or too extreme”. Further, he claims that this problem could be solved by either dropping some of the notions that create crossovers (for example, evolutionary history) or by tolerating some crossovers given that the presence of associated properties does not necessitate the presence of one of the two types of processes. After all, the claim is that the type view concerns some cognitive natural kinds, not all:

The presence and activity of some system 1 (or system 2) does not require that all members of its associated property cluster obtain, any more than, say, the presence of all typical symptoms of influenza is a necessary condition for having flu. The point is that, for natural kinds quite generally, the relationship between underlying mechanism and associated characteristics is far weaker than this. (Samuels, 2009b, p. 138)

But if this is the case, then why is it important that crossover cases are not too extreme or numerous? Let’s consider, for example, the case of two diseases and their symptoms. A set of symptoms A is indicative of disease α and a set of symptoms B is indicative of disease β regardless of whether a subset of symptoms A plus a subset of symptoms B is indicative of yet another disease (δ). Though natural kinds need to be associated to fairly discrete and homogeneous clusters of properties, these clusters need not be completely dissociable from one another. Diseases, for example, tend to have many overlapping symptoms.⁶¹ As in the case of diseases, why not think that there are many cognitive kinds, two of which are described by the dual-process views?⁶² Of course,

61 Just like kinds of animals tend to share properties, as well as fruits and many other natural kinds.

62 The dual-view aims to explain some of the natural cognitive kinds. Why, then, worry about the cases that are left unexplained? Would it be correct to just assume that these describe other cognitive kinds?

we would need to review the “dual” label, but this would be a minor problem compared to the problem of establishing whether there are such two types in the first place.

Crossovers are quite problematic—especially when numerous—because the main justification for positing two types of processes is the uniform clustering of properties. Without such clustering, there is no reason to believe that there are two (or any number of) cognitive types (or natural kinds).

The only way that the numerous dual-process views can support a dual-system view is if the set of properties attributed to each process is (more or less) the same in all the views. That is, there has to be a certain *clustering of properties*. Such clusters can then be explained by two different underlying mechanisms, S1 and S2. If each dual-process view juxtaposes processes with different sets of properties, then there is no clustering and, hence, no underlying cause can be hypothesized. In that case, researchers would have to find another reason for justifying the division of cognitive processes into two *types*.

Samuels recognizes the importance of this issue when stating the third problem faced by the type view.

The Unity Problem: there is no good explanation for the clustering effect. That is, it is unknown why the two types of processes exhibit those particular clusters of properties. Further, assuming that the clustering effect is due to different underlying mechanisms is not sufficient, but it rather begs the question of why such mechanisms give rise to different property clusters.⁶³

63 “Though positing mechanisms is a standard strategy for explaining the existence of property clusters, it does not, by itself, constitute a satisfactory explanation. Rather one needs to specify those features of the proposed mechanisms that account for

According to Samuels, a condition for the existence of a natural kind is that it be associated with some process or mechanism that explains the clustering of properties: “much science seeks to explain the existence of reliably co-varying property clusters by identifying and specifying the structures, processes or mechanisms that –under appropriate circumstances— causally explain the contingent clusters associated with (natural) kinds” (ibid., p. 52). The lack of such explanation constitutes an important challenge to the view, one that has not been overcome despite several attempts (Sloman, 1996; Carruthers, 2006; Evans, 2009).

As I mentioned in section IV, the dual-cluster view supports the dual-process and dual-system views, not the other way around. We infer that there are two types of processes because we see a strong clustering effect, and we explain the duality of processes on the basis of two different underlying mechanisms.

So the dual-system view faces two challenges: (i) to show that there is a robust clustering effect, (ii) to demonstrate that the duality of processes is due to a duality of mechanisms.

In contrast, the dual-process view only needs to show that there are two types of processes, without regard for the difference between the two processes and their different underlying mechanisms. Indeed, some dual-process theories are compatible with a single mechanism employing two different processing modes.⁶⁴ However, a dual-process *type* theory is committed to

such clustering effects. In the present case, we need to specify those characteristics of type-1 systems that yield S1-exhibiting processes, and those properties of type-2 systems that yield S2-exhibiting processes.” (Samuels, 2009b, p. 138)

⁶⁴ Again, all this depends on what we take a mechanism to be. If different processes determine different mechanisms, then this option is not available. But since Samuels and other researchers take mechanism to be equivalent to system, rather than process, I am following that convention here.

the claim that all dual-process views are somehow a consequence of the same phenomenon. Hence, it needs to provide an explanation for the underlying common cause.

In what follows, I will address this last issue: the clustering effect and possible reasons for it.

VIII. The Clustering of Properties

There are mainly two strategies for positing the clustering of properties:

1. Single-property-based clustering. On this strategy, there is one particular property that produces the clustering. For example, in virtue of being fast, a process is also automatic, unconscious, associative, low-effort and high-capacity, etc. (and the opposite for being slow). This is the type of strategy pursued by Kahneman (2011) and Evans (2010).⁶⁵ I will discuss their virtues and shortcomings next.
2. Common-cause-based clustering. On this strategy, most of the properties cluster together because of some common underlying reason or principle that produces the clustering. A process is fast, automatic, and unconscious because of the way in which the process is performed. The same holds for slow, deliberate, conscious processes. As mentioned above, this option distinguishes processes by their computational principles (i.e., micro-architecture).

According to Samuels (2009b), this is a good option for distinguishing the two processes and providing a solution to the unity problem stated above: “one general approach to explaining the type-1/type-2 distinction would be to assume the cognitive mechanisms that underlie them operate

⁶⁵ I focus on these two authors because they hold two of the most well-known, influential views, but this kind of problem is ubiquitous in the literature.

according to different computational principles” (p. 139). However, he later claims that “there is currently no reason to suppose that the type-1/type-2 distinction corresponds to the connectionist/classical divide” and that “it’s very unclear how the classical/connectionist distinction is supposed to explain many of the properties associated with type-1 and type-2 processes” (p. 140).

In Chapter Three I will show how a distinction at the level of micro-architecture can provide an answer to the clustering effect. Later, in Chapter Four, I will show why this distinction only supports a dual-process but not a dual-system view. Before offering this solution, I would like to explain why I am not convinced by other explanations of the dual-system view.

One property based clustering – Kahneman’s view

Let’s consider Kahneman’s (2011) view. This view claims that the clustering of properties is caused by (or at least strongly correlated with) the speed of the process. There are several problems with this view. First, slow and fast processes do not group very neatly. For once, whether a process is fast will depend on context. Deciding which brand of car to buy can be considered a ‘slow’ process if we take into account all deliberations that occur over days, or a ‘fast’ process if we single out the moment in which the person suddenly decides to get the Toyota instead of the Honda. Also, compared to the time that it takes to classify a common object (e.g., a chair) as such, both of the previous processes are slow.

Further, as Kahneman acknowledges, “fast thinking” can be due both to expertise or to heuristic-based reasoning. But these are very different types of processes. As a matter of fact, the same process can be considered type-1 or type-2 depending on the person and the context (notice

that ‘Find a strong move in chess’ and ‘Check the validity of a complex logical argument’ could be the other type of process for a logician who doesn’t play chess well). To this objection someone could respond that the point is to distinguish between cognitively “difficult” or “effortful” processes and “easy” ones. But this type of approach does not produce a very strong view. If the only thing that distinguishes fast processes from slow ones is their speed, how valuable is that distinction? It certainly does not explain nor predict much about those processes. On the other hand, if there is a reason why fast, automatic processes are also effortless and unconscious – and vice-versa for slow, deliberate, effortful, and conscious – then that reason needs to be explained. It is that reason itself that will provide explanatory and predictive power to the view, not the fact that we can divide processes into two different kinds (we can always do that). So, if there are in fact two cognitive types – type-1 and type-2 – then what is the common cause that groups type-1 into one ‘kind’ of process? And what can we infer from this once we know what the common cause is?

Finally, by grouping together very dissimilar types of processes we might erase important differences. For example, ‘fast thinking’ can be due to simplicity, innateness, or the acquisition of skill. Kahneman groups together under type-1 processes all these cases: ‘Understand simple sentences’, ‘Detect that one object is more distant than another’, ‘Find a strong move in chess’ (if you are a chess master).⁶⁶ The process of recognizing a chess position has become automatic (by a transformation that must be explained), while detecting distances is always automatic under normal conditions. Indeed, some processes are fast because they lack complexity, some are fast because they are performed during earlier stages of processing (hence the results are already available,

⁶⁶ See Table 1 in Chapter One.

which would be the case with natural assessments), still others have become fast with practice. Further, the process of automatization might be due to a change in the modality of thinking (from rule-based to associative, for example) or to an increase in the speed of processing.⁶⁷ Finally, certain conditions (for example, tiredness, lack of time, lack of interest) make a process automatic *at times* (mainly because we are not engaging our full attention on that particular occasion). Hence, the variability of reasons why a process is automatic make it very difficult to infer what other properties the process will have.⁶⁸ A comprehensive theory of cognition needs to account for these differences (or explain why they are not relevant). Kahneman does not provide either explanation.

The relationship between speed and other features, like consciousness and control, is also complicated. Sometimes processes will be fast and unaware because of the simplicity of the process: for example, certain associations (salt → pepper) are fast and unaware because (to the best of our knowledge) there are no intermediate steps between input and output, there is just a strong association between them. Of course, this might be a simplified picture of what occurs in the brain. But unlike the case in which processes become associative-like, for this type of association, it seems that the most plausible explanation is that there is a strong (positive) connection between the two representations. Given current knowledge of the workings of the mind, the simplest hypothesis is to assume a strong association between the two items rather than hypothesizing a complex processes involving several steps linking stimulus to response.⁶⁹

67 More importantly, certain processes are automatic and effortless because they have become so over hundreds of trials. But other processes are automatic because they cannot be otherwise.

68 “This is the essence of intuitive heuristics: when faced with a difficult question, we often answer an easier one instead, usually without noticing the substitution.”

69 Of course, there is always the question of the level at which we model the mind. If the focus is at the cellular level, then we might need a very complex model with several steps to model such process. But at a higher, cognitive level, then a simple associative network seems sufficient.

Nevertheless, as I mentioned above, there are some processes that become fast with expertise: we learn how to read, write, add single digits, and the like in a very fast, automatic way, to the point where we are extremely fast and unaware of the intermediate steps—unlike when we first solved these tasks. In these cases, it doesn't seem like a simple association can account for the process.⁷⁰

An empirically correct theory of the mind has to explain this difference: we cannot model processes that have become fast in the same way as those that were (comparatively) fast from the start. We can certainly model the improved process as sharing properties with fast processes, but ideally we would model it in a way that reflects the transition.

An additional problem with Kahneman's (2011) view is the variation in the types of processes that are being compared. First, Kahneman talks about the difference between fast and slow thinking, later referring to it as involving intuitive versus deliberate thought.

I describe mental life by the metaphor of two agents, called System 1 and System 2, which respectively produce fast and slow thinking. I speak of the features of intuitive and deliberate thought as if they were traits and dispositions of two characters in your mind. (p. 12)

Next, Kahneman (2011) focuses on the contrast between processes requiring attention and processes that do not.

⁷⁰ Granted, this is a very complex issue: it might well be the case that for some processes that become automatic what happens is that a network of associations between inputs and outputs is formed and this causes the improved speed. But I do not think that this can account for all the cases. Rather, for some processes there is probably a phenomenon of tuning, whereby better strategies and improved speed of intermediate steps are achieved.

The highly diverse operations of System 2 have one feature in common: they require attention and are disrupted when attention is drawn away. (p. 34)

Yet later he focuses on the difference between conscious and unconscious processes. And then talks about the contrast between automatic, effortless processes and deliberate, effortful ones:

I describe System 1 as effortlessly originating impressions and feelings that are the main sources of the explicit beliefs and deliberate choices of System 2. The automatic operations of System 1 generate surprisingly complex patterns of ideas, but only the slower System 2 can construct thoughts in an orderly series of steps. (p. 21)

However, there is no reason to think that all these contrasts (one based on automaticity, another on consciousness, yet another on speed) pick out the same set of processes. There are no empirical or theoretical reasons that justify these transitions. On the contrary, there is evidence that these properties do not always cluster together. For instance, fast thinking can be rule-based and it can also require attention (e.g., when a mathematician performs a very quick calculation in her head). I will elaborate on this point later. For the moment, it is important to keep in mind that while these distinctions seem to pick out the same dichotomy, there is no evidence that this is the case. Further, even if one showed that the processes that are slow, deliberate, and requiring attention cluster together, one would still need to show that those processes are produced by a different system than those which are fast, intuitive, and do not require attention.

In sum, the basic problem with this type of view is that the distinction between the two systems is not very robust. Kahneman's (2011) view simply says that processes in one system share some common properties, which are opposite to the properties of the other system's processes. The

view divides cognitive processes into two neat categories but, given that one can always come up with groupings of brain processes based one or more common properties, the view lacks theoretical support.

Indeed, there are many other possible divisions of cognitive processes that involve different characteristics from the ones considered by Kahneman. The proposed distinction is not *prima facie* stronger than the rest. So, a list of common characteristics is not enough to propose two different underlying *systems*. Kahneman needs to show that there is something common to all S1 processes and to all S2 processes, other than their speed, controllability, and content. He needs to show that there is an underlying reason for the clustering of properties. So, as it stands, the distinction proposed is not robust enough to assign fast, automatic, associative processes to one system, and slow, deliberate, rule-based processes to another.⁷¹

These problems are not unique to Kahneman's view. Given the wealth of data that a dual-system type theory needs to account for, it is quite difficult to find a unique property that will explain the clustering. Needless to say, cognitive processes are very complex. No particular property can guarantee the appearance of any other property. Even properties that were always clustered together, like attention and consciousness, are beginning to come apart in recent empirical findings.⁷² So it would be quite a bold move to assume that one type of property reliably leads to the presence of not one but a whole cluster of associated properties.

71 They also need to show why these are two systems and not one that employs different types of processes.

72 In a detailed review, Koch and Tsuchiya (2007) show that attention and consciousness can be simultaneously and alternatively present in cognitive processes.

Property-based clustering. Evans' view

According to Evans (2009) what distinguishes the two types of systems is their relation to working memory. Systems that involve working memory inherit some of working memory's properties. For example, since working memory is limited in capacity, its use conditions processes to be slow, serial, high effort, and low capacity. Processes that do not involve working memory are assumed to be fast, parallel, low effort, and high capacity by contrast.

This is a problem for the view (also pointed out by Samuels): we can understand why processes that engage working memory are slow and serial, but it is not clear why the lack of involvement by working memory makes all other processes fast and unconscious, for example. As in the case of Kahneman's view, more needs to be said in order to support a *dual* system view. For instance, Evans needs to rule out the possibility that processes that use working memory are one type, but processes that do not use working memory are of many different types. I will address this issue in the next section.

Property-based clustering. Automaticity, awareness, and control.

It seems that any process involving a resource that is limited, such as working memory or attention, would exhibit some of the prototypical type-2 properties (slow, effortful, serial). Could this be an alternative explanation? Could we divide processes on the basis of their automaticity or their use of attentional resources?

Many of the results in the dual-process literature point to a distinction along this line, which seems to be consistent with other sets of dualities: fast vs. slow, low effort/high capacity vs. high

effort/low capacity, and, to a certain extent, implicit vs. explicit, and conscious vs. unconscious. I believe that the distinction between automaticity and control is indeed fruitful – much more than a distinction based on speed, for example – but it still faces several challenges. I will consider this issue next.

Automaticity and attention

By definition automatic processes are fast, effortless, and high capacity. In this sense, automaticity develops once a process has been rehearsed extensively and does not require executive control (in the form of attention). All kinds of cognitive processes (e.g., walking, reading, playing a sport or an instrument) change in this way: from an effortful, controlled, and slow performance to an effortless, fast, and automatic response. The repeated rehearsal (practice) leads to faster times and the sensation of low effort. This is why automatic processes are fast and low effort: they do not require attention. For example, I can type as I think because I have automatized typing, I can focus my attention in the thinking rather than the typing.⁷³

By contrast, processes that demand attention often seem effortful: attention is a limited resource and the use of this resource—especially when doing something unusual—produces the impression of effort. Or, seen from another perspective: though not all processes that require attention feel effortful, processes that feel effortful require attention. After all, how could a task feel effortful if we are not paying attention? (For example, how could we feel that it is difficult to type/read/hit a ball if we were not paying attention to the process as we perform it?)

⁷³ Similarly, when an advanced golf player makes a shot, she's not thinking of all the particular movements she has to do in order to swing and hit the ball, she can just focus on one thing (e.g., not taking the sight off the ball).

So, processes that are experienced as high-effort involve attention and those that happen while we are paying attention to something else are usually experienced as low-effort. In this way, automaticity also connects with consciousness. Automatic processes are usually unconscious because they do not require attention (i.e., we are not aware of performing them), while processes that are effortful, usually involve attention and are hence conscious.

However, the connection between automaticity and unconsciousness only holds for certain aspects of performing a task. For example, it seems that many tasks (driving, reading, playing a sport, etc.) are performed *automatically* and with absence of awareness. However, this only applies to part of the process of performing those tasks: for instance, though we are not aware of the details of how we're doing what we are doing, we are certainly aware of the process itself. For example, when I type, I might not be aware of how I do it, but I am certainly aware that I am doing it, and I might even be aware of certain movements needed to perform the task.⁷⁴

Someone might object that we are always aware of the overall process, that the question is about the particular sub-processes involved. But this is not always the case, there are times when people completely lose awareness of the action they are performing (this is how some accidents or errors occur): you put sugar in your cup without realizing it, you try to open your office door with your home keys, you walk to the store without “realizing” because you were thinking of something else. It seems that attention and awareness, even for actions, come in different degrees.

⁷⁴ A very simple example of this contrast involves handwriting. When we first learn the rules of handwriting (e.g., dotting the *i* and crossing the *t*), we need to pay a lot of attention to write a sentence neatly. Over time, the handwriting becomes crooked, but also extremely fast and automatic. But if one tries to go back to handwriting neatly (and following the rules we once learnt), the process of writing suddenly becomes very ‘aware’ and difficult (e.g., *A ten year-old is better at this than I am*). In all these cases we are aware that we are writing, what changes is the level of awareness of the movements needed to perform the task. The more neatly we write, the more attention we use, hence the higher the awareness of the process.

Automaticity is also a question of degree. As I mentioned, most of the time, even if we have automatized a process (like typing, reading or walking), no matter how easily we do this, we are aware of performing the action. By contrast, there are some processes of which we have no awareness, ever. This is a different type of automaticity and it is related to the stage in which the processing occurs: “early” processes are usually carried out automatically. The same holds for certain processes that involve particular functions, like vital processes. For these processes, there is no awareness of the process itself.

Hence, it is important to distinguish two types of automatic processes: those done without awareness of the mechanisms that lead to the performance of the process and those without awareness of the process itself (e.g., processes which occur completely under the level of awareness). The latter includes processes like perception, categorization, and, most importantly for our discussion, some kinds of “intuition”.

The contrast is important because this last type of process, especially those involved in reasoning and decision-making, give support to the view that the mind is not unified, but the result of the interaction of opposing processes, some of which the “conscious self” is not aware of. Again, just like we are not aware (at any level) that our brain is controlling our heart beat or categorizing a chair as such, there are many automatic processes that contribute to reasoning and decision making of which we are completely unaware, and this lack of awareness is not the same as the lack of awareness regarding other processes, like hitting a baseball. I might not know exactly how I need to move my arms to swing – and it would probably be worse if I tried to be aware of it

as I do it – but I still know that I am swinging when I do it. I am also paying a lot of attention to this process (unlike when I perceive or categorize things).

Control

Just as automaticity does not imply lack of awareness, awareness does not imply control. Indeed, the relationship between awareness and control is not straightforward. Under one interpretation, processes that require attention and executive control are those that lead to awareness. Since they usually involve working memory, they are also low capacity and slow. Finally, the story goes, we can exert control over such processes by focusing our attention as the process uses working memory and other central resources. However, this view has yet to be proven correct. Many authors, including renowned neuropsychologist Elkhonon Goldberg, claim that conscious control is just an illusion. All processes are under the control of central executive functions and we only create the feeling of intention *post facto*.⁷⁵ Under this second view, awareness and the feeling of control is a result of how the mind constructs the flow of consciousness: we do not focus our attention in a particular process but executive processes distribute attention according to an underlying (so far unknown) mechanism and later construct a coherent view for our “inner eye”.⁷⁶ In a way, Evans (2010) seems to have accepted this as the reality of the mind:

75 As Posner and Fernandez-Duque (1999) put it, “One approach to the question of voluntary control is to argue that it is an illusion that arises out of the competitive activation of a large number of brain systems. What appears to be volition is the result of a complex network relaxing to a particular state. Although without denying the top-down component, this view stresses the bottom-up processes. A different view elaborated in this section is that there is a high-level executive attention network with its own anatomy that works to resolve competition and in the process gives rise to subjective feelings of cognitive control (Norman & Shallice, 1986). This view emphasizes the top-down control.” (p. 44)

76 See Koch and Tsuchiya (2007).

This concept of conscious willing is deeply ingrained in our folk psychology. But things are not as they seem. The experience of willing is a consequence of our intentional actions and not a cause. The contrary view is a compelling but false belief created by our brains. (p. 169)

Of course, we can allow some exceptions to the dualities proposed. So, for example, if automaticity/control allowed us to explain most of the clustering, a dual-system type view based on this contrast would prove fruitful. However, there are many other dualities in the literature that are not explained by differences in terms of automaticity and control—e.g., implicit versus explicit, parallel versus serial, associative versus rule-based.

In the next chapter, I will show that a distinction in terms of different underlying cognitive architectures (i.e., connectionist versus classical) can explain (to a certain degree) the clustering of properties reported in the literature. In particular, it can explain why some processes that are associative are also fast, automatic, effortless, unconscious, and implicit. The same explanation can be provided for processes that are rule-based, slow, deliberate, effortful, conscious, and explicit. Indeed, one should expect certain properties to cluster in the case of associative processes and the opposite for classical processes. If correct, this proposal would let us draw a substantial distinction between the two systems. However, in Chapter Four, I will argue that, though robust, this distinction cannot support a general dual-system view, that is, a macro-architectural *duality of the mind*.

Chapter Three

I. Introduction

The claim that cognition is a kind of computation is one of the most debated issues in philosophy of mind:⁷⁷

No idea in the history of studying the mind has fueled such enthusiasm – or such criticism – as the idea that the mind is a computer. This idea, known as "cognitivism" (Haugeland 1981/1998) or the Computational Theory of Mind, claims not merely that our minds, like the weather, can be modeled on a computer, but more strongly that, at an appropriate level of abstraction, we *are* computers. (Cantwell Smith, 1999, p. 153)

Despite its controversial nature, *Computationalism*⁷⁸ remains a central assumption of cognitive science,⁷⁹ one that is shared by dual-process and dual-system theorists alike. In a similar vein, many

77 For discussions on this topic see Dreyfus (1992), Fodor (1983), Haugeland (1978), Searle (1992), Van Gelder (1995).

78 "Computationalism is the view that intelligent behavior is causally explained by computations performed by the agent's cognitive system (or brain). In roughly equivalent terms, computationalism says that cognition is computation. Computationalism has been mainstream in philosophy of mind – as well as psychology and neuroscience – for several decades." (Piccinini, 2009, p. 515)

79 "The central hypothesis of cognitive science is that thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures. While there is much disagreement about the nature of the representations and computations that constitute thinking, the central hypothesis is general enough to encompass the current range of thinking in cognitive science, including connectionist theories which model thinking using artificial neural networks" (Thagard, 2012, Representation and Computation section).

cognitive scientists assume that the mind is some sort of complex mechanism,⁸⁰ which employs different computational processes.⁸¹

Working under these assumptions, cognitive scientists aim to explain cognition by providing models of its underlying computational processes. These models try to specify in as much detail as possible the processes by which inputs are transformed into outputs (the computations).⁸² However, this common methodology leads to very diverse results. First, computational models range from the very specific (models of the workings of a group of cells) to the most general (models of the brain). Depending on the level of specificity, the details of the model tend to matter more or less.

Second, there are no fix assumptions about the nature of the computational processes employed by the brain. Indeed, the question of how the brain computes is still unsettled. The following quote from Ken Aizawa highlights this issue:

On the hypothesis that cognition is computation, one research goal is the determination of the computational architecture that constitutes the structure and organization of cognition. In other words, if the brain is a type of computing device, one wishes to know exactly what type of computing device. Even on the hypothesis that cognition is a form of effective computation (Turing-equivalent computation), one still wishes to know exactly which of the computationally equivalent forms of computation constitutes the hypothetical computational architecture of cognition.

80 "Though cognitive scientists disagree on many issues, one widespread commitment is that the mind is a mechanism of some sort: roughly, a physical device decomposable into functionally specifiable subparts." (Samuels, 2008, p. 581).

81 See Bechtel and Abrahamsen, 2005.

82 "A cognitive model is intended to be an explanation of how some aspect of cognition is accomplished by a set of primitive computational processes." (Lewis, 1999, p. 526)

Even if diverse forms of computation are equivalent in the sense of computing the same functions, one wishes to know exactly what mechanisms the brain uses to compute those functions. (Aizawa, 2004, para. 2)

Regarding this concern, two main types of cognitive architecture have been proposed: classical and connectionist computation.⁸³ Classical architectures were the first to appear and they were based on the architecture of digital computers. Connectionism⁸⁴ arose in the seventies as an alternative to classicism, when it became clear to many cognitive scientists that it would be extremely difficult to model certain cognitive processes (like pattern recognition) on a classical architecture. As a consequence, there is an ongoing debate about which one of the two architectures can model cognition more accurately.⁸⁵

As I mentioned in Chapter Two, a distinction is in order regarding theories of cognitive architecture. On the one hand, one can understand a cognitive architecture as a global theory of cognition that integrates the models of several cognitive processes. In this case, a model describes the workings of one particular process and a cognitive architecture provides a general, unified explanation of the workings of all the processes. It also provides a way of connecting the different models and showing their underlying similarities. I refer to these theories as models of “macro-architecture”. The other use of the term “cognitive architecture” is more specific and refers to the

83 Some researchers consider dynamical systems as a third alternative.

84 As Dawson and Shamanski (1994) point out, there are many different connectionist architectures. Like them, I will also concentrate on generic connectionism, whose main characteristics have been described in the ‘bible’ of connectionism, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, by Rumelhart, McClelland, and the PDP Research Group (1986).

85 See Bechtel (1987); Chalmers (1993); Churchland and Sejnowski (1990); Dawson and Shamanski (1994); Fodor and McLaughlin (1990); Fodor and Pylyshyn (1988); Smolensky (1988).

computational principles used to model specific cognitive processes. When discussing the merits of these two types of architecture, I will refer to them as kinds of “micro-architecture”.

The main purpose of this and next chapter is to show that there is an underlying confusion between micro and macro-architectural features of the supposed two system types such that for some authors the distinction is at the macro-architectural level while for others is at the micro-architectural level. I will show also that it is not possible to match these two distinctions and that this is the underlying reason why it has been impossible to come up with a reasonably unified dual-system view.⁸⁶ In this chapter, I will address the distinction at the micro-architectural level and show that a dual or hybrid cognitive architecture can better explain many of the results in the literature, particularly the distinction between implicit and explicit processes, the presence or lack of conscious awareness, their automaticity and speed.

II. Marr’s Three Levels of Analysis

When modeling a cognitive process, it is often helpful to distinguish between different levels of analysis. This helps us better determine which of the process’ characteristics we want the model to account for. A minimal model just specifies the input and the output, and characterizes (at least nominally) the transformations that occur between the two. In such a minimal model, it is not

86 The distinction is clear in the way that different surveys of the literature present the relation between dual-process and dual-system hypotheses. Evans (2008) claims that “a number of authors have recently suggested that there may be two architecturally (and evolutionarily) distinct cognitive systems underlying these dual-process accounts” (p. 255). Frankish and Evans (2009) state that “such theories claim that human central cognition is composed of two multi-purpose reasoning systems” (p. 1). And Frankish (2010) says that “dual-process theories hold that there are two distinct processing modes available for many cognitive tasks... Dual-system theories go further and assign these two types of processing to two separate cognitive/reasoning systems, System 1 and System 2 – a view sometimes described as the two minds hypothesis” (p. 914). In one quote the focus is on the structure and mechanisms of central cognition while in the other is about the internal architecture of a specific cognitive process.

important, for instance, how the input and the output are physically implemented. The specific steps of the transformation may also not be relevant. For example, in the case of perception, a neuroscientist might be especially interested in the transduction process (the way in which the visual input is transformed into neural activity). A cognitive psychologist, on the other hand, might be interested in modeling the different transformations of the input along its path in the brain, but she might be indifferent as to how this happens at the anatomic level.

In his book *Vision*, David Marr proposed a distinction among three levels of analysis when studying cognitive processes (again, under the assumption that cognition involves computation): the *computational*, the *algorithmic*, and the *implementational* levels. The first level concerns the function the cognitive process is computing. A theory at this level has the purpose of responding to the question “what is the goal of the computation and why is it appropriate?” The second level concerns the algorithm that is used to compute such function. An analysis at this level answers the question “what kinds of representations are used for input and output and what algorithm transforms one into the other?” Finally, the third level concerns the physical implementation of the algorithm. At this level theorists are seeking to answer the question “how are the representations and algorithms implemented physically?”

For Marr, the distinction provides a general framework for studying all kinds of cognitive processes. Further, the study of each level is fairly independent of the other two. For example, one can study the algorithmic or computational level without taking into account the implementational level.

The debate between symbolic and connectionist architectures is a debate about the kind of *algorithms* that the brain uses to process information. It is not a dispute about what kind of problems or functions the brain computes. The debate takes for granted which problems are being solved. It tries to answer the question of *how* these problems are being solved: by connections in a neural network or by the structured processing of symbols with language like structure? The debate assumes that the difference between the two architectures is at the algorithmic level. The debate is also not about the implementational level since this level concerns how to physically implement the algorithms that are being used to compute the function. In this respect, many researchers from both sides agree that this level is ‘connectionist’ in the sense that there is a huge network of physical connections taking place.⁸⁷

III. The contrast between two types of architecture

As we noted, the notion of computation is itself a source of controversy. Cantwell Smith (1999) remarks:

Within cognitive science, however, cognitivism is usually understood in a more specific, theory laden way: as building in one or more substantive claims about the nature of computing. Of these, the most influential (...) has been the claim that computers are formal symbol manipulators. (p. 153)

⁸⁷ Even Fodor and Pylyshyn (1988), stark defenders of classical architectures, concede that at the implementational level the mind is clearly a large network of connections. But they also claim that this is of no importance for a *cognitive* theory of the mind, i.e., a theory of the algorithmic and not the implementational level.

So, while *computationalism* is the general view that cognition is a kind of computation, the philosophical debate (usually) centers on a specific version of computationalism, that which claims that cognition is classical digital computation.

Classical Computationalism

The so called *classical* view is the product of a long series of developments that took place at the beginning of last century in the following areas: set theory, proof theory, computer (hardware and software) design, and electrical engineering. Though early 20th century computers hardly resembled contemporary PCs, they did exhibit some of the basic properties of a computing device. Programmability, the property that distinguishes a computer from a calculator,⁸⁸ was common in many of the most advanced devices that were designed in the first half of the century. Many computers had also a hard-wired Boolean Logic, and used a binary system to store data. The design of the von Neumann architecture⁸⁹ introduced some defining changes in the field such as a separate compartment to store data and programs from the computing unit (ALU or Arithmetic and Logic Unit). This last feature is important because, for some defenders of classical architectures, it allows for the structured processing of symbols that defines classical digital computation.⁹⁰ Even a simple device like a Turing machine⁹¹ (the paradigm of a classical computer) contains these separate

88 Under certain conceptions of computability, a calculator will also compute. However we're trying to distinguish a particular kind of computing device against general notions of computability.

89 The predecessor of modern computers.

90 Horgan and Tieson (1999) explain: "Classicism is committed to representation-level rules -- that is, programmable rules that apply directly to the formal structure of the representations themselves. But connectionism, at least insofar as it employs fully distributed representations in which local node-activations do not have representational content, is not committed to such representation-level rules as a feature of cognitive architecture." (p. 725)

91 A Turing machine is an extremely simple (idealized) device consisting of: (i) an endless tape divided in squares (each one of these squares contains a symbol out of a finite alphabet), a head that moves along the tape scanning the slots and that can read or write symbols in them, a state register that records the state that the machine is in, and a list of instructions (program) to perform. At any given point, the machine is in one of a finite list of states and reads one out of a finite list of symbols. The

elements: a list of instructions, a tape where the computations take place, and a separate stack for memory. The classical design, based on Von Neumann's architecture is particularly important to cognitive science because the earliest models of artificial intelligence were based on this type of design.

The success of these models plus some important results in Proof Theory and later developments of computer design and programmability, together with Chomsky's innovative approach to language production and acquisition and other general theories of cognition, consolidated the idea of the brain as a structured rule-based symbolic processor.

Connectionism

Unlike a classical computer, a connectionist network is just a vast set of nodes in which the information is stored non-symbolically in the strengths of their connections. Processing takes place via the passage of specific patterns of activation through the network (all units are active to a certain degree at any given point). In processing some input, the network is first supplied with an initial pattern of activation (the input). Then, the active units cause the activation and deactivation of other units, depending on the weight of their connections. At some point the network reaches a stable pattern of unit activation. This new pattern of activation constitutes the result or output of the computation. Depending on how we interpret unit activation, we can take each individual unit to represent an object or property (in this case, we have a localized network), or we can take a complex pattern of activations to represent a single object or property (this is a distributed

program tells the machine what to do considering the state it is in, and the symbol that it is reading. Despite its simplicity, a Turing machine like this one is capable of computing any function that is computable by more structurally complex computers. In fact, the Church-Turing thesis claims that any effectively computable function can be computed by some Turing machine.

network). What usually distinguishes one connectionist model from another —besides the amount of nodes in each layer and the connections between these nodes— is the function that determines the passage of activation through the nodes. In such models there is no need for storage, retrieval, and manipulation of symbols in a structured way.

IV. The Debate

In a famous paper, Fodor and Pylyshyn (1988) claimed that rule-based systems (unlike connectionist models⁹²) exhibit three main characteristics: (i) they are systematic; (ii) they are productive; and (iii) they are compositional. Such systems are said to be systematic because their rules apply to all instances of the same category of inputs. For example, a system that it is capable of processing the input ‘John loves Mary’ is also capable of processing the stimulus ‘Mary loves John’. Since language and thought share this property, Fodor and Pylyshyn (1988) insist that any candidate for the title of ‘cognitive architecture of the mind’ has to be an architecture that is able to model this property. Another property shared by language and thought is productivity: they can both produce an unlimited amount of sentences/thoughts based on a finite set of symbols and a finite set of rules. Mathematical thinking is also productive in this way: it is always possible to represent larger numbers by stringing together more numerals. Finally, we say that a symbolic system is compositional when the content of complex symbols is a function of the content of its parts and the structural relations between them. Language and thought also share this third property. Therefore,

92 “The style of processing carried out in such models is thus strikingly unlike what goes on when conventional machines are computing some function.” (Fodor & Pylyshyn, 1988, p. 4)

only an architecture that can express these properties is a plausible architecture for our cognitive processes.⁹³

But despite these criticisms, connectionism still presents some advantages over classical architectures, especially when it comes to modeling certain processes, like pattern recognition, that classical architectures find very difficult to model. Also, connectionism can explain some important aspects of certain cognitive processes that classical architectures can't account for. One such aspect is time frame. Neural networks are better able to perform certain tasks in a time/step frame that it's similar to that of a mental process. The 100 step principle by Feldman and Ballard (1982) establishes that many basic cognitive tasks require less than 100 steps to be completed. Given that a typical neuronal process takes a few milliseconds to be completed, and the whole of these basic tasks takes some hundreds of milliseconds, this gives a rough approximation of a hundred steps per task. While classical models often have a hard time performing these tasks in the required amount of steps, connectionist networks can easily settle into a result inside this limit (Churchland & Sejnowski, 1990).

Connectionism can also handle contextualization, holistic representation of the data, spontaneous generalization, and statistical patterns better than classical architectures. And, like in human cognition, many networks can take deformed or incomplete inputs and still give an adequate answer (while classical models have a hard time dealing with incomplete or defective inputs).

93 Let us now turn to what counts as a representation. This debate is less clearly delineated. According to some authors, only structures that have a language-like combinatorial syntax, which supports a compositional semantics, count as computational vehicles, and only manipulations that respect the semantic properties of such structures count as computations (Fodor, 1975; Pylyshyn, 1984). This suggestion flies in the face of computability theory, which imposes no such requirement on what counts as a computational vehicle (see Piccinini, 2012).

Further, connectionist networks can be partially damaged/inactive without suffering a major loss in performance, something quite unattainable for classical computers (Churchland & Sejnowski, 1990).⁹⁴

All this suggests that there are big advantages in including connectionism as a possible architecture of the mind. In the next section I motivate a ‘hybrid’ architecture, partly classical, partly connectionist.

So, which one of the two architectures is more appropriate to model the mind’s cognitive processes? On the one hand, we know that a universal Turing machine can compute any function that is computable, so there is no room for preferring connectionism to classic processors on grounds of computational power. On the other hand, though there was a time when connectionism was thought to be less computationally powerful than classical processors, it is actually not the case. A three layered network can approximate any function (given enough units in each layer) (Dawson, Medler, MacCaughan, Willson, & Carbonaro, 2000). So in terms of computability power, the ground has been leveled and there is not much room for preferring one architecture to the other for this type of reasons.

Since formal results cannot decide the matter, the suitability of each architecture to model a particular process has to be decided on the basis of empirical results. According to Lewis (1999), cognitive scientists can use a variety of empirical constraints to determine the suitability of a model, including functionality requirements, patterns of errors, response times, eye-tracking data,

⁹⁴ “Neuroscientists often note the lack of neurological realism in connectionist networks. These networks often have too little recursion, far too much inhibition, unrealistic learning algorithms, simplistic transfer functions, and no analog to the large number of neurotransmitters and hormones which affect human cognition.” (Eliasmith, 2004, para. 5)

participants' verbal reports, and learning data (p. 526). When discussing the merits of connectionism, I also mentioned time and function degradation, reliability, and learning patterns. We use all these aspects of human performance to assess each architecture's success in reproducing it. After all, connectionism became a clear challenger of classical architectures because it could model certain cognitive processes (like pattern recognition) that were very difficult to model in classical architectures.

To sum up, defenders of both classical and connectionist architectures find support for their view (i) in formal results, (ii) in the fact that their architecture is 'good' at modeling a certain process that the other architecture can't be said to model well. So, a great part of the cognitive architecture debate revolves, *pace* Fodor,⁹⁵ around the current capacities of these architectures to model human performance in an empirically adequate way. In the next section, I will review some of these empirical results to support a (partly connectionist, partly classical) 'hybrid' view. Next I will show that a 'hybrid' approach can deal better with the empirical results concerning both type-1 and type-2 processes, and therefore model more accurately some important aspects of these cognitive processes, like reaction times, sequences of steps, degradation of performance.

95 Before moving to the next section, let me point out that Fodor and Pylyshyn's (1988) approach doesn't per se prevent the possibility of a dual architecture of the mind. On the contrary, Fodor's modularity thesis actually seems to support such a view. What they fervently oppose is the idea that 'high cognition' or our 'general thinking capacity' can be modeled by connectionism. They allow that other brain processes, such as perception, are not obviously classical and can perhaps be modeled in a connectionist way. Hence, when assessing the plausibility of a hybrid architecture of the mind, it is important to clearly specify which processes are being modeled by which architecture. I will come back to these issues in section five, where I bring together Fodor's modularity view, the architecture debate, and theoretical aspects of the dual system view.

V. The solution: a hybrid approach

So, assuming that there are two types of processes as described in section one, is it possible to provide a more robust notion of the distinction in terms of two underlying cognitive architectures? I think the answer is yes. In Chapter One, I gave some examples of the empirical data that led psychologists to propose this dual view, and I explained why this data can be explained in terms of the working of two processes. I now propose to reassess the data in terms of a dual-process view understood as a ‘dual architecture’ view. I will show that a ‘hybrid’ architecture not only fits these results, but it also explains why participants’ reasoning processes and responses show that particular set of characteristics, as described at the beginning of this chapter.

As it will become clear soon, a connectionist architecture perfectly fits the results from type-1 processes, while a classical architecture is ideal for modeling the outcomes of type-2 processes. This will not come as a surprise if we take into account that type-1 processes are supposed to process information on the basis of associations, while type-2 processes transform symbols according to their structure. In this section, I show the fit between each process and each architecture. In the next section, I will elaborate more on why we can assume each type of process to have certain sets of properties given that they have these two different micro-architectures.

Linda the Bank Teller

As I pointed out above, it seems that participants are basing their responses on how representative or similar the description of Linda is to that of a feminist bank teller, on the one hand, and to that of an average bank teller, on the other. If this is the case and participants are

evaluating similarity to form their judgments, then connectionism can certainly model this behavior quite well, since similarity and representativeness assessments are usually based on previously stored associations between the items being compared. As I pointed out in my discussion of the two cognitive architectures, one of the key strengths of connectionism is its capacity to model cognitive processes whose computations mainly assess how similar are the features of the new item to the features of previously encountered items. Pattern recognition, for example, is one such case (Rumelhart, McClelland, & the PDP Research Group, 1986).

So, it seems that most participants in the Linda experiment are using some kind of ‘associative’ process to come up with a response. Under this assumption, a ‘hybrid’ architecture hypothesis could very well explain the correct and incorrect responses. If one of the processes is mainly associative, then its results are going to be quite fast but also heavily influenced by previous associations of the stimulus with the target responses. If Linda is mainly associated with someone that is very active in the feminist movement, then the connection between the description of Linda and a feminist bank teller is going to be stronger than the connection between Linda and a bank teller.

Of course, a classical architecture can model very well the correct responses, i.e., those in accordance with probability theory. Though different models might produce different rankings depending on how they deal with lack of information, what we can know for sure is that they wouldn’t rank the conjunction as more likely than the conjunct. This is because probabilistic rules are the types of rules that these models can process very well (in some cases, the models are solely

based on the use of logical and probabilistic rules so they obviously can handle correctly this type of reasoning).

Could a classical architecture model well the actual performance of most participants? Since modeling of cognitive processes is done in digital computers, we do know that if we can create a connectionist model, then there is a way of transforming such a model into a series of rule based computations and obtaining the same results. As I said in section three, what this means is that any cognitive process that can be modeled in a connectionist fashion can also be modeled in a classical way. This does not mean that the classical model will be equal to the connectionist in terms of processing time, simplicity, number of assumptions, etc. Nor does mean that the classical model will be empirically adequate. In particular, the following fact needs to be taken into account: classical processors (in the form of digital computers, for example) can process an enormous amount of steps in a very small time. This is not the case for humans. As I mentioned above, there is a limit of a 100 steps for most simple cognitive processes. So, if we want to argue in favor of classical architectures we need to show that such an architecture can model cognition in a step frame similar to that of human processes.

The question then becomes the following: is it possible to create a purely rule-based model that will produce biased responses (like the ones in the data) in a reasonable amount of steps? One thing we should keep in mind is that people only rank one of the conjuncts as less likely than the conjunction, the other conjunct ('Linda is feminist') is systematically ranked as more likely than the conjunction. This last point makes it more difficult to reproduce the empirical results in a classical architecture. If people ranked both conjuncts as less likely than the conjunction, then we could just

assume that somewhere in their brains there is a rule saying ‘rank conjuncts as less likely than conjunctions’. But this is not what’s going on. It is the specific semantic content of a conjunct and the associations between this content and other beliefs what determines whether it is ranked as more or less likely than the conjunction. And something like this is extremely difficult to model in simple classical architecture because it requires the processing of a vast amount of information (namely the impact of all those associations in the assessment of overall probability).⁹⁶ On the contrary, a connectionist network finds this very easy to do since all that information can be encoded in the distribution of units in the network, and the strengths of connections between them.

Another option is to have a general rule that says ‘replace calculations of probability by calculations of similarity/representativeness’. This option would in principle make the model work, but it has two main flaws. First, it is still a problem to determine how the model is going to get these similarity assessments (as we said above this is difficult to calculate in a classical model). Secondly, and most importantly, why would there be such a rule in the first place? Let’s remember that we are not just trying to model our cognitive processes in order to implement them artificially. We are trying to come up with an empirically adequate model of how the mind really works. So, if the mind does actually process its inputs like a classical processor, it does not make sense to assume that there is a special rule telling it to process the inputs by replacing probability calculations with assessments of similarity (especially when the replacing calculation seems to be so much more complicated for such an architecture!). Adding this rule would just be an ad-hoc attempt to match the results of the model to the empirical results. This is not theoretically sound.

⁹⁶ And this wouldn’t be empirically adequate.

To sum up, similarity assessments are done very quickly and apparently without requiring many computations or steps, so for the time being connectionist networks are the best models for them (Chalmers, 1993; Smolensky, 1988; Churchland & Sejnowski, 1990). On the other hand, classical architectures are perfect to model deliberate rule-based reasoning since this is the type of computation at the basis of this architecture. Since people who have gone through a lot of formal training usually perform better in these tasks, we can assume that it is the engagement of a rule-based type of reasoning like the one that is taught in formal training that is responsible for the correct responses.

Finally, a further consideration for assimilating the two responses to different architectures is the disparate phenomenology of each judgment process. Judging that Linda is more likely to be a feminist bank teller is the result of a fast, automatic, consciously opaque process. Participants don't have conscious access to why they think that way (except for knowing that the information in the description points out in this direction). By making this judgment participants are not straightforwardly following any rule (of the kind 'if someone got a B.A. in philosophy then they are quite likely to be an active feminist') but rather 'getting a feel' for what it is more plausible, what makes more sense. One of the main differences between classical and connectionist architectures is that the first works in a step-by-step basis, while the second carries out many parallel computations at the same time. Even if we took a connectionist network to have some intermediary computational steps (for example the result of the passing of activation between nodes in the first layer and nodes in the second layer), in many cases these intermediate states of the network do not really represent anything of interest. This is due to the fact that most processing in a connectionist

network takes place at a sub-symbolic level. If this is in fact correct,⁹⁷ then intermediate steps are non-symbolic and hence are not meaningful to the conscious mind. By contrast, in a classical architecture, each step is the result of the transformation of a symbolic input into another symbolic output. So, at any given step, one should be able to determine what the result of the last transformation is.

These differences between the two architectures could explain why we don't have access to intermediate steps of computation when our reasoning process is type-1, but we do have access when it's type-2. In other words, the opacity to consciousness of type-1 processes is to be expected on our model; I return to this point below. It also explains why most of the time, we can follow and reproduce our reasoning process in the second case but not on the first. So, at least in the Linda experiment, a 'hybrid' architecture would explain much better all these aspects of the empirical results than positing just one single architecture.

Evans's Logic Experiments

Classical architectures are of course ideal for modeling rule-based behavior. They were first designed with that purpose in mind, so modeling assessments of logical validity in a classical architecture is extremely easy.⁹⁸ However, we need to keep in mind that what is easy to model is *correct* behavior. The problem arises when the architecture needs to model 'less than perfect'

97 There is some controversy over this. See Fodor and Pylyshyn (1988), Chalmers (1993), Smolensky (1988), Churchland and Sejnowski (1990).

98 In all these cases we are dealing with extremely simple arguments, so concerns about computability constraints are not really relevant.

behavior. In particular, if responses are incorrect in a very systematic way, we can't just add some noise to the model and expect it to model incorrect responses in this particular way.

Again, the problem is quite similar to the one we encountered in the Linda example: incorrect responses do follow a pattern but it is not a pattern that it's easy to describe using some simple⁹⁹ set of rules (it's rather the result of strong associations learned over a person's life). So, we could in principle add a rule in the model that says: 'when the conclusion of the argument is believed strongly by the participant, respond 'valid'; and when the conclusion is obviously false for the participant, respond 'invalid'.' But the problem with this approach is the same that I mentioned above: it is a very artificial solution to a problem that can be easily addressed in a connectionist architecture.

In sum, while a classical architecture can model type-2 processes very well, one can't really model type1 processes unless we add an ad hoc rule to the model to match the predictions of the model with the empirical results.

Would a connectionist network do better at modeling heuristic behavior? Definitely. With proper training, a network could learn to produce a response based on the strength of previous beliefs. For this it could use as a guide how strong are the associations or inhibitions between the concepts that appear in the conclusion (so if the association between 'vitamin' and 'nutritious' is very strong then the proposition 'Vitamins are not nutritious' would be believed strongly to be false).

99 So that it doesn't violate the 100 step constraint mentioned above.

Could it also model type-2 responses? In the same way that a computer can reproduce the workings of a neural network by transforming its processes into structured manipulation of symbols, a neural network can compute logical functions by using the passage of activation between its units as logical gates. So, in principle, a connectionist network could also model type-2 responses. But before supporting a single connectionist architecture for this task, we need to take into account Fodor and Pylyshyn's remark that a connectionist architecture that processes its inputs in the way just described is only connectionist at the implementation level - not at the algorithmic level. At the algorithmic level, it is for all purposes the same as a classical processor. Their point is quite simple indeed: they claim that given the biological structure of the brain, it is obvious that at the implementational level the mind is connectionist. But the discussion is not about that level, but about the algorithmic level, which is the level that concerns the 'cognitive' aspect of these processes (and not the physical one). So a big network of connections between units does not by default constitute a connectionist architecture at the algorithmic level. If the architecture processes information according to rules based on the structure of symbols, then the architecture is classical regardless of physical structure.

So, the advantage of connectionism (when not implementing a classical processor) lies in its ability to model type-1 processes. In this case it can use semantic associations (and not logical operations) to produce an empirically adequate response. By contrast, a classical processor can model type-2 processes in a very simple way by taking into account the syntactic structure of the input rather than its semantic associations.

So, as I mentioned already, there is no formal result showing that either architecture can't model a certain cognitive process. However, it is also clear from the history of cognitive modeling that a connectionist architecture does a great job at modeling statistical patterns, but has a hard time at modeling exceptionless inferences of the kind we find in logical arguments (unless it does it by imitation of classical processors, that is, by setting its connections to work as logical gates). On the other hand, classical architectures are made for modeling rule-based processes, but they are quite weak at modeling content based associations.

In Chapter One, I presented a wide variety of reasoning experiments that support a dual view. One type of process is fast, automatic, effortless, and consciously opaque; and the other is slower, deliberate, effortful, and consciously transparent. The first works by processing associations and the second by processing rules.

There were two main reasons for explaining the results in terms of the involvement of two types of processes. Firstly, participants' incorrect responses were highly systematic, suggesting that they were not the product of a random mistake, but of the systematic employment of a different approach. When solving reasoning problems, people often rely on the use of simple strategies instead of using formal methods. The predominance of this 'heuristic' process is evident: even participants who have the ability and training to solve the puzzles correctly often rely on the use of heuristics. Secondly, even after recognizing that the heuristic leads us to a biased judgment, we still feel the pull of the two responses. This suggests, as in the Müller-Lyer case, that there are two different processes trying to solve the same problem.

The next step was to distinguish the two types of processes in a way that explained and predicted the empirical results. In order to do so, I turned to a famous debate in cognitive science about the cognitive architecture of the mind. I then endorsed Marr's (1982) separation of the three levels of analysis of a cognitive process (the *computational*, the *algorithmic*, and the *implementational*) and I claimed that what distinguishes the two types of processes is their algorithmic level. Type-1 processes use associations, while type-2 processes are rule-based. Since the debate between symbolic and connectionist architectures is a debate about the kind of *algorithms* that the brain uses to process information, it is directly pertinent to the dual process view. If the difference between the two processes is at the algorithmic level, then one way of explaining this is to assume that one process is classical, while the other is connectionist. That is, to propose a 'hybrid' architecture for the mind. Finally, I reviewed some of the examples in section two and showed that this 'hybrid' view not only fits the data, but it also predicts and explains it.

Let us now review some of the advantages of a dual-process view understood in terms of two underlying architectures. As in any other science, psychological models need to be empirically adequate. The 'hybrid' architecture view that I endorse in this paper not only makes predictions but these are also confirmed by the data. In particular, this view predicts two important sets of things:

1. when and why participants give a correct answer to a formal problem (given measures of intelligence);
2. why each of the two processes has the properties it has.

Concerning the first point, the dual-process view I have presented predicts that for most reasoning tasks the number of correct responses will depend on: (i) which type of process is

answering the task, (ii) how good are the participants in solving the task, and (iii) if both processes get involved, how much conflict there is between the answers of each process.

For example, if participants are solving an abstract logical problem, then the dual-system view predicts that participants with higher measures of intelligence will produce more accurate responses because (i) the process engaged in answering the task is type-2, and (ii) these processes' capacity is correlated with measures of general intelligence.¹⁰⁰

As it stands, any dual-system theory can predict that performance will probably decrease if there is conflict between the two processes' answers (it can actually decrease a lot depending on which system gets to respond). It can also predict that performance will improve if the two processes agree. In both these cases, the comparison is with the scenario where only one system is engaged.

For example, let's consider a logical task where participants need to determine whether the argument is valid or not. When the arguments are devoid of semantic content, then 'smarter' participants perform better. However, when researchers add semantic content to the arguments, the results can go both ways: in some cases, performance (of all participants) increases, in other cases, performance (of all participants) decreases. Why is this the case? As I said above, the dual theorist can hypothesize that in the first case it is due to conflict between the two processes' responses, while in the second case it's due to their agreement. But why does this happen? Why does the addition of semantic content hinder performance in some cases and enhance it other times?

100 As I mentioned above, this is not surprising given how intelligence tests are constructed.

Well, unless one has a theory of how and why the two processes work differently, it will be very hard to predict and explain this kind of result.

The view I defend does explain (and predict) this result. This view claims that one of the processes is connectionist and therefore works by processing associations (broadly construed), while the other is classical and processes symbols according to rules. If that's the case, then we can expect type-1 to involve semantic content, previous knowledge & experience, and other strong associations, to come up with a response. We can expect type-2 processes to produce a response mainly based on the syntactic structure of the inputs (and not on its content).

So if the nature of the task is such that there is a conflict between what a syntactic analysis suggests and what a semantic one does, then we can predict that, regardless of their intelligence,¹⁰¹ participants will perform generally worse than on the 'abstract' version of the task (one which doesn't have semantic content).¹⁰² And this is exactly what's going on in Evans' logic experiments: the concretization of the task, precisely because it adds semantic content that conflicts with the rule based response, hinders all participants. On the contrary, in those experiments where semantic content was added that 'supported' the logical assessment, all performances improved (see Evans, 2003). This is to be expected given the 'hybrid' architecture view.

Conversely, we can explain the difference in performance in the Wason selection task between the 'concrete' and the 'abstract' scenarios in the same way. While the 'abstract' scenario does not provide any background information on how to solve the task, participants depend on

101 Cognitive conflict usually affects everyone regardless of intelligence, though not always to the same extent.

102 The presence of conflict usually results in slower reaction times and poorer performance.

their rule-based system to solve the task correctly. Since most participants are not very good at abstract logical puzzles, the responses are generally wrong. On the other hand, the ‘concrete’ scenario provokes strong associations between the task and past experience. Because the experience that is invoked is very familiar to most participants, then in this ‘concrete’ scenario, type-1 processes provide the same answer that type-2. The convergence of the two responses is therefore the cause of the dramatic improvement in performance.

Now, let me turn to the second important prediction of the *hybrid* view: why each of the two processes has the properties it has. First of all I would like to distinguish between two types of properties that are commonly assigned to each system: the observed and the unobserved. Type-1 processes’ observed properties are: fast, automatic, effortless, consciously opaque. Type-1’s unobserved properties are: parallel and associative. Type-2’s observed characteristics are: slow, deliberate, consciously transparent, and effortful; while its unobserved properties are: rule-based and serial (though its phenomenology certainly points in this direction).

The reason for drawing this distinction is that the unobserved characteristics are, as the name suggests it, properties which have been inferred from the observed properties but that are strictly unobserved in the data. For example, it’s not a certain fact that type-1 is associative. This is an inference that we make from different aspects of the data, including its observed properties. Similarly, it is not certain that type-2 is rule-based, though it certainly appears to be so given its phenomenology.¹⁰³ In my view, it is a consequence of proposing an underlying hybrid architecture

¹⁰³ The phenomenology gives some support for this claim but it’s certainly not a proof of it.

that the two processes have the (above mentioned) unobserved properties. It is not an independent fact.

Then if we assume that type-1 processes are connectionist, how does this explain why it is fast, effortless, and consciously opaque? Connectionist networks are very fast because they are highly parallel, so computation doesn't require many steps (though it might require an enormous amount of units in each layer).¹⁰⁴ Associative, parallel computations are (phenomenologically) effortless because we are not consciously aware of how we carry them out. Though they of course 'consume' brain resources, we are not consciously aware of them, so we don't 'feel' the burden of having to carry out the computation. Also, we don't have to consciously employ rules. In a network the passing of information is constant, and no decisions need to be made. The machine simply 'runs itself'. Why are these processes consciously opaque? Well, probably because the intermediate steps of the computation are not representative of anything of importance or significance to us, or perhaps even interpretable to us. There is nothing to represent as 'intermediate' results; the computation is carried out all at once (in a certain sense). Moreover, many researchers agree that connectionism is a theory of sub-symbolic processing. If this is right, then we have our explanation of why type-1 processes are opaque: intermediate steps are non-symbolic and hence have no 'meaning' to our conscious mind.¹⁰⁵

Unlike in connectionist networks, classical architectures process the input by applying 'structure-sensitive' rules. They are quite slow compared to neural networks because the processing

¹⁰⁴ Actually, a large enough network can approximate any function with only three layers. See Harman and Kulkarni (2007).

¹⁰⁵ Many researchers agree that connectionism is a theory of sub-symbolic processing. If this is right, then we have our explanation of why S1 is opaque: intermediate steps are non-symbolic and hence have no 'meaning' to our conscious mind. See Chalmers (1993), Smolenski (1998), Churchland and Sejnowski (1990).

is serial. This also means that, at any given point, more than one transformation of symbols could be taking place, which requires some kind of exercise of ‘control’. This is probably the reason why type-2 processes are deliberate and more effortful. Finally, this ‘conscious’ exercise of control over the steps of the computation is what makes many type-2 processes consciously transparent (though of course we have to assume that there are many other ruled-based processes going on in the mind to which we have no access).

VI. The Advantages of a Hybrid Architecture

It seems then that a ‘hybrid’ architecture can explain the participants’ responses better than a single (classical or connectionist) architecture. And this is not because a single architecture (of either kind) cannot reproduce the correct input-output connections. It is rather because one single architecture cannot reproduce the empirical data without resorting to ad-hoc modifications. Further, even without these ad-hoc modifications, each architecture is particularly weak at modeling the results of one of the two processes (while the other architecture predicts and explains the empirical results quite well). So, it is a much better theory of the workings of the two processes (both on empirical and theoretical grounds) that which assumes that each system is best modeled by a different architecture.

More importantly, a hybrid architecture of the mind shows the complexity of human cognition: cognitive capacities are not isolated boxes in one big cognitive container. Rather, many areas and processes intervene in what we currently classify as ‘memory’, ‘learning’, ‘reasoning’, etc.

As we add empirical data on human performance¹⁰⁶ we will keep fine tuning these models. But so far, it has proven impossible to choose one architecture to model all processes.

And, as a matter of fact, the best current models of the whole of cognition use hybrid architectures where classicism and connectionism are combined to produce the most complete picture of the mind. For example, Sun and colleagues (Sun, Merrill, & Peterson, 2001; Sun, Slusarz, & Terry, 2005) propose a computational model, CLARION, that combines associative and rule-based components.

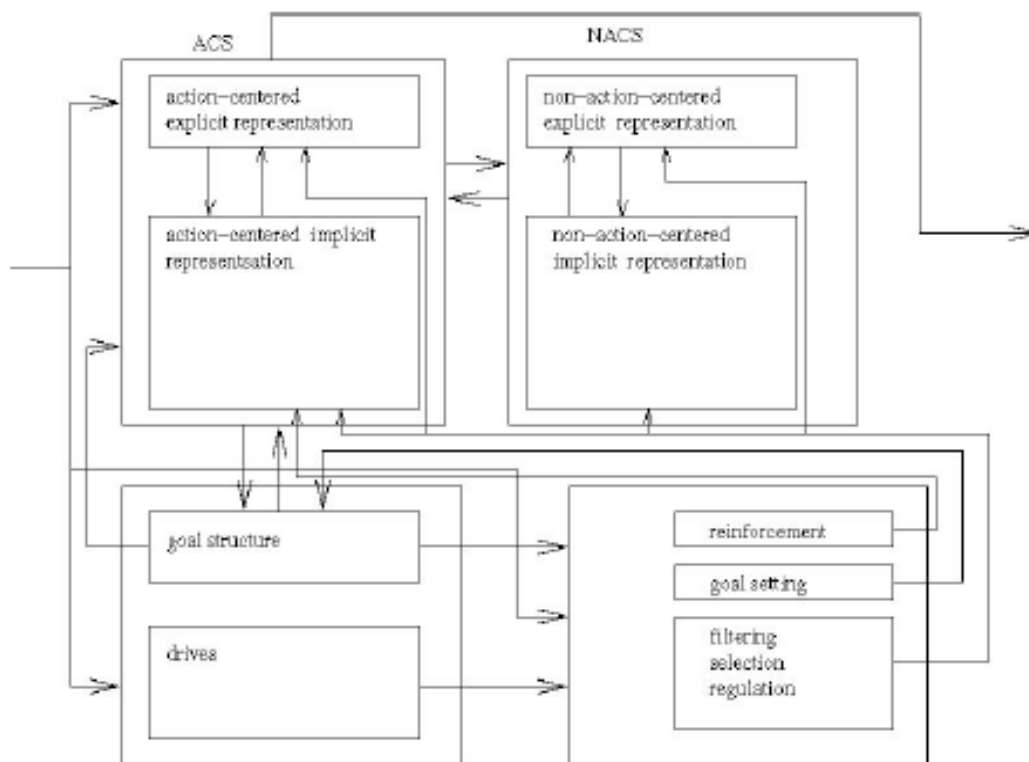


Figure 4. The CLARION architecture. (Taken from Sun, Lane, & Matthews, 2009, p. 245)

¹⁰⁶ This involves data from cognitive development, expertise, neural localization, neuropathology, etc.

The motivation for using a hybrid architecture stems from many of the empirical results described in the dissertation,¹⁰⁷ especially those contrasting implicit and explicit learning. The more experience the system acquires, the better tuned the associations become and the more automatic and less requiring of central control the whole process is. Regarding this type of hybrid model, Sun et al. (2009) remark that:

A general pattern discernible from human data (especially those from the implicit learning literature) is that, if to-be-learned relations are simple and the number of input dimensions is small (in other words, if the relations are salient to participants), explicit learning usually prevails; if more complex relations and a larger number of input dimensions are involved, implicit learning becomes more prominent (Mathews et al. 1989; Reber 1989; Seger 1994; Sun et al. 2001; Sun et al. 2005). [...] This pattern has been demonstrated in artificial grammar learning tasks, serial reaction time tasks, and dynamic control tasks. Seger (1994) further pointed out that implicit learning was biased toward structures with a system of statistical relations. Thus, the implicit learning mechanism appears more structurally sophisticated and able to handle more complex situations. (p. 246)

In sum, the advantages of a hybrid architecture are several:

- Classical architectures can account for some important features of *some* cognitive processes, such as compositionality, systematicity, and productivity.

¹⁰⁷ “Hybrid connectionist-symbolic models constitute a promising avenue to the development of more robust, more powerful, and more versatile architectures for both cognitive modeling and intelligent systems.” (Sun, 1996, p. 99)

- However, classical architectures assume that cognitive processes can be described in terms of mathematical functions, and can be modeled like any other function is modeled in computer science, that is, by specifying the rules that determine the transformation of inputs into outputs. This is a conceptual mistake, not all cognitive processes are describable by computable mathematical functions. We can think of three cases:
 - The process can be described as a function if we introduce other variables into the input, but this function is not (Turing) computable.
 - The process itself is not describable in a functional manner because there is an element of randomness in the environment that cannot be accounted for in computational terms.
 - The process cannot be described in functional terms because the same input might be classified in different ways depending on other factors. These factors cannot be included as part of the input because they depend on the particular experiences of the participant (these types of models are implemented by the SVM framework).
- Even in those cases where the computability of the function is not an issue, there might be other reasons (for example, related to results in cognitive psychology) that make us prefer a different type of architecture.

Finally, further support for a hybrid architecture comes from recent neuro-scientific research. Goldberg (2009) claims that both brain hemispheres work simultaneously to address a myriad of tasks: the right hemisphere is a general processor that addresses the task whether it has specific knowledge of it or not, the left hemisphere forms ‘fine-tuned’ nets that with time learn to classify specific inputs. The more the brain addresses a classification problem, the better it becomes

to the point that the left hemisphere's response will compete and overturn the right hemisphere's 'general response' (Goldberg, 2009, Chapter 6).

VII. Some Concluding Remarks

In this chapter I have shown that it is possible to distinguish the two types of processes by their underlying micro-architecture and show, at the same time, why the processes' properties cluster in the way described in the literature. The next step would be to distinguish the two systems on the basis of this micro-architectural distinction: there are two types of systems, S1 and S2, which produce two types of processes, type-1 and type-2, which are distinguished by their different computational principles, connectionist versus rule-based. Such a dual-system view would provide a robust characterization of the notion of system and the distinction between the two types. However, I do not think this is a viable solution. I will briefly state some of the reasons and will address them in more detail in the next chapter.

Can hybrid micro-architecture view support a dual-system type view? The answer is 'no' and the main reasons are the following:

1. A hybrid or dual micro-architectural view claims that processes can be distinguished on the basis of the computational principles they employ to transform an input into an output.

Systems, on the other hand, are distinguished by their neural substrates, their function, and the processes they use to transform inputs into outputs. So, in principle, a system whose neural substrate and function are unique could engage several processes, not all associative or rule-based. That is, nothing precludes the possibility of there being hybrid systems (systems that

employ both processes in their normal functioning), rather than dual-systems (systems that alternatively employ two kinds of processes). If different systems can employ one and the same type of process, why can't one and the same system employ different types of processes? There is no reason to preclude this option a priori. In Chapter Four I will show that it is a much more reasonable view of the cognitive architecture of the mind to posit the existence of hybrid systems, rather than dual ones.

2. As mentioned in Chapter Two, the dual-system type view relies on a robust clustering of properties for each system. If a system's properties are determined by its processes' properties, then most of the systems' processes need to share a sufficient amount of properties from the cluster. Further, if the dual-system view claims that systems are of two types, then there should be two main clusters of properties, one for each system. However, as I will show in the next chapter, there are systems that show properties from both clusters. This leads to the conclusion that either these systems employ both types of processes or that there are yet other kinds of processes, processes that share properties with both types. This would not refute the work exposed in this chapter – which merely attempts to show why each architecture tends to produce a certain cluster of properties. Rather, it would suggest that a process' properties might depend in factors beyond its underlying micro-architecture (that is, the micro-architecture explains to some extent the clustering but does not determine it).

In Chapter Four, I will address these issues in more detail. I will also examine the fourth interpretation of the dual-system view, the dual macro-architectural theory. This view claims that the macro-architecture of the mind is one of duality, either two minds or two sets of systems, some

type-1 and some type-2. Beyond the problems faced by the dual-system type view, I believe that there are issues with this macro-view. For once, this view does not take into account the importance of a third type of process, executive processes, which need to be distinguished from the types of processes/systems usually studied in the literature.

Chapter Four

In the previous chapter I have defended the claim that processes can be distinguished on the basis of the cognitive architecture that best models them. At the end of the chapter I claimed that this distinction, at the micro-architectural level, cannot support the dual-system type view. There are several reasons for this.

1. First, it is quite probable that most systems, in their current definition, involve both types of processes. Indeed, most of the cognitive mechanisms that are referred to as ‘systems’ (e.g., declarative memory, language production, logical reasoning) involve very complex set of processes, not just a simple, unified transformation of input into output. Also, many systems rely heavily on common processes that are modeled by different micro-architectures.¹⁰⁸
2. Secondly, even if we constrained the definition of ‘system’ to refer to a set of processes modeled by only one architecture, we would still need to account for the interaction of these smaller systems with other systems in a way that accounts for the *migration* of processes and conflicting simultaneous responses. For reasons that I will spell out below, this would be very difficult if we collapse the notion of system to that of process. These two phenomena (migration and conflict) can be better explained when we allow for hybrid systems (systems that employ and depend on both types of processes) rather than dual-systems.

¹⁰⁸ See Evans (2010), Chapter 2.

3. Finally, the dual-system view depends on the clustering of properties in a way that a dual-process view doesn't. The dual-process view explains the clustering of properties by proposing different underlying architectures for different processes. However, it does not need to account for exceptions since phenomenal properties (like effort or consciousness) can be due to other reasons beyond micro-architecture. Indeed, the clustering of properties gives plausibility to the hybrid view, but the view doesn't rely on it. By contrast, a dual-system type view (like the one defended by Samuels and other researchers) relies entirely on the clustering effect: without the clustering effect, there is no reason to suppose that there are two kinds of systems. Further, the lack of a robust clustering raises the question of how to account for processes that have properties from both clusters—e.g., are there other types of system that produce those processes? Then why a *dual* and not a *many* systems view?

In order to address these issues and provide an empirically and theoretically adequate view of the macro-architecture of the mind, there is some important conceptual work to be done regarding the notions of *system* and *process*. We need to establish individuation principles for both entities and specify the relation between them. For example, do the properties of a process determine/define the characteristics of the system that underlies the process? Or is it more appropriate to individuate processes by the characteristics of the system? Do different systems engage the same type of processes? Does one system underlie many types of processes? I will address these concerns in the next section. Then I will re-assess the viability of the dual-system type theory in the context of a more stringent definition of the notions of 'system' and 'process'.

I. Systems and processes revisited

Confusing processes with systems

The conceptual confusion regarding *systems*, *processes*, *mechanisms*, and *processing modes* is ubiquitous. In what follows I illustrate the ways in which different authors confuse these notions.

Evans and Frankish

In his earlier work, Evans (1984, 1989) distinguished heuristic and analytic *processes*. But in later work he sometimes referred to these processes as the analytic *system*:

The characteristics that determine analytic system intervention, other than cognitive ability, are dispositional. People may choose to engage in effortful analytic thinking because they are inclined to do so by strong deductive reasoning instructions (Evans, 2002) or, perhaps, because they have personal motivation. (Evans, 2006, p. 383)

A key feature of hypothetical thinking is the tendency of the analytic system to hold on to representations that are merely *good enough*, leading to such characteristic features as the endorsement of fallacious inferences in deductive reasoning and apparent confirmation biases in hypothesis testing. (Ibid., p. 383)

And in his writings with Frankish Evans talks about *processing modes* and *processing mechanisms*, which makes the picture even more complicated:

These theories come in different forms, but all agree in positing two distinct processing mechanisms for a given task, which employ different procedures and may yield different, and sometimes conflicting, results. (Frankish & Evans, 2009, p. 1)

Finally, Evans assimilates dual-system and dual-process views, processes and processing modes:

What dual-process theories have in common is the idea that there are two different modes of processing, for which I use the most neutral terms available in the literature, System 1 and System 2 processes (Kahneman & Frederick 2002, Stanovich 1999). (Evans, 2008, p. 256)

Clearly, the notions of *system*, *process*, and *processing mode* need to be distinguished, or an explanation for their assimilation needs to be provided.

Frederick and Kahneman

The two authors claim that a system is just “a label for collections of cognitive processes that can be distinguished by their speed, their controllability, and the contents on which they operate” (Kahneman and Frederick, 2005, p. 267). They also claim that “the first section introduces a distinction between two families of cognitive operations, called *System 1* and *System 2*” (ibid., p. 1).

Finally, in his acceptance speech for the Nobel prize, Kahneman (2002) explains that Tversky and himself “held a two-system view, which distinguished intuition from reasoning. Our research focused on errors of intuition, which we studied both for their intrinsic interest and for their value as diagnostic indicators of cognitive mechanisms.” And later he talks about “the

differences between the two modes of thought” and says that “there is considerable agreement on the characteristics that distinguish the two types of cognitive processes, which Stanovich and West (2000) labeled System 1 and System 2” (p. 2).

Frederick (2005) claims that “many researchers have emphasized the distinction between two types of cognitive processes” (p. 26). He also adds that “System 1 processes occur spontaneously and do not require or consume much attention” and that “System 2 processes—mental operations requiring effort, motivation, concentration, and the execution of learned rules” (ibid.). Though these claims are consistent—System 1 and System 2 are collection of processes—they make unclear the difference between a system and a process, since S1 is not just the collections of processes exhibiting type-1 properties, it can sometimes refer to one particular type-1 process, sometimes to all of them.

Reber

Reber (1989) defines implicit learning as the “process by which knowledge about the rule-governed complexities of the stimulus environment is acquired independently of conscious attempts to do so” (p. 219). According to this quote, learning is a process rather than a mechanism/system. But later Reber claims that “not much is known about how conscious, explicit processing systems interact with the implicit and unconscious” (p. 224). He also mentions the “the explicit system” and the “tacit system” in other sections of the same article (p. 224).

Samuels

Samuels (2009b) sometimes uses the terms ‘processes’ and ‘systems’ interchangeably (“positing two processes or systems with broadly similar characteristics”, p. 129), sometimes he moves from claims about the duality of processes to the duality of systems (“dual-process theories are in the business of providing (at least partial) answers to questions about what cognitive systems there are, and what properties they possess”, *ibid.*). Samuels also associates the notions of ‘system’ and ‘mechanism’ and claims that systems and mechanisms “subserve” cognitive processes:

To a first approximation, what the *Token Thesis* maintains is that each mind contains two *particular* cognitive mechanisms or systems. The first (...) subserves those cognitive processes that tend to exhibit the S1- property cluster. The second mechanism (...) subserves those cognitive processes that tend to exhibit the S2-property cluster. On this view, then, each human mind exhibits a fundamental, bipartite division into particular systems. (*Ibid.*, p. 133)

Samuels also mentions that dual-process theorists believe that the two types of processes reveal the existence of two types of mechanisms (a claim he later agrees with):

Thus according to dual-process theorists, not only do cognitive process exhibit disjoint clusters of properties, but this putative fact also provides reason to posit a bipartite division between cognitive *mechanisms*. (*Ibid.*, p. 131, his italics)

Further, Samuels explicitly says that the clustering of properties for both types of processes is due to two different underlying mechanisms, a claim that is taken for granted rather than explained:

Though positing mechanisms is a standard strategy for explaining the existence of property clusters, it does not, by itself, constitute a satisfactory explanation. Rather one needs to specify those features of the proposed mechanisms that account for such clustering effects. (Ibid., p. 141)

One could argue that the clustering of properties just shows the existence of two types of processes, not of two mechanisms.

Sloman

Sloman's (1996) dual-system view of reasoning was particularly influential in the late '90s. Though the view was restricted to reasoning processes, Sloman also defends the existence of two systems of reasoning, rather than one system that employs different strategies. Interestingly, Sloman distinguishes the two systems on the basis of the computational principles underlying their processes, a view similar to the dual-process view defended in Chapter Three. According to this view, System 1 processes information based on regularities of the environment, while system 2 processes inputs according to abstract rules. Sloman adds that "associations capture structure not by indicating how to calculate it but by representing it directly" (p. 4). Clearly this view individuates systems by their processes' computational principles. Indeed, Sloman (1996) claims that: "My central claim is about the principles that govern computation in the two systems, not the details of the systems' processing" (p. 2).

Again, this kind of view reduces the individuation of systems to that of processes: a system is individuated by the characteristics of its processes. In this type of view, the difference between the two systems is a difference in the processes employed. This is conceptually confusing since it is

unclear what the role of the notion of ‘system’ is. Sloman (1996) claims that “I therefore limit my use of the term associative system in reasoning to mean a cognitive system that draws inferences on the basis of similarity and contiguity” (p. 4). Clearly, a system is individuated by its computational processes. Later he adds that “I have said two forms of reasoning exist and can be differentiated by the computational principles that they implement” (p. 5). Sloman talks about ‘forms of reasoning’, which leads to more conceptual confusion. Are these *forms of reasoning* supposed to be ‘systems of reasoning’, ‘reasoning processes’, ‘principles of reasoning’ or something else? The lack of clarity regarding these notions just emphasizes the difficulties faced by the dual-system view.

Smith & DeCoster

Following Sloman, Smith and DeCoster (2000) proposed a dual-system view of moral judgment and decision making. They claim that each system implements a different *processing mode or strategy*,¹⁰⁹ while System 1 draws on associations, System 2 draws on symbolically represented rules. They also proposed a dual-theory of memory mechanisms. In their view, there is a slow learning memory mechanism that creates associations, which S1 uses in its computations. S1’s processing mode is associative and “it takes the form of pattern-completion or similarity-based retrieval from the slow learning memory system, cued by salient features of the input” (p. 110). There is also a fast learning memory mechanism that helps S2 make use of recently acquired rules. S2’s processing mode is rule based and it also relies on the slow learning mechanism: “rules may be stored in either processing system, depending on such factors as how frequently they have been encountered (i.e., just one or two times vs. many times) and over what length of time (i.e.,

109 Yet another term to add to the list of unclear notions (Smith & DeCoster, 2000, p. 108).

whether consolidation has had time to occur)” (p. 110). These quotes reflect the kind of confusion so present in the literature. The authors talk interchangeably about “a system with two processing modes”, “two systems”, “two mechanisms”, and “processing system” (a mix of the two notions!).

Individuation conditions for systems

These are not isolated cases. Most of the dual-process and dual-system literature is plagued with this kind of conceptual confusion. In the dual-system literature – and in cognitive psychology in general – the most common usage of the term ‘system’ is to refer to a mechanism or collection of processes responsible for a specific cognitive capacity. For example, the memory system is the mechanism or set of processes that underlies the capacity to form memories, the visual system is the set of processes that process visual information. At the neurological level, a system is the particular brain structure where those processes take place.¹¹⁰

It seems that both notions can refer to either a more general or a more precise unit, and this is the reason why, in some of their meanings, they are taken to be synonymous. It is important though that, for the purpose of analyzing the plausibility and merits of the dual-system view, we distinguish the two notions.

As I mentioned in Chapter Two, there isn’t much theorizing about systems, nor their relation to processes. I will follow Schacter and Tulving (1994) and Keren and Schul (2009) in proposing the following characterization:¹¹¹

110 At this level of explanation, if we can dissociate the performance of two brain areas, they are part of different systems (or at least sub-systems).

111 As I mentioned in Chapter Two, Schacter and Tulving (1994) claim that systems are distinguished by the type of information they process, the operations they use to process the information, and their neural substrates, and that one system can use more

A system is a specialized unit (or complex mechanism) that employs different types of processes to transform specific inputs into outputs. It is defined by the type of information it transforms (the informational content), the operations it uses to transform them (the processes), and the brain localization where the transformation takes place (the neural substrates). More importantly, a system is defined by its functional properties.

In this last point I follow Samuels (2009b) and others (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004; Oaksford & Chater, 2001) for whom the function or *goal* of a system is crucial for its individuation.¹¹²

In sum, a system is defined and individuated by its neural substrates, its informational content, its computational principles, and its functional properties. Given that there are several aspects that determine the individuation of a system, at times these might conflict. What happens if two sets of process share functional but not computational properties? What if two sets of processes are computationally the same but play different roles in cognition? One could insist on separating all these processes into different systems, however, this strategy would lead to a rapid multiplication of systems. I think that a good solution is to emphasize functional aspects and, if needed, use computational aspects to decompose a system into relevant sub-systems. After all, computational considerations are important but not the primary concern of cognitive science, which seeks to

than one type of process. Keren and Schul's (2009) claim that a system is a specialized unit that transforms specific inputs into outputs, and that a process is a component of the system. Hence the same process can run in different systems and with different contents.

112 "For psychological mechanisms are, by very widespread consensus, individuated by their functional and computational properties." (Samuels, 2009b, p. 136)

characterize cognitive capacities. It is only secondary to this goal to address questions about cognitive processes' computational principles.

The way in which we identify systems shows the importance of functional aspects: researchers speak of 'memory systems', 'learning systems', 'language system', etc. and not of 'classical system' and 'connectionist system'. Clearly, identifying how a system processes its information is subsidiary to identifying the function of the system.

Of course, some researchers might endorse a more fine-grained ontology, which includes systems and sub-systems.¹¹³ I think that this could be a good solution for the current data: systems are complex mechanisms that include many sub-systems whose computational principles and informational content differ in some respect. Again, the distinction between systems and sub-systems is a lax one and generally depends on the level of analysis of the theory proposed.

Someone might respond that the main point of the dual-system view is to claim that beyond individuation by function, one needs to distinguish systems by their computational properties. Under this view, there is no unique learning system, there is an *associative* learning system and a *classical* learning system. There is an *associative* reasoning system and a *classical* reasoning system, etc. I think that for certain capacities, the division on the basis of computational principles can be useful, especially if it is accompanied by a distinction in neural substrates (as is the case with differently located memory systems). But I do not think that this is a solution that generalizes to all higher functions.

113 Similar to what happens with processes and systems, the distinction between systems and sub-systems generally depends on the level of analysis of the theory proposed.

Next I will mention some problems related to the view that the two types of processes originate in different systems.

1. An overly complex ontology of systems.

As mentioned above, a system is more than any particular process it engages. Let's suppose a specific neural structure is associated with a particular cognitive function (e.g., declarative memory) and that function involves a set of key processes (encoding, rehearsal, and retrieval). There might be reasons for modeling some processes classically and some not. Still, we want to say that they are all part of the same system. Indeed, Schacter and Tulving (1994) point out that memory encoding processes are a part of the memory system, not a system in themselves.¹¹⁴ If each process constitutes its own system, we would have an extremely rich ontology of systems, one difficult to justify theoretically.

2. Some systems would cease to exist.

There are certain cognitive capacities that are too complex to be modeled by a simple process. Insofar we want to keep referring to certain capacities as 'systems', it would be very difficult to account for the operations of such systems in only a classical or connectionist way. For example, language is one such capacity. If we were to distinguish systems by their computational principle, then there would be no *language system*. Indeed, there is no (so far) proven way to model language

¹¹⁴ "The notion of a memory process too must be distinguished from that of a memory system. A memory process refers to a specific operation carried out in the service of memory performance. Processes such as encoding, rehearsal, activation, retrieval, and the like are constituents of memory systems but are not identical with them. Indeed, there are good reasons to believe that particular memory processes may participate in the operations of more than one memory system." (Schacter & Tulving, 1994, p. 12)

processes in just a classical or connectionist way.¹¹⁵ While semantic and pragmatic processes are easier to model in a connectionist way, grammatical processes are more easily modeled classically.

3. It defies current cognitive architectures.

Lately the trend in cognitive modeling has been towards the integration of connectionist and classical modeling. Many current cognitive architecture employ both types of processes for one and the same system. In Chapter Three I presented in detail Sun et al.'s (2009) Clarion architecture, a dual-system model for learning in which the two systems (and their subsystems) employ both associative and classical processes. Finally, one of the advantages of this type of hybrid architecture is that it allows to account for the interaction between different sub-systems. Finally, the use of two kinds of representation for the same system accounts for differences in accessibility¹¹⁶ and consciousness.¹¹⁷

4. It doesn't explain the *migration* of processes.

According to most researchers, type-2 processes can become type-1 with practice and the acquisition of expertise. But if type-2 processes are serial, conscious and rely on working memory

115 On the one hand, language is a fast, automatic, intuitive and consciously opaque process (we don't know how we do it). On the other hand, it clearly employs rule-based operations. So, some of its characteristics make language a S1 type of reasoning, some make it a S2 process. Further, according to Berry and Dienes (1993), we can unconsciously learn rules without having explicit access to them. This would be consistent with our knowledge of grammar rules, which is fast, automatic, and unconscious though not necessarily associative.

116 "The inaccessibility of implicit knowledge is captured by (subsymbolic) distributed representations (e.g., provided by a backpropagation network; Rumelhart et al, 1986). The accessibility of explicit knowledge is captured by symbolic/localist representations. Thus, this distinction there is intrinsic instead of assumed. This is a more principled way of accounting for the accessibility difference." (Sun et al., 2009, p. 20)

117 "A central thesis of the theoretical model outlined earlier is that direct accessible representation, along with explicit manipulability (on directly accessible representation), constitutes the essence of consciousness (see Sun, 1997, 1999). Thus, it naturally embodies the difference between the conscious and the unconscious through the use of localist/symbolic and distributed representations in different levels of the model, and provides a plausible grounding for the notion of consciousness." (Ibid., pp. 20-21)

for their performance, in what sense can we say they have become type-1, if they are later supposed to be associative, parallel, unconscious, and not relying on working memory? It would seem this new process is a completely different one, not a modified one. However, if we take systems to be complex units of several processes, then the acquisition of skill or expertise would consist in the recruitment of more type-1 processes and less type-2. So, for example, when a chess player becomes more experienced, that means that when making a play the processes involved in that play would be proportionally more type-1 than type-2. For example, the player could have acquired recognition skills for complex patterns of board positions and hence rely more on that type of process rather than rule-based thinking (more typical in novice players). In sum, a hybrid architecture for systems explains the migration of processes in a simple way: depending on circumstance and expertise, a system might rely more on one particular type of process. So, a system might seem type-1 because it relies (mostly) in type-1 processes but it also uses type-2 processes. In fact, the migration of processes can be due to a bigger reliance on one type of process versus another.

Hence, while current computational models give support to the idea of a distinction at the level of processes¹¹⁸ (the main point of Chapter Three), I believe that this distinction is not enough to support a distinction at the *system* level. There are several reasons to keep the connectionist/classical distinction as a distinction at the level of processes and not systems. First, systems are different entities than processes. A system is much more than the implementation of a

118 Samuels (2009b) agrees: "At any rate, the most plausible computational models to have emerged from cognitive science and artificial intelligence –for planning, deductive reasoning, causal explanation, and so on— vary considerably both in the sorts of computations that are taken to be relevant and the representational formats that are used (Russell & Norvig, 2003). But if this is so, then to the extent that we take attempts at computational modeling seriously, we should suppose such processes differ in important computational respects." (p. 136)

particular process and it is therefore not uniquely individuated by it. Secondly, many systems are likely to contain several sub-systems and hence combine connectionist and rule-based processes. Finally, allowing one and the same system to employ both types of processes explains two difficult cases for current dual-system views: system that show a clear mix of properties (like language), and the automatization of certain tasks (sometimes referred as ‘migration of processes’).

In sum, there is no reason to posit a difference at the level of systems to explain the difference at the level of processes. Further, I think that separating the two claims – the dual-process thesis from the dual-system thesis – allows for a more empirically adequate micro and macro-architecture. I will address issues relating to macro-architecture in section III.

II. Some final remarks on the dual-system view

For all the reasons that I have stated in the previous section, I do not believe that a distinction at the level of computational principles is a good basis for distinguishing two *types* of systems. However, there could be other reasons for distinguishing the two types of systems, like different involvement of working memory (Evans, 2009) or different levels of analysis (Frankish, 2004). I do not think that either of these options is a good alternative to support a dual-system view. On the contrary, I think that all these distinctions (computational principles, use of working memory, level of automaticity) are important notions that can explain some of the clustering. Further, they can produce fairly similar sets of contrasting processes. But neither distinction is sufficient by itself to divide processes into two unique types. Hence, they are not a good basis for

supporting a distinction at the system level: without a robust clustering of properties, there are no reasons to support a dual-system view.

The dual-system *type* view depends on the clustering of properties in two ways. First, there has to be a consistent clustering so that we can infer the existence of two (and *only* two) types of processes. Secondly, the clustering needs to be due to some reason that leads us to reasonably suppose that there are two kinds of mechanisms producing the clustering effect. Hence, the dual-system type theory would face considerable criticism if any of the following claims were true:

- There is no strong clustering of properties.
- There is a strong clustering of properties but it indicates the existence of more than two types of processes.
- There is a strong clustering of properties, which indicates the existence of only two types of processes, but the clustering does not indicate the existence of two types of systems.¹¹⁹

Regarding the first and second claims, there are several reasons why one might think that the clustering of properties is not robust enough, or that it shows that there are more than two types of processes. As I mentioned in Chapter Two, even a reduced list of properties (e.g., speed, consciousness, effort, and control) do not neatly divide into two kinds. If we add to this list several other properties that are usually mentioned in the literature – like capacity, content, rules of operation, etc. – then we are definitely faced with a plethora of processes, some having most type-1

119 Indeed, according to Samuels, we infer the existence of an underlying mechanism (or system) to explain the clustering (in the same way that we infer the underlying virus to explain the cluster of symptoms). Hence, without clustering, there is no reason to infer an underlying mechanism or system.

properties, some type-2, and some both. Further, processes that share characteristics from the two clusters present a challenge for the view because they lead to the natural conclusion that there are other types of systems (beyond S1 and S2) or that the clustering of properties is not a proof of the existence of two cognitive kinds, S1 and S2. (Below I will mention a few examples where the main characteristics of each cluster come apart.) Secondly, even though one can explain the clustering of properties for certain processes given the modeling tools used, the clustering is not sufficiently consistent that one can ignore the exceptions. Cognitive processes are highly diverse and only when considering a selective sample is there enough consistency to suppose a common underlying cause. As I mentioned in Chapter II, in the case of the dual-system theory, the issue at stake is whether one can posit the existence of two natural kinds, System 1 and system 2. So far, the only reason proposed for doing so is the claimed existence of processes with two different clusters of properties.

Regarding the third claim, Samuels (2009b) believes that a robust clustering of properties clearly indicates the presence of two mechanisms, each responsible for one kind of process. Under this view, the two types of processes are the products of two types of systems. These systems are spread in the brain and produce the different dual-process views: two systems for reasoning, two systems for decision making, two systems for social cognition, and so on. The problem with this approach is that it assumes that the data is always identifying the same two types of processes, and hence that the clustering of properties is always due to the same reason. I disagree with this. I think that, though familiar in nature, dual-process views are describing more than two types of processes and that the clustering of properties for each type of process is due to different reasons.

Let's consider, for example, type-1 processes. These processes are fast, automatic, low effort, high capacity, parallel, associative, implicit, and domain specific. Is there a unique reason for this clustering of properties? Not always. Some properties can give rise to similar clusters of properties, which means that 'the clustering effect' can actually be the result of different phenomena and cannot be attributed to one and the same cause (as with the influenza symptoms).

So, for example, automaticity leads to faster times, low effort, less attention and less feeling of consciousness. On the other hand, associative processes are also fast, unconscious, low effort and high capacity and, many times, automatic. Yet, implicit processes can be associative, automatic, unconscious and low effort. Finally, an unconscious process will certainly require low effort and be automatically engaged, but we couldn't a priori say whether it's associative, fast, and domain specific. But all these characteristics can come apart.

What about type-2 processes? Can we find a common cause for the other clustering of properties? Again, I think there isn't such common cause. Rather, when processes critically involve a resource that is limited then they tend to exhibit similar characteristics. For example, attention is a limited resource, so the involvement of attentional processes will necessarily lead to a bottleneck effect, where the process becomes slower, effortful, serial (only one step can be processed at a time). Similarly, the use of working memory has the same effect. So does the feeling of consciousness: the stream of consciousness is a unique thread that introduces thoughts and impressions one at a time. Any process that passes the threshold of consciousness is going to be presented slowly and serially to the mind.

There is the temptation to bundle together these three notions (attention, consciousness, and working memory) and assume that it is one and the same type of process that share all those qualities, and that there is a unique system that produces these kind of processes. However, there are many reasons not to do so. As evidence shows, consciousness and attention come apart, and working memory is a system of its own, that is used differently than attention.¹²⁰ Finally, as Evans (2009) and Samuels (2009b) point out, there are definitely many systems that rely on working memory and compete for attentional resources.

Finally, there is an additional reason why binding together a process' properties to its system's properties is problematic. When we consider how to describe performance of a certain task (learning a list of items, judging the validity of an argument, etc.), assuming that we want to assign the performance of the task to a specific system (a learning system, a reasoning system, etc.), different aspects of the performance can manifest conflicting properties. In that case it would be difficult to specify which one of the differing properties should be assigned to the system at large. For example, several tasks (driving, reading, playing a sport, etc.) are performed *automatically* and with absence of awareness. However, this only applies to part of the process of performing those tasks: for instance, though we are not aware of the details of how we're doing what we are doing, we are certainly aware of the process itself. That is, when I type, I might not be aware of how I do this, but I am certainly aware that I am typing and might even be aware of certain movements needed to type (similar claims can be made about automaticity).

¹²⁰ Recent research suggests that there might be more than one working memory system:

In sum, if systems are defined by their processes' properties, then the task's performance would have to inherit the properties of the system. The problem arises when we encounter tasks that combine some of these properties. Then, what system is in charge of these tasks? Or, alternatively, how do we define the task? It is difficult to say whether a task is automatic or controlled, if some parts of performing the task are automatic and some are controlled. Similarly for awareness, consciousness, and speed. And the problem is amplified when we consider that properties are supposed to cluster for processes, and hence for systems and their tasks. That is, even if we can assign one property of each dichotomy (fast/slow, aware/unaware, explicit/implicit, etc.) to the task, the task might combine properties from both clusters, making it hard to assign it to either S1 or S2.

For these reasons (I will mention a final objection in section III), I believe that talk of 'System 1' and 'system2' should be abandoned. It is fruitful to talk about the different systems in the brain but I do not think they are simple enough to be divided into two different types. I do think that talk of type-1 and type-2 process is fruitful. At that level a distinction can be drawn. We can say that processes that have several of the characteristically type-1 properties are type-1 processes and vice-versa for type-2. Still, there are processes that combine properties from both types and that need to be explained in other terms.

Finally, there is a third type of process that has not received a lot of attention (since it is not responsible for one particular cognitive faculty) but that is crucial to understand macro-architectural questions, given its role in coordinating the whole of the cognitive activity. These are executive processes, which include control and conflict resolution processes.

In Chapter Two, I mentioned that there are two notions of cognitive architecture that are relevant to the dual-system literature: micro and macro-architecture. I also mentioned that there are several options for what it means to defend a dual-system view:

1. A dual-system token theory.
2. A dual-system type theory.
3. A dual-process theory (or a dual theory of micro-architecture).
4. A dual theory of macro-architecture. This view claims that the mind, as a whole, is dual. It claims that the mind contains a collection of systems, either S1 or S2, which are in charge of all higher cognitive functions.¹²¹

I will address issues relating to the macro-architecture of the mind next.

III. The Macro-Architecture of the Mind

A final consideration against dual-system views is that it doesn't support a convincing macro-architectural view of the mind. The dual-system view distinguishes two kinds of systems: S1 is in charge of fast, automatic, low effort, unconscious processes, while S2 underlies slow, controlled, high effort, conscious processes. This bipartite view of the mind raised, from the beginning, the problem of the coordination between the two types of systems. Some researchers thought the two types of systems to be competitive in their interaction, other thought that they were alternative or complementary.¹²² Both macro-architectures raise the following question: how

¹²¹ Of course, the plausibility of this view depends on providing a good account of the distinction between the systems.

¹²² Several proposals were offered though none successful. The two most common solutions were the competitive and the default-interventionist architecture. In the competitive architecture S1 and S2 compete for overt control of action, depending on the framework diverse aspects of the nature of the task decide which one wins (needless to say, this is not a very satisfactory

does the interaction get modulated? What (system? process?) decides which system wins in case of conflict?¹²³ And, if both types of systems are supposed to work cooperatively, in what sense is a dual-system or dual-mind view explanatory of conflicting intuitions/responses?

Indeed, the dual-system view omits a crucial component of the mind: executive processes (type-3 processes from now on). Executive processes can be divided (based on current neurophysiological evidence) in control processes (type-3a) and conflict resolution processes (type-3b).¹²⁴ These processes—mainly based in the frontal lobes¹²⁵—are the cornerstone of our cognitive capacity. They run the show behind the curtains and without them, nothing works. As I mentioned, these processes were not studied in the literature because they are not part of any dual-process view. When they were mentioned, they were confused with type-2 processes (see Evans, 2006), though they are not the same. One thing is a slow, controlled process that uses working memory to produce a result to a question (for example, ‘what is 5 times 27?’). In order to perform the steps that will lead to the answer, another type of process, a control process, needs to coordinate the action and give control to the mathematical process of working memory and other central resources. So, central executive processes are not the same as central cognitive capacity. They run the show that is the whole brain, rather than one particular spot.

view). The other proposal, the default-interventionist, states that S2 is in charge of overall control and of conflict resolution. For several reasons, this is not a satisfactory view. For a review of the problems associated to both views, see Evans (2009) and Keren and Schul (2009).

123 Let’s not forget that an important source of support for the dual-system view comes from conflicting responses (e.g., the one documented in the heuristics and biases literature).

124 Other subtypes of executive processes might be added to the list as research is carried out about the functions of frontal lobes and other structures that play a role in executive control.

125 Especially the Dorso-lateral prefrontal cortex (DLPFC) and the Anterior cingulate cortex (ACC).

This kind of processes are becoming increasingly interesting to cognitive psychology and neuroscience. In his book *The New Executive Brain*, Goldberg provides an overview of all up-to-today known functions of executive processes. It seems that there is little they don't do or control. So, I think that it is time to add type-3 processes (which themselves could be divided into control and conflict resolution processes) to the picture of the mind. Until we understand better how this processes work, we will not be able to understand how we resolve conflict and why, sometimes, we are left doing what is not in our best interests.

So, at the macro-architectural level, a dual view is not sufficient to explain the complex interaction of higher cognitive systems. Several researchers have recently retracted their strong dual-system views and acknowledged the importance of type-3 processes.¹²⁶ For example, Frankish and Evans (2009) recently pointed out, regarding future avenues of research in the dual-system literature: "In short, it is likely that future two-systems theories will need to posit multiple kinds of cognitive processing." (p. 28)

And Evans (2009) has adapted his model of heuristic and analytic processes to include type-3 processes, proposing the following model:

¹²⁶ Stanovich (2009) mentions type-3 processes but, instead of focusing on executive control, he focuses on individual differences in cognitive style, that is, on the participant's tendency to employ type-2 processes over type-1.

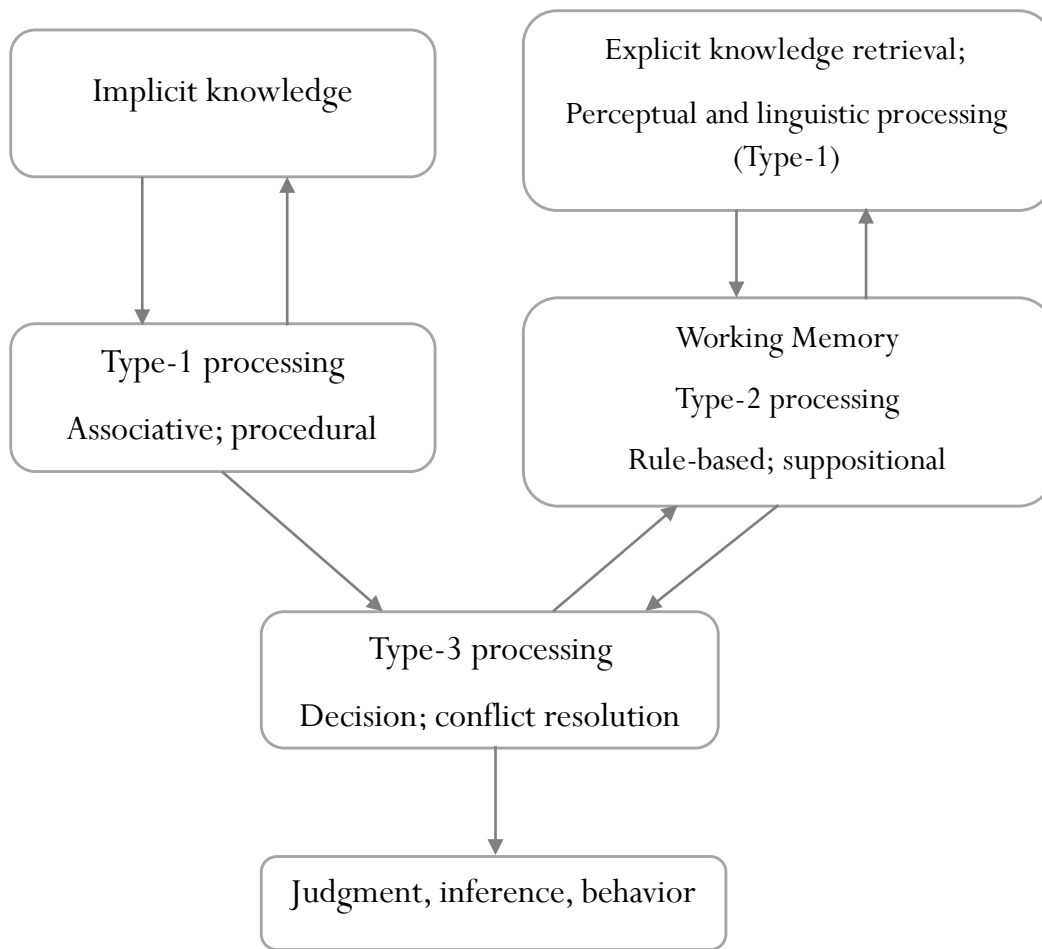


Figure 5. Evans' model of the interaction between type-1, type-2, and type-3 processing.

(Taken from Evans, 2009, p. 49)

According to Evans (2009), the rationale for introducing type-3 processes is the conscious nature of type-2 processes. While type-2 processes are by definition conscious, the processes that ultimately control behavior cannot themselves be conscious since these are the processes that coordinate between conscious type-2 processes and automatic type-1 processes. Though I do not agree with Evans' reasons for positing type-3 processes, I think his proposal is on the right track. However, I would introduce some modifications to his model. In particular, I think a correct view of the macro-architecture of the mind needs to account for the high level of interaction between both

types of processes, as well as the coordination that type-3 processes exercise among them.¹²⁷ I

propose the following instead:

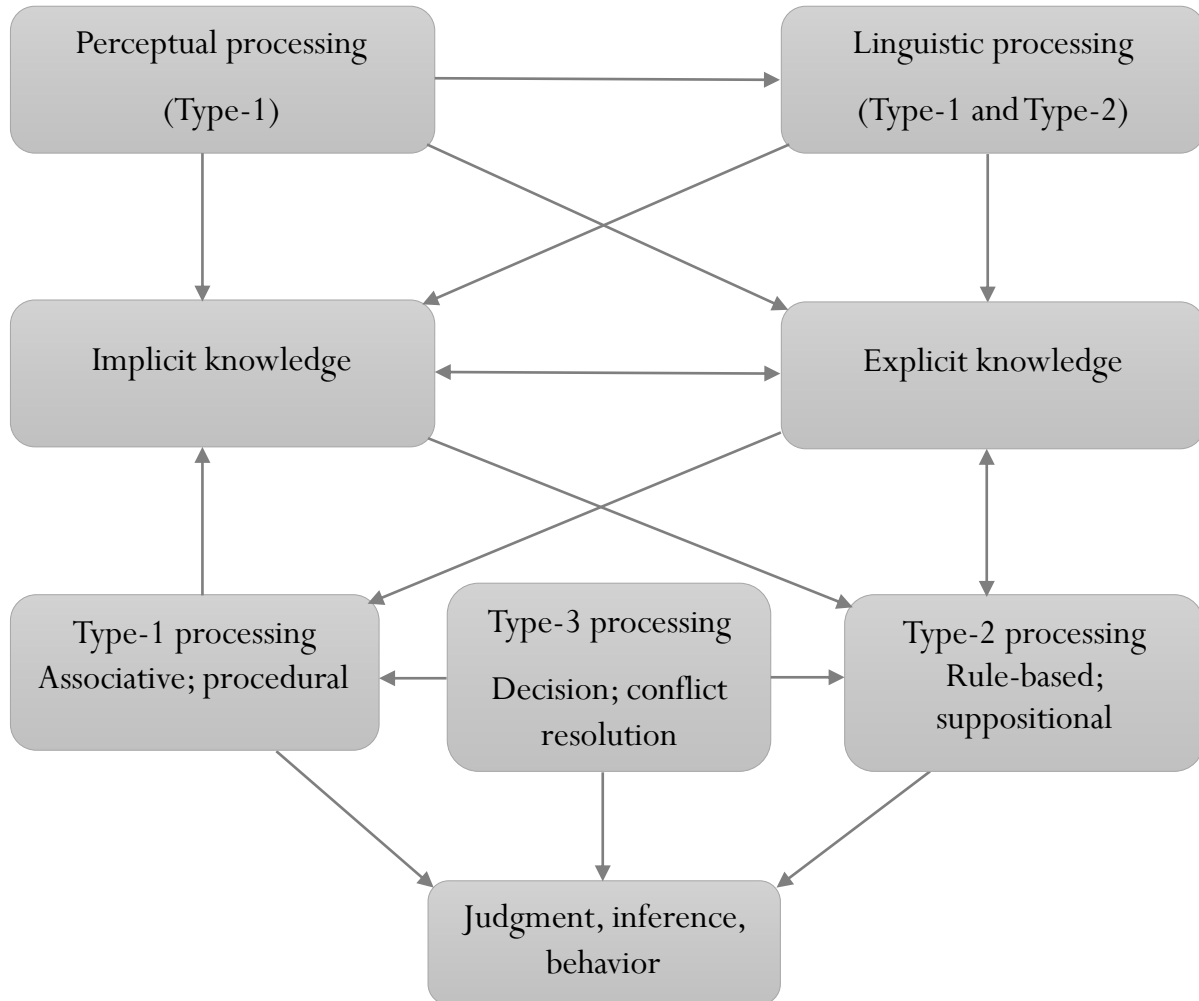


Figure 6. Proposed model of the interaction between type-1, type-2, and type-3 processes.

¹²⁷ In this I agree with Kahneman (2011) who claims that “the control of attention is shared by the two systems.” (p. 22)

Further, this model of the macro-architecture of the mind is compatible with the dual-process view proposed in Chapter Three. Some executive processes could be modelled classically, while other processes (e.g., conflict resolution) could be modeled in a connectionist way (for example, affective processes could influence conflict resolution by providing negative or positive feedback in the way of reinforcing or negative associations). Whether this is the case is a matter for further investigation.

To conclude, it seems that a crucial challenge for future research will be to develop a comprehensive, coherent view of executive control. An investigation of central executive processes will be key to provide unity and consistency to the fragmented picture of the mind that results from the dual-process literature. It will be central to achieve this goal that psychologists develop a more sophisticated view of the interaction of central executive processes with automatic processes on the one hand and other, more effortful types of thinking.

Finally, despite the complexity and fragmentation of our minds, we are rarely aware of this inner turmoil. By contrast, most of the time we live under the impression that our conscious reasoning processes actually control our inner life. A thorough account of the interaction between executive and other types of processes will be central to explaining why and how we end up with a unified picture of our inner minds, how agency is centralized into one set of actions, decisions, (conscious) beliefs and judgments, and not a multitude of them. Hence the next step in developing a complete, coherent macro-architecture of the mind is the detailed study of executive processes.

IV. The significance of the dual-system literature

Despite its many shortcomings, the research described above is important philosophically and psychologically for many reasons.

Psychological significance

I think that the main contribution of the dual literature is emphasizing the complexity of all the processes involved in higher cognition. It shows that there is an inherent plurality of processes involved in reasoning, judgment, decision making, and cognition in general. It also shows that controlled, conscious behavior is only one part of the picture. Dual-process and dual-system views highlight the importance of automatic processes at the higher cognitive level.

Some of the research (especially Kahneman's recent work) also shows that there are two types of *automatic* processes. Some automatic processes can be purposely summed up (as when a good chess player uses 'her intuition' to decide the next move). By contrast, some processes intervene automatically without any deliberate decision to engage them and, most dangerously, sometimes without awareness of the role they play (as when automatic biases overcome reflexive thought).¹²⁸

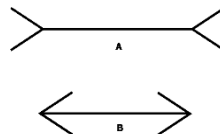
Further, it provides a framework for explaining why we end up with conflicting responses in many situations involving reasoning and problem solving. The traditional *single mind* view can explain why we might come up with different solutions at different times but it struggles to explain

¹²⁸ That is, the behavior itself might be automatic but deliberately summed up (for ex., when I decide to go to the kitchen to make a tea). By contrast, some behavior is automatic and under the level of awareness. So, against Frankish, the contrast involves automatic processes that we decide to use and those that intervene 'automatically'.

the feeling of having two simultaneous opposing solutions. A dual system-view can better explain this feeling. The explanation is similar to that offered for visual illusions¹²⁹ and similar phenomena. There are two independent systems which produce conflicting answers. Because the mind finds no way to resolve the conflict, it creates a feeling of a “dual pull” or two opposing feelings of rightness.¹³⁰

Another reason to keep the distinction is related to historical changes to psychological accounts of the nature of “higher” cognition. Traditionally, psychologists and philosophers alike considered thought as a unified capacity whose workings are transparent to us. In a way, only type-2 processes had a place in an architecture of high cognition. But again and again it has been proven that unconscious, automatic processes influence our behavior and decision making. Ideally, we will be able to provide an architecture of the mind that describes all the processes that intervene in producing overt behavior (along with their importance, their susceptibility, etc.). Once a full picture is in place, it might be possible to see the unity in it. However, for the moment it is not possible to provide a single framework for our higher cognitive processes. A dual approach is a step

129 Consider, for example, the Müller-Lyer illusion:



Though the two lines definitely look different, we can use a ruler to check that they have actually the same length. This shows that there are (at least) two faculties in the brain capable of assessing the length of the lines: the visual system, which gives rise to the (wrong) perceptual assessment, and the cognitive system, which uses knowledge about rulers and the two measurements to establish the opposite conclusion. Notice that knowing that the two lines have the same length doesn't influence at all our perception of them. This is because some aspects of visual processing are cognitively impenetrable (cannot be influenced by cognitive processes).

130 Here is an example: when participants rate Linda as more likely to be a feminist bank teller than a bank teller, they clearly find this rating more compelling than the one in the opposite direction. But when experimenters explain to the participants why this is a mistake according to probability theory, participants generally agree with the experimenters' reasoning and change their ratings. However, their previous answer somehow still remains compelling.

in the right direction. It emphasizes the complexity of the mind and the consequences of such complexity (for example, conflict, instability, unreliability).¹³¹ Finally, the dual-process literature also points out that many – if not most – processes that people engage in happen under the level of awareness but have a fundamental significance in everyday life. The vulnerability of this *divided mind* is clearer under a dual (or multi) process view than under a single mind view.

Philosophical significance

The classical view of the mind posits a unified agent with unified beliefs and desires and a single, unified cognitive capacity in charge of judgment, decision making, planning and problem solving. The classical, rational agent decides what to do on the basis of these unified capacities and only allows for conflicting judgments and behaviors as the result of the misuse of these capacities and not of the intrinsic nature of the mind.

The literature on dual-process views put this thesis to the test and showed that the mind is far from being unified. The claim that there are (at least) two ways of knowing, deciding, and reasoning challenges considerably the classical view. Further, there are no easy solutions to the conflict depicted in the dual views literature. That is, it is not viable solution to claim that one system/process is rational and that the other is systematically wrong. First, assuming there are at least two ways of deciding and two ways of desiring, how could we choose among solutions? To a certain extent, a dual (or multiple) system view gives rise to the same rationality issues as when comparing two individuals' choices. If we are really of 'two minds', then how do we decide which

¹³¹ An often mentioned example: people walk slower after reading a text about elderly people. Of course, they do not realize that they walk slower nor that they have been primed with stereotypes of old age.

mind makes the correct choices? It would be similar to deciding which one of two agents is right, when they are actually choosing differently on the basis of different probability distributions and different preferences.¹³² Secondly, even for those cases that have an independently established correct response, the literature consistently shows that no mind, system, or process has a monopoly of truth. On the contrary, the relative accuracy of each type of process depends on several factors, including the nature of the task, motivation and expertise levels, and time constraints.

The classical view faces yet another challenge. As Kahneman (2011) points out, “rational agents are assumed to make important decisions carefully, and to use all the information that is provided to them” (p. 413). Empirical results from dual-process views show that this is far from the reality of our minds: it might seem to the conscious mind that deliberate processes employ all available resources, but even rule-based, controlled reasoning is limited by those pre-attentive processes providing the informational content. Further, intuition seems random and unsupported because most of its processes are unconscious. However, type-1 processes have resources that are not available to the conscious mind. For example, in Chapter One I mentioned several learning experiments by Reber that show that people have difficulties tapping into implicitly learned rules but that they can nevertheless use that knowledge in unconscious processing. This difference in resources is sometimes beneficial (as when we use unconscious expert judgment) but other times it’s detrimental (as when heuristics bias our judgments).

132 The comparison inter minds/systems is akin to the one that we make inter participants. If two agents decide to travel by different means given their different likes and dislikes of transportation means, how could be more or less rational one system’s choice over another’s, if they really respond to different underlying preferences?

The classical view also emphasizes the superiority of abstract, logical, rule-based, and conscious reasoning (the prototypical type-2 process). However, the literature has shown that other reasoning processes, of a more intuitive kind, can sometimes be better at tracking the truth or reaching an adequate solution. For example, there are some problems that are better solved with a ‘pattern recognition’ type of approach. The research on the shortcomings of type-1 reasoning (mainly the heuristics and biases literature) emphasizes the perils of relying on intuition. However, there is also a great deal of research¹³³ that shows the shortcomings of applying slow, deliberate reasoning to problems that are not easily solvable that way (given the inherent limitations of our minds). In particular, problems that involve a large number of variables are difficult to solve in a rule-based way – precisely because the mind can only hold/consider a small number of variables at a time. In such cases, a holistic/impressionistic approach can become more useful. Similarly, putting an emphasis on knowing and analyzing why we take certain decisions can hamper the decision making process: though our minds are capable of unconsciously processing many features at the same time, they cannot ‘bring to mind’ all of them. So, by forcing ourselves to explicitly consider the reasons why we make certain decisions/have certain preferences, we constraint the problem to a size that is manageable for the conscious mind but inadequate for the characteristics of the problem.¹³⁴

133 See Gigerenzer and Murray (1987); Gigerenzer, Todd, and the ABC Research Group (1999), Wilson and Schooler (1991), Dijksterhuis (2004, Bos, Nordgren, & von Baaren, 2006).

134 When making very complex decisions most participants did better when deliberating without paying attention to the problem (Dijksterhuis et al., 2006). Similar results were obtained by Mikels, Maglio, Reed, & Kaplowitz (2011) when considering emotion-focused decision making against deliberate decision making (the task involved choosing between cars, apartments, vacations).

Indeed, the use of heuristics or intuitions can be fruitful in many situations. After all, the default setting of a finite mind is one of constraints: under constant constraints of time, resources and motivation, automatic and/or effortless processes provide an edge. Many times (if not most of the time) it is better to go with intuition, which provides a good enough solution, than be stuck trying to consciously come up with a difficult to arrive solution. That is why we value expertise so much: it is better to count on an automatic mechanism that will produce pretty good results most of the time, than to rely on a highly capable mechanism, which nevertheless needs to learn every move from scratch and consider every possibility every single time.¹³⁵

In recent decades, methodological issues relating to philosophical inquiry, particularly the origin of philosophical judgments and intuitions, have become a popular topic of discussion. The literature on dual-process and dual-system views has important contributions to make to this discussion, since the contrast between these two types of judgments is quite similar to the contrast between type-1 and type-2 processes. Further, it might be the case that philosophers' preferred line of argumentation concerns mainly the use of *analytic reasoning* (or type-2 processes). Though this type of reasoning is available to the 'ordinary man', it might not be his method of choice when solving the type of problems that philosophers reflect on. Indeed, there could be a default predisposition to use *intuition* (or type-1 processes) for many problems, while philosophers would tend to override this predisposition and engage in deliberate, conscious, analytic reflection.¹³⁶ In this

135 Of course, not all judgments by 'experts' are due to expertise. Experts are subject to the same biases as the rest of us, especially when time is constrained. This is yet another reason to distinguish between type-1 expert processes, heuristic processes, and biases.

136 For example, let's assume that there is a moral reasoning module that helps us decide how to act in a given situation. If philosophers' assessments on how to act mainly involved type-2 processes, there is a sense in which their normative models wouldn't be adequate if people were invariably predisposed to use other type of processes instead.

way, a dual-process view can provide a good explanation for some meta-philosophical problems, such as why sometimes philosophers' theories conflict so much with the ordinary man's intuitions. Empirical research on the nature, interaction, and conflict among reasoning processes and systems can help clarify some debates in philosophy that so far have been decided solely on the basis of *armchair* considerations.¹³⁷ Finally, though support for a dual view might not change philosophers' beliefs about the inconsistency of such claims, it might change their view on which truth conditions to assign to those claims. Relatedly, it might help decide what kind of circumstances grant justification for making such claims.

Finally, there are further rationality issues that need to be considered in the context of dual-process views. I would like to raise a concern inspired by Antonio Damasio's work¹³⁸ on the relation between brain damage to the frontal cortices and pro-social behavior. Close examination of patients with this type of damage led Damasio to hypothesize that although these patients have all their general cognitive abilities intact (such as general intelligence, memory, logical reasoning abilities), they have nevertheless lost the capacity to act in their best interest. Somehow, the brain damage

137 Here is an example from epistemology. Lately some philosophers have considered the following problem: while in everyday circumstances we are certain of the truth of knowledge claims like 'I know I will fly to Europe next week', 'I know that I will attend a conference next semester', 'I know that my car is parked outside my apartment', there are related claims whose truth we claim not to know, such as 'I will not have a fatal car accident on Tuesday', or 'My car hasn't been just stolen from the parking lot'. The problem seems to be that our knowledge of the first type of claims depends on our knowledge of the second type. So, how can we claim to know the first ones and not the second ones? I will not go into details about the various accounts that have been proposed to explain this phenomenon. I just want to point out that a dual theory of cognition provides a good explanation of why this phenomenon arises in the first place: everyday knowledge claims of the type described above are the result of type-1 reasoning. Such claims are fast, automatic, and intuitive. It is just as intuitive to believe that you know that your car is parked outside your apartment if you just left it there, as it is to concede that you don't know that a burglar hasn't just stolen it, if someone points out this possibility to you. This might be due to a domain encapsulation issue, for example. Maybe, when making everyday knowledge claims, we don't consider related circumstances unless someone points them out to us. On the other hand, philosophers' perspective on the issue clearly seems to be the result of type-2 reasoning: a controlled, slow, rule-based account of such inconsistencies.

138 Damasio, Tranel, & Damasio (1990); Damasio (1994); Anderson, Bechara, Damasio, Tranel, & Damasio (1999); Bechara, Damasio, & Damasio (2000).

causes them to lose the capacity to associate affective responses and emotions with the anticipated consequences of their actions. As a consequence, it is difficult for them to avoid behaviors that will cause them great losses or damage. Given these results, it is not farfetched to assume that people might be equipped, on the one hand, with reasoning processes which update and calculate probabilities according to decision theory's norms. On the other hand, people also seem to have a different set of processes (e.g., affective) that motivates the agent to act according to those calculations. Under this hypothesis, type-2 processes do all the calculations and establish all the appropriate expected utilities for different courses of action, but it is the affective system the one that motivates the agent to act accordingly. But, what if this motivation came in degrees? Such findings would change the way we assess people's rationality. And what about people who had the motivation and their cognitive faculties intact, but lacked the resources to make a connection between the two? In a sense, this seems to be the actual problem with the brain damaged patients studied by Damasio. They do have the appropriate emotional responses (these were measured in gambling tasks), they just can't associate them with the anticipated consequences of their actions. In any case, all I want to point out is that empirical findings from the two systems view can influence our positions concerning rationality issues.

These are very general points on how dual-process views relate to issues in philosophy. In Chapter Five, I will provide a more detailed example of how this view relates to a specific philosophical discussion. I will describe some recent empirical findings concerning moral decision making and its relation to dual-system views of moral judgment. After I summarize the empirical

results, I will address issues relating to the methodological design of the experiment and discuss the suggested implications that these results have for philosophical debates about morality.

Chapter Five

In Chapter One I presented the main psychological findings that support a dual-process view. According to this view, there are two types of processes underlying higher cognition—including reasoning, judgment, decision making, and learning. Type-1 processes are (generally) fast, automatic, low-effort, and high capacity. Type-2 processes are slow, controlled, high effort, and low-capacity. In Chapter Three I showed that the differences between the two types of processes can be due to different underlying cognitive architectures. Indeed, many type-1 processes run in parallel and process information on the basis of learned associations and regularities, as a consequence they usually occur under the level of consciousness. By contrast, type-2 processes involve working memory, run serially and according to consciously stored rules. In Chapter Four I argued that the dual-process view is not sufficiently robust to support a dual-system view. There are three main reasons. First, the distinction at the micro-architectural level only explains the clustering of properties to a certain degree. Not all type-1 processes can be modeled by a connectionist architecture (e.g., language processes that are fast, effortless and automatic but not associative), and similar remarks can be made for type-2 processes. Indeed, the goal of Chapter Three was to show why processes with different micro-architectures exhibit the clustering described in the literature, not to assume that all instances of clustering are due to a micro-architectural distinction. Secondly, there are independent reasons for separating the properties of a system from the properties of their processes and there are also reasons for preferring an architecture with hybrid systems, rather than solely classical or connectionist ones. Finally, there are considerations relating to the macro-architecture of the mind that support a plural rather than a dual-system view. In particular, there are

strong reasons to introduce a third type of process related to executive control and conflict resolution.

Against this background, I will consider a very influential set of experiments by Greene et al. (2001, 2004) on the psychological mechanisms underlying moral decision making and judgment. According to the authors, there is not one moral decision making system but *two*.¹³⁹ Greene and colleagues claim that one of the systems relies, for the most part, on the emotional response generated by the input. By contrast, the other system uses “higher” cognitive processes. Further, the researchers claim that the first system is generally responsible for deontological judgments and the second for consequentialist ones. They also claim that, when in conflict, these two mechanisms compete to come up with a response. In a related paper, Greene (2003) suggests a number of implications that these results have for moral philosophy (in particular, regarding the suitability of the two normative theories, consequentialism and deontology).

The aims of this chapter are several. First, I will briefly describe the experiments and summarize the empirical results. I will then criticize some of the authors’ choices of methodological design and discuss the suggested implications that these results have for philosophical debates about morality. I will show that Greene oversimplifies the type of processing involved in emotion-based

139 Greene (his website): “The goal of my research is to understand how moral judgments are shaped by automatic processes (such as emotional “gut reactions”) and controlled cognitive processes (such as reasoning and self-control).”

“Our research indicates that there is no dedicated “moral sense” or “moral faculty.” Instead, moral judgment arises from interactions among dissociable cognitive systems, many of which—perhaps all of which—are not specifically dedicated to moral judgment.”

“I have proposed a “dual-process” theory of moral judgment according to which characteristically deontological moral judgments (judgments associated with concerns for “rights” and “duties”) are driven by automatic emotional responses, while characteristically utilitarian or consequentialist moral judgments (judgments aimed at promoting the “greater good”) are driven by more controlled cognitive processes. If I’m right, the tension between deontological and consequentialist moral philosophies reflects an underlying tension between dissociable systems in the brain.”

responses. For example, just because a process is fast and automatic it does not mean that it has an underlying simple and/or incorrect algorithm (think, for example, of grammatical assessments which are fast and automatic but highly reliable and probably very complex in structure). I will also explain why in many of his remarks, Greene is presupposing a consequentialist view of morality and therefore begging the question about which moral mechanism is more reliable¹⁴⁰ (assuming there are two such mechanisms and not a plurality of processes). Though Greene is cautious not to claim that the data shows the superiority of consequentialism, he does claim that they lead him to ‘shift the balance’ and take a more consequentialist approach. I will argue that this shift, though acceptable, it is not supported by the empirical results.

I. Moral Psychology and Moral Philosophy

The significance of facts about our psychology for ethical theorizing is contentious.¹⁴¹

Despite these disagreements, we have in recent years seen an increased predisposition among moral philosophers to draw upon moral psychological facts in defending their ethical views. Empirical attentiveness is not new, but two aspects of this approach are new. Firstly, although for a long time ethicists have made claims about common sense intuitions, beliefs and judgments, only recently

¹⁴⁰ I will talk of ‘reliability’ and not correctness/accuracy to avoid contentious meta-ethical questions. My reasons for this will become apparent.

¹⁴¹ Gil Harman’s (1985) distinction between naturalistic and autonomous approaches to ethics is still apt. Naturalists prioritize the reduction of ethics to natural phenomena, usually via psychology. Accordingly they think that psychology constrains ethics, and hence is prior in the order of theorizing. Autonomists think that ethical theorizing involves distinctive methods and standards, which constrain theories of moral psychology, and hence ethics is prior in the order of theorizing. Autonomists object to the naturalists’ attempts to get an ethical ‘ought’ from a psychological ‘is’.

have theorists actually gone out and tested these claims (and with some surprising results!).

Secondly, the weight given to psychological facts in some contemporary ethics is unprecedented.

This methodological turn has stimulated the development of psychological experiments in many areas, including people's reward and punishment attitudes, the influence of particular situations in people's moral attitudes, people's responsibility judgments, and many other areas of moral psychology (Doris, 2002; Knobe, 2006; Sober & Wilson, 1999).

There is also a growing interest in neuro-ethics – that is, in measuring the brain activity of participants while they answer moral questions or make moral decisions. In particular, one line of research uses functional magnetic resonance imaging (fMRI) to better understand which areas become more active during moral reasoning (Greene et al., 2001, 2004; Haidt, Koller, & Dias, 1993; Moll, de Oliveira-Souza, Bramati, & Grafman, 2002; Moll et al., 2005; Moll, de Oliveira-Souza, & Zahn, 2008; Schaich-Borg, Lieberman, & Kiehl, 2006).

II. The Philosophical Background

Consequentialism and deontology are two of the dominant contemporary views in moral philosophy. Consequentialism is the view that one should act so as to maximize value of some kind. It is concerned with outcomes, and not specifically with types of actions. Deontology is the view that certain rules or intentions or characters or action-types are the ones we ought to follow/inculcate/perform, irrespective of consequences.

A series of papers by Philippa Foot and Judith Jarvis Thomson (among others) opened up a debate about moral dilemmas and the role of our intuitions in supporting deontology as a consistent moral view. The first paper presented a now famous scenario known as ‘the trolley problem’.

You are at the wheel of a runaway trolley quickly approaching a fork in the tracks.

On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing, the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. Is it morally permissible for you to hit the switch in order to avoid the deaths of the five workmen? (Foot, 1978)

As it is stated, it seems morally permissible to press the switch and kill¹⁴² one person in order to save five. On the other hand, the same judgment doesn’t seem intuitively correct in the “footbridge case”:

A runaway trolley is heading down the tracks towards a group of five railway workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks in between the approaching trolley and the five workmen. Next to you on this footbridge there is a stranger who happens to be very large. The only way to avoid the deaths of these five workmen is to push this stranger

142 I’m using the expression ‘to kill’ instead of ‘to let die’ (or any other similar one) for describing these scenarios because it is not my intention to get into a discussion about the appropriateness of the distinction and other related issues.

off the bridge and onto the tracks below where his large body will stop the trolley.

The stranger will die if you do this, but the five workmen will be saved. Is it morally appropriate for you to push this stranger off the bridge and onto the tracks in order to avoid the deaths of the five workmen? (Thomson, 1976)

Many philosophers agree that while it is permissible to press the switch it is not permissible to throw the man over the bridge. This split in judgment is philosophically hard to justify because, at first sight, there doesn't seem to be a morally significant difference between the two cases. Hence it is possible that what is producing a difference in judgment is a morally irrelevant factor.

In any case, whether morally irrelevant or not, there are clearly some factors that are producing the difference in judgment. But which are they? This is a question that can be studied empirically in at least two different ways. One option is to create many scenarios related to the ones described above and to compare people's responses to each of them. By analyzing these responses, we can find out which changes in the scenario create a difference in judgment. We can also investigate which brain areas are involved when deciding about the different cases and try to extrapolate from this information the factors that cause the difference in moral judgment. Greene and colleagues have taken both paths. On the one hand, the authors have gathered extensive data on people's judgments in a broad set of scenarios, all variations of the trolley problem (Greene et al., 2008, supplementary materials). They have also developed a series of experiments – which I will discuss in detail in the next section—that use fMRI to record participants' brain activity while reading, reasoning, and making a decision about moral dilemmas (Greene et al., 2001, 2004).

Before I describe the fMRI experiments in detail, I would like to clarify some notions that play a key role in Greene's interpretation of the experimental results. In the paper "The Secret Joke of Kant's Soul" Greene claims that consequentialism and deontology are not just two different moral views, but rather two 'psychological natural kinds'. That is, according to Greene, there are two different mental mechanisms that can be used for moral reasoning and judgment. One of them roughly produces our characteristically deontological judgments (e.g., judgments about right actions), while the other gives rise to our characteristically consequentialist judgments (e.g., judgments about good outcomes). In his own words, deontology and consequentialism "are philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking that have been part of the human repertoire for years." (Greene, 2007, pp. 37-38)

Because he is interested in the psychological natural kinds (rather than the philosophical theories) his definitions of deontological and consequentialist judgments are based on the functional role that these mechanisms play in our mental life. According to Greene, a 'characteristically' consequentialist judgment is one that considers only the consequences of an action, disregarding the means used to achieve those consequences. By contrast, a deontological judgment is one that takes into account characteristically deontological notions, such as 'duty' and 'rights' (ibid., p. 38). An example of a characteristically consequentialist judgment would be 'It's morally permissible to save five lives at the expense of one', while the judgment 'no matter what the consequences are, it is not morally permissible to kill a person in order to save more lives' is typically deontological.¹⁴³

143 These two judgments are respectively people's most common responses to the "trolley problem" and the "footbridge dilemma".

This distinction is crucial to Greene's work because his experiments focus on the differences in brain activity when reaching characteristically consequentialist and characteristically deontological judgments. It is also important for him to establish how to interpret higher activation in specific brain areas. In particular, areas previously associated with emotions, reasoning, planning, and decision making are of special interest since Greene wants to test the claim that deontological judgments are driven by emotions while consequentialist judgments are the product of 'genuine' moral reasoning.

So how does he distinguish 'emotion' from 'cognition'?¹⁴⁴ Clearly, there are important differences between the two. Still, it is quite difficult to define what differentiates them and what they have in common. Greene claims that cognition – in the broad sense of 'information processing' – includes emotion. But there is a notion of "cognition" which is defined in contrast to emotion. The key difference is that "cognitive" representations¹⁴⁵ lack the affective valence that emotional representations have.¹⁴⁶ They are inherently neutral and therefore more apt for flexible behavior:

Highly flexible behavior requires "cognitive" representations that can be easily mixed around and recombined as situational demands vary, and without pulling the agent in sixteen different behavioral directions at once. [...] Stereotyped behavior, in contrast, doesn't require this sort of flexibility and therefore doesn't

144 A category of which 'moral reasoning' would be an instance.

145 From now on, I will follow Greene's convention of using quotation marks when using the term in this restricted sense, while dropping them when referring to general information processing.

146 It is highly controversial whether such a thing exists. Just because a representation is formal or symbolic it doesn't mean that it is not included in a network of associations. So, for example, any use of the 'mother' representation might have an emotional valence just in virtue of being easily associated with several other representations, as well as affective content.

require “cognitive” representations, at least not to the same extent. (Greene, 2007, p. 40)

Based on previous experimental results, activation in the dorsolateral surfaces of the prefrontal cortex and parietal lobes is interpreted by Greene and colleagues as engagement in reasoning, planning, and other “cognitive” processes.¹⁴⁷ By contrast, activation in the amygdala and in the medial surfaces of the frontal and parietal lobes will be associated with emotion (ibid., pp. 40-41).

III. The Experiments

In the first study, Greene et al. (2001) compared brain activity during judgment of scenarios that involved: (i) personal moral dilemmas, (ii) non-personal moral dilemmas, (iii) non-moral dilemmas. The category of ‘personal’ moral dilemmas was specially developed by the researchers for these studies. The authors hypothesized that certain moral dilemmas had a special status because of the emotional responses they generated. They describe such cases as involving an action that (a) leads to serious bodily harm, (b) is performed by a particular person or group and is directed to a particular individual or group, and (c) where the harm is not the result of deflecting an existing threat but involves clear agency on the actor’s part.¹⁴⁸ An example of such a case would be the

¹⁴⁷ This is not quite right since the contrast proposed by the authors usually matches the type-1/type-2 distinction. However, the areas associated to “cognitive” (type-2) processes are areas involved in executive control (type-3 processes). Confusing type-2 processes with control processes is quite common in the literature.

¹⁴⁸ About this distinction, Greene et al. (2004) say: “The rationale for distinguishing between personal and impersonal moral violations/judgments is in part evolutionary. Evidence from observations of great apes suggests that our common ancestors lived intensely social lives guided by emotions such as empathy, anger, gratitude, jealousy, joy, love, and a sense of fairness (de Waal, 1996), and all of this in the apparent absence of moral reasoning. [...] Thus, from an evolutionary point of view, it would be strange if human behavior were not driven in part by domain-specific social-emotional dispositions. At the same time, however,

footbridge dilemma, in which the agent can only stop the trolley by pushing the man over the bridge onto the tracks.

One of the aims of this first study was to show that the strong emotional responses that participants feel when they are considering the personal moral scenarios play a very important role in the decision making. The following findings seem to support this claim:

- Personal moral dilemmas in comparison with the other two conditions yielded increased activity in areas associated with emotion: medial frontal gyrus (medial BA 9 and 10), posterior cingulate gyrus (BA 31), angular gyrus, bilateral (BA 39), and a decrease in activity in areas associated with working memory: middle frontal gyrus, right (BA 46) parietal lobe, bilateral (BA 7/40).
- Non-personal moral dilemmas when compared with non-moral dilemmas yielded only an increase in activity in medial frontal gyrus (medial BA 9 and 10) and posterior cingulate gyrus (BA 31).

The researchers also noticed that there was an increase in reaction times when participants made a consequentialist judgment in personal moral dilemmas. According to them, this suggests that the strong emotional response against judging the action appropriate has to be overridden by “cognitive” areas and that this causes the decision process to take longer than when participants rely on their emotional reaction.

humans appear to possess a domain-general capacity for sophisticated abstract reasoning, and it would be surprising as well if this capacity played no role in human moral judgment.” (p. 389)

This hypothesis was tested by Greene et al. (2004). The purpose of the study was to show that there are two alternative paths for moral decision making and judgment. One involves emotional processes and is fast and automatic. The other takes into account the emotional response but it also involves “cognitive” processes. When there is a conflict between the two mechanisms, exercise of cognitive control is needed to override the automatic mechanism.

For the type of dilemmas that involved a high emotional component, the authors hypothesized that activation in brain areas associated with abstract reasoning and cognitive control would correlate with a characteristically consequentialist judgment while increased activation in emotion areas (but not “cognitive” areas) would correlate with deontological responses.¹⁴⁹ They tested the following two hypotheses (they found the expected brain activity in both cases):

1. Increased reaction times in personal moral dilemmas are due to conflict between a strong emotional response and a more abstract ‘consequentialist’ reasoning. This conflict is realized through an increased activity in the ACC (anterior cingulate cortex) and in the DLPFC (dorso-lateral prefrontal cortex). The first area has been related to the presence of cognitive conflict, while the second is associated with abstract reasoning processes.¹⁵⁰
2. In considering these dilemmas, cognitive control is exercised against the emotional response and in favor of a consequentialist response. Therefore, an increased presence of activity in

149 As I will elaborate in the critical part of this chapter, this expectation can be a particular result of the experimental design and not an expectation based on the nature of deontological/consequentialist moral judgment. The stimuli were produced in such a way that the ‘emotional’ response coincided with the deontological response. Whether this is always the case because the deontological judgment is the emotional response, or whether it is a particular result of the experimental design is open to discussion. I will say more about this in the next section.

150 It is also related to general cognitive control (i.e., the coordination of different cognitive processes). This is the area underlying type-3a processes, as described in the previous chapter (type-3b processes are associated with ACC, the other area mentioned). For a thorough analysis of executive processes and the frontal lobes, see Goldberg (2009).

DLPFC is observed in the trials where the participants considered a personal moral violation to be appropriate rather than in the ones that they considered it inappropriate.

In summary, while the purpose of the first study was to test the hypothesis that different types of moral dilemmas produce different patterns of neural activity, this second study was expected to show that differences in the pattern of neural activity (for the same class of moral dilemmas) correlate with differences in moral judgment. According to the authors, the experimental results in the second study support their hypothesis.

- The comparison between difficult personal moral dilemmas and baseline¹⁵¹ yielded increased activity in areas associated with emotion: medial prefrontal cortex (BA 9/10), posterior cingulate/precuneus (BA 31/7), bilateral superior temporal sulcus (STS)/inferior parietal lobe (BA 39).
- The comparison between difficult versus easy personal moral dilemmas showed increased activity bilaterally in: anterior DLPFC (BA 10/46), inferior parietal lobes (BA 40/39), and increased activity in: ACC (BA 39), and posterior cingulate cortex (BA 23/31).
- When they compared consequentialist moral judgments to non-consequentialist judgments, they found an increase in activity in anterior DLPFC (bilaterally) (BA 10/46); right inferior parietal lobe (BA 40), a more anterior region of the posterior cingulate (BA 23/31). And, in a whole brain analysis, increased activity was found in right anterior DLPFC (BA 10), left inferior parietal lobe (BA 40), posterior cingulate (BA 23/31), right superior temporal

¹⁵¹ Baseline activation is usually recorded at the beginning of the experiment, while the participant lies in the fMRI machine waiting to perform the task.

gyrus (BA 22/42), right middle temporal gyrus (BA 21), and left inferior temporal gyrus (BA 19).

In sum, Greene and colleagues claim to have shown that two distinct mechanisms (involving different areas of the brain) are usually engaged in moral decision making and that they compete to come up with a solution. According to Greene (2003), the persistence of the philosophers' debate between consequentialist and deontological perspectives is due to the structure of the human brain: the social-emotional responses inherited from our ancestors incline us to a deontological perspective, while the more recently developed structures for abstract thinking and higher cognition support a consequentialist response.

III. Criticisms

First, I will address the issue of the philosophical implications of the experimental results. Greene never states directly that we should draw philosophical conclusions from the results (he is very cautious about making general or unqualified claims). But there is an array of claims in Greene (2003, 2007, et al. 2009) which point all in the same direction: that consequentialism provides a better standard for moral decision making than other views (including deontology) because intervention by areas related to higher cognition mainly leads to consequentialist judgments. In his words:

For me at least, understanding the source of my moral intuitions shifts the balance, in this case as well as in other cases, in a more Singerian, consequentialist direction. As a result of understanding the psychological facts, I am less complacent about my all-too-human

tendency to ignore distant suffering. Likewise, when I understand the roots of my retributive impulses, I am less likely to afford them moral authority. (Greene, 2007, p. 76)

As I said above, Greene is cautious not to suggest that we should all do so, or that this is the most ‘rational’ attitude. However, the implication is clear. So next I will evaluate the claims that lead him to ‘shift the balance’ and take a more consequentialist approach. I will argue that this change, though acceptable, it is neither required nor supported by the data.¹⁵² Greene will have to keep looking if he wants to convince us of the superiority of the consequentialist approach, even if just as a ‘guiding’ theory.

Here are some of the claims that Greene uses as a basis for *shifting the balance* towards consequentialism:

1. Emotion (the source of our deontological judgments) is an automatic response to the moral stimulus while “cognition” (the source of our consequentialist judgments) is flexible and affectively neutral and therefore can be used to model a variety of situations.

Though Greene doesn’t explicitly say so, he seems to be implying that “cognition” is more adequate for moral reasoning and decision making because it is ‘genuine’ moral reasoning. By contrast, judgments based on emotions are automatic and inflexible, so we’re left to assume that this makes them inadequate or unreliable. But this point is clearly unjustified since there is a wide range of mental processes that are automatic, but not for that reason less complex or accurate. For example, judgments of grammaticality are automatic but they are also extremely reliable and

¹⁵² And they also contradict many theoretical assumptions.

certainly quite complex (so complex that we still haven't completely figured out how they are carried out).

It could well be the case that those 'deontological' judgments that seem to be automatically based on the affective valence of the stimulus are actually the result of some complex mechanism, similar to a 'moral grammar'.¹⁵³ In fact, a more complex view on emotional processes would be more compatible with the results described in the dual system literature. Many of the claims that Greene makes are in accord with this theory. He assigns features to these emotional responses that are prototypical of type-1 processing (intuitive, fast, effortless, and consciously opaque). He also describes "cognitive" processes as evolutionarily recent mechanisms that are part of a domain-general system for abstract reasoning. These distinctions are quite consistent with the picture proposed by dual theorists and it is clear that Greene is using this theoretical background to assess the empirical results.

Nevertheless, unlike other authors' work in the area, Greene tends to considerably diminish the value and complexity of the first type of process and of its role in moral reasoning. In fact, he regards emotion-based judgments as automatic reactions that do not constitute 'genuine' moral reasoning (Greene, 2007, p. 36). So, one first problem with Greene's psychological picture is that it puts too much emphasis on a distinction between "cognitive" and emotional areas that is not well supported empirically (or theoretically).

Also, I would like to observe that one should always be cautious about categorizing brain areas as simply "cognitive" or emotional, even if the distinction reflects previous experimental

¹⁵³ This view is defended by Mikhail (2008) in his response to Greene (2007).

results (that is, the “cognitive” areas have been recorded as particularly active during “cognitive” tasks and the emotional areas during tasks involving emotions). There is still much research to be done before it is clearly established what the complete list of functions of these areas is.

2. Deontological judgments are directly based on the emotional response to the stimulus.¹⁵⁴

In Greene’s experiments, deontological judgments are mostly correlated with high activation of emotional areas. But it is an oversimplification to assume that this is always the case. For example, Bolender (2001) makes the exact opposite claim. According to him, many deontological judgments run against our emotional reactions and require a great deal of cognitive control to be accepted. Certainly, many deontological precepts are hard to accept at the emotional level. For example, Kant’s deontological views included claims such as ‘one ought to turn one’s mother to the police if she has committed a crime (no matter how harmless)’. Clearly, this is a case where one would have to overcome a strong emotional aversion in order to accept the claim. The fact that deontological judgments occur after increased activation in emotional areas (and not in the ACC) in Greene’s experiments might just be a result of the experimental design.

3. The exercise of cognitive control over the automatic emotional response is correlated with consequentialist judgments.

Similarly to the point above, I think that this might be a specific consequence of the experimental design. In particular, the moral dilemmas used in the experiment seem to be constructed in such a way that whenever a participant judges the action as morally permissible, she

¹⁵⁴ Consequentialist judgments require the exercise of cognitive control if they prescribe an action that has a negative affective valence.

is making a consequentialist judgment. In contrast, the non-consequentialist response is always associated with rejecting the action (taking it as morally impermissible). Therefore, the data is based on scenarios where the exercise of cognitive control is required for choosing the consequentialist answer, not the deontological one.

So, one alternative explanation is the following: perhaps in general it is cognitively more demanding for people to judge a controversial action as morally permissible (as in the example above about turning your mother to the police). It could well be the case that it is cognitively easier to condemn the action than to accept it. For example, people could have some kind of internalized rule that says that if an action is of a dubious moral status, then it can't be accepted until there is a firm enough reason to allow it. Before concluding that consequentialist judgments are generally the result of the exercise of cognitive control over an emotional response, one would have to conduct an experiment where the dilemmas are such that allowing the action requires taking the non-consequentialist perspective. One could then confirm whether these responses are just based on emotions or whether they also require the exercise of cognitive control.

For example, it seems to me that in the following scenario, people would find it easier to condemn the action from a consequentialist point of view, but easier to allow it from an emotional point of view:

It is wartime and you and your two young children, ages three and five, are living in a territory that has been occupied by the enemy. At the enemy's headquarters is a doctor who performs humiliating and degrading experiments on children that most of the times also lead to death. He intends to perform experiments on one of your

children, but he will allow you to choose which of your children will be experimented upon. You have twenty-four hours to bring one of your children to his laboratory. If you refuse to bring one of your children to his laboratory he will find them both and experiment on both of them. You realize that the only way of saving one of your children from a horrible death is to kill them both. Is it appropriate for you to kill both of your children in order to save one of them from a horrible death?¹⁵⁵

If someone decided that the action is morally permissible, they would be taking a non-consequentialist stance.¹⁵⁶ It also seems that an affirmative answer would require a lot of cognitive control (given the amount of conflict that such a dilemma produces in the reader). But, if that were the case, if non-consequentialist judgments also involved “cognition” and the exercise of cognitive control, then Greene’s analysis of the data would mostly reflect his experimental methodology rather than our moral psychology.

4. Nature has provided us with moral emotions because they made us fitter in the environment in which we evolved. However, in today’s world many of the factors to which nature made us sensitive and which condition our moral judgments are morally irrelevant.

The claim that some of these factors are morally irrelevant begs the question. For example, it is not a given that geographical distance is morally irrelevant. It might seem irrational to maintain that I am morally obliged to save a drowning child in front of me but I am not morally obliged to

155 This example has been adapted from one of Greene’s examples (itself inspired in the book *Sophie’s Choice*).

156 Assuming that saving the life of one child outweighs the disutility caused by the pain and humiliation of the other child.

save a drowning child far away from me. However, whether this is in fact irrational depends on our choice of moral theory. (There are some data that plausibly must be accommodated by any moral theory, e.g., that egregious torture is wrong; but the irrelevance of geography is not such a datum. Careful writers on ethics such as David Lewis (1986) and Frances Kamm (2000) have argued that geography is ethically relevant.) Situational factors may be relevant for the generation of all reasons, including moral ones.) Even though many people have an intuition that geography is morally irrelevant, we have other intuitions too, such as our intuition that we ought to help the near child but it is not the case that we ought to help the distant child. Greene needs to do more to show that the former intuition has better credentials than the latter intuition, just in virtue of the brain areas that are engaged in supporting that intuition.

5. Wherever nature has been nearsighted and provided us with moral intuitions that are based on morally irrelevant factors, we can use “cognition” to compensate for this.

The problem with this claim is quite obvious: there is no reason to believe that “cognitive” processes¹⁵⁷ will provide better guidance on what to do when we have reasons to doubt our evolutionary instincts. For example, let’s suppose that distance is in fact a morally irrelevant factor. It remains an open question whether we should endorse the emotional response in the case of the drowning child and apply it to the case of the child dying of AIDS in Africa, or whether we should reverse the first judgment given our attitude to the second case. Greene (2007) assumes that we should take the first path and endorse “a more Singerian, consequentialist direction” (p. 76). I

157 For obvious reasons, I do not agree with Greene’s terminology. I do not think there are “cognitive” processes as he describes them, nor that there is a partition between these and emotional processes only. However, for argumentative purposes, let’s assume that he is talking about prototypical type-2 processes (i.e., slow, deliberate, conscious, effortful, etc.).

suppose many other philosophers would do the same. But where does this certainty come from? Couldn't we use "cognitive" processes to determine that in *neither* case we are obliged to save the child? In fact, isn't this type of "cognitive" reasoning the cause of so many disagreements in moral philosophy (e.g., the failure of philosophical ethical theories to 'converge' (Williams, 2011))? Aren't these "cognitive" processes responsible for the claim that there are actually no moral facts?¹⁵⁸

I agree that being sensitive to morally irrelevant factors would be a regrettable flaw of moral emotions. But this does not imply that in those cases we are better off using the "cognitive" mechanism.¹⁵⁹ This approach could lead us, as a matter of fact, to produce more damage than good.¹⁶⁰ Again, given that Greene believes that the "cognitive" mechanism generally leads us to consequentialist judgments, he assumes that when conflict arises the "cognitive" response should be endorsed. But this presupposes that consequentialism is the most appropriate stance, since there is no reason to assume that the "cognitive" mechanism is more reliable than the emotional one.

6. Just because moral emotions made us fitter (in the environment in which we evolved) it doesn't mean that we need to pay special attention to them when making moral judgments.

I generally agree with this claim, but I would add that the same can be said of "cognitive"¹⁶¹ processes. They are the most recent product of our evolutionary development –and what distinguishes us from other primates–, still they might be quite unreliable at making moral

158 Something Greene cannot deny since he's a moral antirealist himself.

159 Assuming there is one such higher cognitive mechanism. But, as shown in this dissertation, cognition is much more complicated than just a set of two systems. Further, the distinction Greene proposes is even more restricted: type-2 processes versus emotional processes, which are under any interpretation only a subset of type-1 processes.

160 After all, many moral atrocities have been committed under the guidance of well-thought out and reasonable arguments.

161 As above, I will understand this terminology to refer to characteristically type-2 processes.

judgments. We can't just assume that a more recently developed system is by default better or more accurate in its outcomes. One could easily assume the opposite and hypothesize that an evolutionary older system is probably more developed and that its processes are more stable and less susceptible to failure. Of course, it is generally agreed that later developed areas and systems have provided us with better tools to solve complex reasoning tasks. Yet it is still unconfirmed that for **all** reasoning and decision making, evolutionary recent systems are by default better than older ones.

Let's consider, for example, the following situation:

A man is lost in a forest in the middle of the night and he needs to decide which of two paths he should take in order to get back into town. He has a map and a compass. He looks at both paths in the map and he realizes that path A is shorter and somehow more direct than path B. Nevertheless, when he is about to take path A, he has the intuition that it might be dangerous or in some way unreliable, making it worse than path B. So, he decides to take path B and he arrives safely into town some hours later.

In this case, we can see why the intuitive answer might actually be a better solution for the problem. Though the choice is based in just an intuition and not a conscious, deliberate reasoning process, this could be the type of situation where it is better to rely on intuitions generated by a long evolutionary adaptation process than on highly complex but very unreliable (given the situation) reasoning strategies. A very well planned strategy, though reasonable, could end up being a bad response to a situation that has so many unknown factors.

Similarly, many social situations are cases where intuitive processing can be better at guiding behavior. For example, until we know someone well, we need to think much about the intentions, attitudes, beliefs and desires of that person in order to fully understand her actions. As our knowledge of the person increases, all these reasoning processes become more intuitive, fast, and automatic, until we found that we have a clear immediate understanding of the person's attitudes without needing to consciously reflect on it. Yet another example from social cognition is the ability to understand a joke or some special kind of humor. When we first watch a T.V. show or become new friends with another person, we might not understand their humor and idiosyncrasies (indeed, we might need to think and analyze their jokes in order to understand them and assess whether they are funny). But as time passes and we become more familiar with them, our reactions become automatic and judgments about them become intuitive.¹⁶²

In all these situations, it seems that the activity is first performed with the help of arduous, slow, and deliberate cognitive processes. As skill is developed and proficiency acquired, the processes become automatic, fast, and unconscious. This supports the claim that for certain types of reasoning and decision making type-1 processes are more appropriate, since they show the acquisition of expertise. It is also possible that some cognitive processes are always type-1. This seems to be particularly the case for many processes involved in social cognition. And, given the close connection between moral and social issues, it wouldn't be surprising if type-1 processes were better at solving moral dilemmas than type-2 processes.¹⁶³

162 This type of development usually occurs during childhood and early adulthood. Hence, we usually don't remember that making these judgments was effortful at one time. However, if we move to another country or somehow change social contexts, the need to develop these skills reemerges.

163 This line of thought has resonance in contemporary work on moral particularism (see Dancy, 2004; Hooker & Little, 2001).

I would like to make a final comment regarding evolutionary arguments in favor of newer “cognitive” mechanisms. When making such arguments, one needs to take into account that moral emotions were also shaped by evolutionary processes. So there are other (for some comparable) reasons which can lead us to claim that moral emotions are more accurate in guiding behavior and judgment. For example, emotions might work like a highly reliable, pattern recognition system for anti-moral behavior. Nature could have shaped our emotional reactions so that ‘moral’ behavior produces positive emotions and ‘immoral’ behavior produces negative emotions.¹⁶⁴ As shown in previous chapters, type-2 processes are low capacity and quite unreliable. By contrast, pattern recognition is high capacity and highly reliable (when shaped by expertise). So, given the complexity of morality and the number of variables involved in any particular situation, emotion-based moral intuitions might actually be more reliable for judging the morality of an action or event.¹⁶⁵ Of course, we don’t have any proof that emotions are good ‘recognizers’ of morality but the same can be said of slow, conscious, and controlled reasoning.

7. Some moral emotions seem to have been shaped by morally irrelevant factors, hence we should rethink our moral commitments by engaging “cognitive” capacities.

My final criticism is related to the psychological plausibility of Greene’s picture. One of the reasons that has motivated philosophers to do research on moral psychology is the inherently practical aspect of moral philosophy. Some philosophers think that the empirical adequacy of an ethical theory does not undermine its normative force. Yet other philosophes think that empirical

¹⁶⁴ Hume (1939/1978) held this position.

¹⁶⁵ In this picture, emotions are the way in which nature provides us with fast and reliable moral intuitions.

inadequacy is represents a significant challenge to any ethical theory, especially if the theory cannot be used as a good guidance for behavior either.¹⁶⁶

Another consideration is the following: even if we assume that there is one mechanism that is generally more accurate in its moral judgments, we cannot assume that it will be more reliable in motivating us to act in that way. A wide range of empirical data clearly shows that emotions play a key role in motivating and directing behavior.¹⁶⁷ Many empirical experiments also show that damage to our emotional areas can have a great impact in behavior, even to the point of leading us to destructive, amoral, and/or antisocial conduct. Greene agrees that emotions are crucial to moral behavior, yet he does not explain how this is compatible with assuming a purely “cognitive” perspective. That is, he does not explain why we need not worry that this *cognitive* turn will lead us to a motivational dearth.

IV. Greene’s Experimental Results and the Dual-System View

In the previous section, I showed why Greene’s argumentation is flawed. Instead of showing that processes leading to consequentialist judgments are superior in guiding moral judgment and decision making, he assumes it and looks for confirmative evidence. He claims that the superiority of certain processes is due to their more recent evolutionary development. Clearly this is not sufficient evidence for the claim that he seeks to defend. Evolution could have built us in such a way that we have specialized mechanisms for social and moral cognition that are different than those

¹⁶⁶ A relevant example is John Doris’ work on the empirical inadequacy of ‘virtue theory’ (Doris, 2002).

¹⁶⁷ For relevant reviews, see Angie, Connelly, Waples, and Kligyte (2011), Blanchette and Richards (2010), Mitchell (2011).

mechanisms for solving mathematical problems. Much more research is needed before we can draw this kind of conclusion (assuming we can decide these matters based on empirical evidence).

Why is it then so tempting to assume that “cognitive” (type-2) processes are generally more accurate than intuitive/automatic (type-1) processes? One reason is related to early developments on the dual-process view literature, particularly the heuristics and biases literature. Originally, a general way of seeking evidence for a dual-process view was to conduct research that showed a possible conflict between the two types of processes. The reason for this methodology was simple: if one can show that for the same problem two different, opposing answers arise, then a natural explanation is that there are in fact two processes dealing with the same problem in different ways. If the problem is supposed to have only one correct answer, then at least one of the two processes is making some kind of mistake. But in order to establish which process gives the wrong answer, there has to be an independent way of assessing the correctness of each response. This has led psychologists to choose tasks and problems that have a known mathematical, probabilistic, or logical solution. But, since by hypothesis this type of reasoning is prototypically type-2, what is left to do is to test whether under certain circumstances type-1 processes produce a different answer. By experimental design, the ‘intuitive’ answer will be always the wrong one.¹⁶⁸ And for this reason many researchers conclude that the intuitive/automatic answer is generally unreliable.

This methodology is very popular in the literature and it is inspired by Kahneman and Tversky’s (1973, 1982) early research on the heuristics and biases in judgment under uncertainty.

¹⁶⁸ It is remarkable the parallels between this methodology and the one used by Greene: here too the experiments are designed so that one answer (the consequentialist judgment)

Given the nature of their research, it is no surprise that many of their conclusions included a criticism of intuitive judgment and type-1 reasoning. Nevertheless, Kahneman later emphasized the importance of not reducing System 1 (the collection of type-1 processes) to an ineffective, biased, old form of reasoning. In Kahneman & Frederick (2002) the authors state that ‘although System 1 is more primitive than System 2, it is not necessarily less capable. On the contrary, complex cognitive operations eventually migrate from System 2 to System 1 as proficiency and skill are acquired’. Similarly, several researchers have strongly defended the adequacy of intuitive and heuristic processes for solving complex problems.¹⁶⁹

In sum, the supposition that when the two types of processes provide conflicting answers only one of them (mainly type-2) is correct is generally unfounded and only right for some domains. So, we should not extend the conclusions of the original research to all types of reasoning. In particular, we should remain unbiased about which of the two processes is generally right, and which of the two is better at solving specific kinds of problems (e.g., moral dilemmas).

Beyond these considerations, it is important to keep in mind that much more research needs to be conducted to investigate the interaction of the two processes, how conflict resolution is achieved, why some processes prevail on certain circumstances, and on how to independently establish which process is correct in case of conflicting answers. The latter is particularly important if we want to avoid the oversimplified view according to which type-2 processes are always more reliable than type-1.

¹⁶⁹ See Gigerenzer (2007), Myers (2002), Gladwell (2005), Dijksterhuis et al. (2006).

V. Concluding remarks

Greene's hypothesis that there are two alternative paths for moral judgment, one relying on "cognitive" processes and one relying on emotional reactions, can easily be related to the type-1/type-2 distinction. Further, his analysis of the data suggests that in the case of dilemmas with a high emotional content, participants resolve the conflict between the two processes by engaging areas in the frontal lobes (especially the ACC). This also offers support to the claim made in the previous chapter about the existence and importance of type-3 processes.

Greene's main points are that the exercise of cognitive control tends to favor consequentialist judgments, and that this fact shows that consequentialist judgments (and the mechanism generating them) are more accurate.

I have put forward several arguments against Greene's conclusions. I would like to emphasize two. First, it is far from clear that the exercise of cognitive control over evolutionarily older mechanisms and in favor of newer processes shows the superiority of the chosen response. Further, it is unclear that slow, deliberate, and conscious reasoning is generally more accurate in reasoning, judgment, and decision making tasks. In fact, there are some reasoning and decision making tasks for which intuitive responses seem to be more reliable, especially in the realm of social cognition. Given the close connection between moral and social issues, it wouldn't be surprising if type-1 processes were also better able to handle moral dilemmas. Hence, we cannot assume that 'rule-based reasoning' is better at solving moral conflict just because it is better at solving other type of problems (e.g., mathematical puzzles). As I mentioned, we need to find an independent way of assessing the correctness of both judgments. Then we need a clear method for establishing what

types of processes are providing each answer. Only then we will be able to draw some conclusions about the aptitude of each type of process in solving moral dilemmas. There is no reason to assume, as Greene does, that “cognitive” processes will provide more reliable answers in moral judgment and decision making.

Secondly, it seems clear to me that Greene is assuming the correctness of the consequentialist judgment and only looking for a way to “confirm” and justify this.¹⁷⁰ He claims that emotion based judgments are the result of once effective but now defective evolutionarily evolved mechanisms, a questionable claim that begs the question about the appropriateness of different normative views. For example, Greene maintains that moral emotions are naturally flawed because they are sensitive to some morally irrelevant factors. But this is not a theoretically neutral claim, it will depend on our choice of moral theory. These type of claims can only be experimentally confirmed if we agree on which factors are morally irrelevant and which aren’t. Only then we can empirically confirm whether emotion based intuitive moral judgments are sensitive to them.

¹⁷⁰ Ironically, Greene falls prey to confirmation bias, one of the most studied biases in the heuristics and biases literature.

Conclusion

I began this dissertation by showing that a wealth of psychological data suggests that two different processes are involved in higher-level cognition. One process (type-1) is fast, automatic, low-effort, high-capacity, implicit, and unconscious; the other (type-2) is slow, deliberate, high-effort, low-capacity, explicit, and conscious. Among others, dual-process views have been proposed for learning, memory, moral and social judgment, reasoning, and decision making.

Some researchers have sought to unify these views under a general dual-system theory. The generic dual-system view claims that most (if not all) dual-process views can be subsumed under a general dual-system framework. According to this view, there are two general systems of cognition. One system (System 1) underlies type-1 processes, the other system (System 2) underlies type-2 processes. These two systems operate sometimes cooperatively and sometimes competing for cognitive control.

The main problem with the generic dual-system view is the lack of a robust characterization of the notion of *system*, one that can explain and predict the empirical results. Indeed, the dual-system view claims that there are two types of processes that manifest a cluster of opposite properties and the two types of processes manifest opposite characteristics precisely because they are underlid by two different systems.

In the dissertation I show that there is no non-ad hoc way of keeping the distinction between the two systems. Appeals to underlying cognitive architecture, the use of working memory, and levels of analysis are all useful to explain some of the data and make sense of the complexity of processes involved in higher cognition. However, none can provide an explanation for the high

variability of processes described in the literature. I argue that the reason why these (and other accounts) fail to explain the available data is that there isn't a unified and consistent opposition of two cognitive types but several types of processes.

My own view is that there is no privileged distinction, no two systems, but a set of properties that are cognitively significant because they identify important differences in how we learn, process, and use information. And even though these differences do not always align, there are some types of processes that are particularly important because they embody the prototypical construal of the contrast between *intuition* and *reasoning*. I show that these two *preferred* types of processes (i.e., slow, effortful, conscious thinking and fast, effortless, unconscious intuition) many times owe the clustering of properties to different underlying cognitive micro-architectures (that is, different computational principles). Using Marr's distinction of three levels of analysis of a cognitive system, I showed that the difference between these two types of processes concerns the algorithmic level. That is, I showed that the computational principles used by each process are different. Type-1 processes are better modeled by connectionist architectures while type-2 processes are better modeled classically. In this way, I endorsed a 'hybrid' micro-architecture of the mind, partly connectionist, partly classical. To illustrate the merits of this view, I carefully discussed some of the main psychological experiments which support a dual-process view and showed that these and similar results can be explained and predicted by the 'hybrid' proposal that I defend. I also established that neither a purely connectionist architecture nor a purely classical one can explain and predict these results as well as the hybrid view. Still, I added that there is no strict correspondence between underlying micro-architecture and phenomenological properties and this is the reason why there are processes that share properties with the two preferred types.

Finally, I showed that this view lacks a central component: executive processes. At the level of macro-architecture (the general cognitive architecture of the mind) there is another important type of process (type-3 processes) involved in conflict resolution and cognitive control, that are crucial for understanding the interaction and cooperation of all the mind's systems.

All these factors render the *duality of the mind* hypothesis highly implausible: there are more than two types of processes and more than two kinds of systems under any plausible interpretation. However, dual-process and dual-system views have played a crucial role in advancing our understanding of the complexity of the cognitive architecture of the mind.

To conclude, I argued that despite its many shortcomings, the dual-system view delivers important insight about the nature of the mind and the complexity of higher cognitive processes and for this reason it is philosophically and psychologically significant. In the last chapter, I considered one example of the application of the dual-system view to philosophical discussion: dual-processes views of moral judgment and decision making. I showed that there is no way to decide the better aptitude of a type of processes based on the performance of that process in unrelated tasks. Indeed, type-2 processes can be more accurate in many mathematical, logical, and reasoning tasks, but type-1 processes can be more accurate in tasks that involve a high number of variables or tasks that are computationally too demanding for conscious rule-based thinking. Further, as one acquires expertise in a domain, performance becomes more automatic and processes 'migrate' from type-2 to type-1. Hence, when it comes to assessing accuracy, the first thing to consider is whether intuitive (i.e., fast, automatic) processing is due to expertise and not just 'sloppy/lazy' thinking. If that is the case, then one should try to use that expertise, especially if there are no well-known, simple algorithms for solving the same tasks. After all, chances are that using that expertise will end

up being more reliable than using slow, deliberate thought (which faces its own set of problems).

Finally, for those cases in which there is no known method for acquiring expertise, nor a set of well-known rules, the question remains open as to which is the best method for solving the problem.

This seems to be the case for moral reasoning.

References

1. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, 111(4), 1036-1060.
2. Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in the human prefrontal cortex. *Nature Neuroscience*, 2, 1032–1037.
3. Aizawa, K. (2004). In E. Hochstein (Ed.), *Dictionary of Philosophy of Mind*. Retrieved from <http://sites.google.com/site/minddict/computational-architecture/>
4. Angie, A. D., Connelly, D., Waples, E. P., & Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: A meta-analytic review. *Cognition & Emotion*, 25(8), 1393-1422.
5. Aristotle (2000). *Nichomachean Ethics*. (R. Crisp, Trans.). Cambridge: Cambridge University Press.
6. Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York, NY: Guilford.
7. Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken and Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York, NY: Guilford.
8. Bargh, J. A. (2005). Bypassing the will: Towards demystifying behavioral priming effects. In R. Hassin, J. Uleman & J. Bargh (Eds.), *The new unconscious*. Oxford: Oxford University Press.
9. Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, 36, 147-168.
10. Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype priming on action. *Journal of Personality and Social Psychology*, 71, 230-244.
11. Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3), 295-307.
12. Bechtel, W. P. (1987). Connectionism and the Philosophy of Mind: An overview. *The Southern Journal of Philosophy*, Supplement 26, 17-41.
13. Bechtel, W. P., Abrahamsen, A. A. (2005). Explanation. A mechanistic alternative. *Studies in History and Philosophy of Biology and Biomedical Sciences*, 36, 421-441.

14. Berry, D. C., & Dienes, Z. (1993). *Implicit learning*. Hove: Erlbaum.
15. Blanchette, I., & Richards, A., (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning, *Cognition & Emotion*, 24(4), 561-595.
16. Bolender, J. (2001). A two-tiered cognitive architecture for moral reasoning. *Biology and Philosophy*, 16, 339-56.
17. Boyd, R. (1991). Realism, antifoundationalism, and the enthusiasm for natural kinds. *Philosophical Studies*, 61, 127–148.
18. Carruthers, P. (2006). *The architecture of the mind*. Oxford: Oxford University Press.
19. Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299, 999-1000.
20. Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5), 752-756.
21. Chalmers, D. (1993). Why Fodor and Pylyshyn were wrong: The simplest refutation. *Philosophical Psychology*, 6, 305-319.
22. Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken and Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73-96). New York, NY: Guilford.
23. Churchland, P. S., & Sejnowski, T. (1990). Neural Representation and Neural Computation. *Philosophical Perspectives*, 4, 343-382.
24. Damasio, A. R. (1994). *Descartes' Error. Emotion, Reason, and the Human Brain*. New York, NY: Grosset/Putnam.
25. Damasio, A. R., Tranel, D., Damasio, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioral Brain Research*, 41, 81–94.
26. Dancy, J. (2004). *Ethics without principles*. Oxford: Clarendon Press.
27. Dawson, M., Medler, D., MacCaughan, D., Willson, L., & Carbonaro, M. (2000). Using extra output learning to insert a symbolic theory into a connectionist network. *Minds and Machines*, 10, 171–201.
28. Dawson, M., & Shamanski, K. (1994). Connectionism, confusion, and cognitive science. *The Journal of Intelligent Systems*, 4, 215-262.
29. De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, 17, 428-433.
30. Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1, 568-570.
31. Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. *Journal of personality and social psychology*, 56(1), 5-18.

32. Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology*, 87, 586–598.
33. Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & von Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, 311, 1005-1007.
34. Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.
35. Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: The MIT Press.
36. Eichenbaum, H., & Bunsey, M. (1995). On the binding of associations in memory: Clues from studies on the role of the hippocampal region in paired-associate learning. *Current Directions in Psychological Science*, 4(1), 19-23.
37. Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. Oxford: Oxford University Press.
38. Eliasmith, C. (2004). Connectionism. In E. Hochstein (Ed.), *Dictionary of Philosophy of Mind*. Retrieved from <https://sites.google.com/site/minddict/connectionism/>
39. Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, 49(8), 709-724.
40. Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4), 451-468.
41. Evans J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Brighton: Erlbaum.
42. Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science*, 7, 454-459.
43. Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, 13(3), 378-395.
44. Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255-278.
45. Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33-54). Oxford: Oxford University Press.
46. Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford: Oxford University Press.
47. Evans, J. St. B. T., Barston, J. L., & Pollard P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295-306.
48. Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11(4), 382-389.
49. Evans, J. St. B. T., & Frankish, K. (Eds.) (2009). *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.

50. Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove: Psychology Press.
51. Evans, J. St. B. T., Over, D. E., & Handley, S. J. (2003). A theory of hypothetical thinking. In D. Hardman & L. Maachi (Eds.), *Thinking: Psychological perspectives on reasoning, judgement and decision making* (pp. 3-22). Chichester: Wiley.
52. Feldman, J., & Ballard, D. (1982). Connectionist models and their properties. *Cognitive Science*, 6, 205-254.
53. Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
54. Fodor, J. (1983). *The Modularity of Mind*. New York, NY: Crowell.
55. Fodor, J., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35, 183-204.
56. Fodor, J., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: a Critical Analysis. *Cognition*, 28, 3-71.
57. Foot, P. (1978). The problem of abortion and the doctrine of double effect. In *Virtues and Vices*. Oxford: Blackwell.
58. Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.
59. Frankish, K. (2009). Systems and levels: Dual-system theories and the personal–subpersonal distinction. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 89-107). Oxford: Oxford University Press.
60. Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10), 914-926.
61. Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1-29). Oxford: Oxford University Press.
62. Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42.
63. Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy*, 105(10), 634-663.
64. Gigerenzer, G. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin*, 119, 23–26.
65. Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. London: Penguin.
66. Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
67. Gigerenzer, G., Todd, P. M., & ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.
68. Gilovich, T., & Griffin, T. (2002). Introduction. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 1-18). New York, NY: Cambridge University Press.

69. Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.
70. Gladwell, M. (2005). *Blink*. New York, NY: Little Brown and Company.
71. Goel, V. (2003). Evidence for dual neural pathways for syllogistic reasoning. *Psychologica*, 32, 301–309.
72. Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, 11, 435-441.
73. Goel, V., Buchel, C., Frith, C., & Dolan, R. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage*, 12, 504–514.
74. Goldberg, E. (2009). *The new executive brain: Frontal lobes in a complex world*. Oxford: Oxford University Press.
75. Greene, J. D. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience*, 4(10), 846-850.
76. Greene, J. D. (2007). *The secret joke of Kant's soul*. Cambridge, MA: MIT Press.
77. Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364-371.
78. Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144-1154.
79. Greene, J. D., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389-400.
80. Greene, J. D., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105-2108.
81. Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of personality and social psychology*, 65(4), 613.
82. Harman, G., & Kulkarni, S. (2007). *Reliable reasoning: Induction and statistical learning theory*. Cambridge, MA: MIT Press.
83. Harman, G. (1985). Is there a single true morality? In D. Copp & D. Zimmerman (Eds.), *Morality, Reason and Truth* (pp. 27-48). Totowa, NJ: Rowman and Allanheld.
84. Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences*, 1, 215-226.
85. Hooker, B., & Little, M. O. (Eds.). (2000). *Moral particularism*. Oxford: Clarendon Press.
86. Horgan, T., & Tieson, J. (1999). Rules and representations. In R. Wilson & F. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 725-726). Cambridge, MA: MIT Press.
87. Hume, D. (1978). *A treatise of human nature*. (L. Selby-Bigge & P. Nidditch, Eds.). Oxford: Oxford University Press. (Original work published 1739)

88. James, W. (1950). *The Principles of Psychology* (Vol. 1). New York, NY: Dover. (Original work published 1890)
89. Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5), 1449-1475.
90. Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Macmillan.
91. 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* (pp. 49-81). New York, NY: Cambridge University Press.
92. Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267-293). Cambridge: Cambridge University Press.
93. Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.
94. Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237-251.
95. Kamm, F. M. (2000). Does distance matter morally to the duty to rescue. *Law and Philosophy*, 19(6), 655-681.
96. Keren, G., & Schul, Y. (2009). Two Is Not Always Better Than One: A Critical Evaluation of Two-System Theories. *Perspectives on Psychological Science*, 4(6), 533-550.
97. Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130(2), 203-231.
98. Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, 11(1), 16-22.
99. Leibniz, G. W. (1981). *New Essays on Human Understanding*. (P. Remnant & J. Bennett, Trans. & Eds.). Cambridge: Cambridge University Press. (Original work published 1765)
100. Lewis, D. K. (1986). *On the plurality of worlds*. Oxford: Blackwell.
101. Lewis, R. L. (1999). Cognitive modeling, symbolic. In R. Wilson & F. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp.525-527). Cambridge, MA: MIT Press.
102. Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, 126, 109-137.
103. Lieberman, M. D. (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In J. Forgas, K. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44-67). New York, NY: Cambridge University Press.
104. Lieberman, M. D. (2007). The X- and C-Systems: The Neural Basis of Automatic and Controlled Social Cognition. In E. Harmon-Jones & P. Winkielman (Eds.), *Social neuroscience: Integrating biological and psychological explanations of social behavior* (pp. 290-315). New York, NY: Guilford Press.

105. Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
106. Mikels, J. A., Maglio, S. J., Reed, A. E., Kaplowitz, L. J., (2011). Should I go with my gut? Investigating the benefits of emotion-focused decision making. *Emotion*, 11(4), 743-753.
107. Mikhail, J. (2008). Moral Cognition and Computational Theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 3). Cambridge, MA: MIT Press.
108. Mitchell, D. G. (2011). The nexus between decision-making and emotion regulation: a review of convergent neurocognitive substrates. *Behavioural Brain Research*, 217(1), 215–231.
109. Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *Neuroimage*, 16(3), 696-703.
110. Moll, J., de Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition. *Annals of the New York Academy of Sciences*, 1124(1), 161-180.
111. Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, 6(10), 799-809.
112. Myers, D. G. (2002). *Intuition: Its powers and perils*. New Haven, CT: Yale University Press.
113. Nagel, J. (2007). Epistemic Intuitions. *Philosophy Compass*, 2(6), 792–819.
114. Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4), 608-631.
115. Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in cognitive sciences*, 5(8), 349-357.
116. Piccinini, G. (2009). Computationalism in the philosophy of mind. *Philosophy Compass*, 4, 515–532.
117. Piccinini, G. (2012). Computation in physical systems. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition). Retrieved from <https://plato.stanford.edu/archives/fall2012/entries/computation-physicalsystems/>
118. Posner, M. I., & Fernandez-Duque, D. (1999). Attention in the human brain. In R. Wilson & F. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 43-46). Cambridge, MA: MIT Press.
119. Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT press.
120. Reber, A. S. (1965). Implicit learning of artificial grammars. Unpublished master's thesis, Brown University.
121. Reber, A. S. (1966). A perceptual learning analysis of probability learning. Unpublished doctoral dissertation, Brown University.
122. Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, 6(6), 855-863.
123. Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology*, 118(3), 219-235.

124. Reber, Arthur S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford University Press.
125. Reber, A. S., Allen, R., Allen, P. J. (1999) Implicit versus Explicit Learning. In R. Sternberg (Ed.), *The Nature of Cognition* (pp. 475-513). Cambridge, MA: MIT Press.
126. Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, 7(1), 1-75.
127. Roberts, M. J., & Newton, E. J. (2002). Understanding strategy selection. *International Journal of Human-Computer Studies*, 54(1), 137-154.
128. Rumelhart, D., McClelland, J., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.
129. Samuels, R. (2008). Psychology. In S. Psillos & M. Curd (Eds.), *The Routledge Companion to the Philosophy of Science* (pp. 581-593). New York, NY: Routledge.
130. Samuels, R. (2009a). Delusion as a natural kind. In M. R. Broome & L. Bortolotti (Eds.), *Psychiatry as cognitive neuroscience: philosophical perspectives* (pp. 49-79). Oxford: Oxford University Press.
131. Samuels, R. (2009b). The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 129-146). Oxford: Oxford University Press.
132. Schacter, D. L., & Scarry, E. (2000). Introduction. In D. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 1-10). Cambridge, MA: Harvard University Press.
133. Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 1-38). Cambridge, MA: MIT Press.
134. Schaich-Borg, J., Lieberman, D., & Kiehl, K. A. (2008). Infection, incest, and iniquity: Investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, 20(9), 1529-1546.
135. Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
136. Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.
137. Smith, B. C. (1999). Computation. In R. Wilson & F. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 153-155). Cambridge, MA: MIT Press.
138. Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: distributed social cognition. *Psychological review*, 116(2), 343-364.
139. Smith, E., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration links to underlying memory systems. *Personality and Social Psychology Review*, 4, 108-131.
140. Smolensky, P. (1988). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26 (Supplement), 137-163.

141. Sober, E., & Wilson, D. S. (Eds.) (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.
142. Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Psychology Press.
143. Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55-88). Oxford: Oxford University Press.
144. Sun, R. (1996). Hybrid connectionist-symbolic modules: A report from the IJCAI-95 workshop on connectionist-symbolic integration. *AI Magazine*, 17(2), 99-103.
145. Sun, R. (2008). Introduction to computational cognitive modeling. In R. Sun (Ed.), *Cambridge handbook of computational psychology* (pp. 3-19). Cambridge: Cambridge University Press.
146. Sun, R., Lane, S. M., & Mathews, R. C. (2009). The two systems of learning: An architectural perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 239-262). Oxford: Oxford University Press.
147. Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*, 25(2), 203-244.
148. Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological review*, 112(1), 159.
149. Thagard, P. (2012). Cognitive Science. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition). Retrieved from <http://plato.stanford.edu/archives/fall2012/entries/cognitive-science/>
150. Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59, 204-217.
151. Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), 207-232.
152. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.
153. Tversky, A., & Kahneman, D. (1982a). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153-160). New York, NY: Cambridge University Press.
154. Tversky, A., & Kahneman, D. (1982b). Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-98). New York, NY: Cambridge University Press.
155. Tversky, A., & Kahneman, D. (1983). Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probabilistic Judgment. *Psychological review*, 90(4), 293-315.
156. Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92(7), 345-381.

157. Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology I* (pp. 106-137). Harmandsworth: Penguin.
158. Wason P. C., & Evans J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, 3, 141-154.
159. Williams, B. (2011). *Ethics and the Limits of Philosophy*. London: Fontana.
160. Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, 107(1), 101-126.
161. Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of personality and social psychology*, 60(2), 181-192.