UNIVERSIDADE FEDERAL DE MINAS GERAIS
SAMUEL DE CASTRO BELLINI LEITE

PREDICTING AND REFLECTING: A DUAL FRAMEWORK FOR DUAL PROCESS THEORY

Belo Horizonte
2017

SAMUEL DE CASTRO BELLINI LEITE

PREDICTING AND REFLECTING: A DUAL FRAMEWORK FOR DUAL PROCESS THEORY

Tese apresentada ao programa de doutorado em Filosofia, área "Lógica, Ciência, Mente e Linguagem" da Universidade Federal de Minas Gerais, com orientação do Prof. Dr. André Joffily Abath e co-orientação do Prof. Dr. Keith Frankish.

Belo Horizonte
2017

100
L533p
2017

Leite, Samuel de Castro Bellini

Predicting and reflecting [manuscrito] : a dual framework for dual process theory / Samuel de Castro Bellini Leite. - 2017.

213 f.

Orientador: André Joffily Abath.
Coorientador: Keith Frankish.

Tese (doutorado) - Universidade Federal de Minas Gerais, Faculdade de Filosofia e Ciências Humanas.

Inclui bibliografia

1.Filosofia – Teses. 2..Raciocínio – Teses. 3. Processo decisório - Teses. I. Abath, André Joffily. II. Frankish, Keith. III. Universidade Federal de Minas Gerais. Faculdade de Filosofia e Ciências Humanas. IV. Título.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

FOLHA DE APROVAÇÃO

Predicting and Reflecting: A Dual Framework For Dual Process Theory

SAMUEL DE CASTRO BELLINI LEITE

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em FILOSOFIA, como requisito para obtenção do grau de Doutor em FILOSOFIA, área de concentração FILOSOFIA, linha de pesquisa Lógica e Filosofia da Ciência.

Aprovada em 03 de maio de 2017, pela banca constituída pelos membros:

Prof. André Abath - Orientador
UFMG

Prof. Richard Samuels
Ohio University

Prof. Ernesto Perini Frizzera da Mota Santos
UFMG

Prof. Marco Aurélio Sousa Alves
UFSJ

Belo Horizonte, 3 de maio de 2017.

—Dedicated to all emerging production in Brazilian
Philosophy of Mind and Cognitive Science

AGRADECIMENTOS

Gostaria de agradecer primeiramente ao meu orientador André Joffily Abath por todos os comentários, incentivos, discussões, oportunidades, e principalmente por ter acreditado tanto em mim e no meu trabalho.

À Ernesto Perini Frizzera da Mota Santos pelo o apoio e abertura desde o início, procurando sempre facilitar meu percurso e por leituras da tese com excelentes sugestões.

Aos professores da linha de pesquisa Lógica, Ciência, Mente e Linguagem por serem perfeitos exemplos da eficácia entre tradição e filosofia temática, agradeço em especial, Túlio Roberto Xavier de Aguiar e Patrícia Maria Kauark Leite.

Ao grupo de Filosofia da Mente da ANPOF por sempre apoiarem minha participação e discutirem meu trabalho em diversos eventos.

Ao grupo de CLE de auto-organização por me receber de braços abertos em diversas oportunidades em seus seminários.

Gostaria de agradecer aos meus pais José Luiz Bellini Leite e Heloisa de Castro Leite pelo amor, atenção e respeito que tiveram durante os meus anos como estudante e também por batalharem além do limite pela possibilidade da minha pós-graduação.

Aos meus irmãos Saul de Castro Leite e Sarah de Castro Bellini Leite por literalmente salvarem minha vida mais de uma vez durante esse tempo.

Ao meu amor, meu filho Joao Felipe Mamede Bellini Leite que com apenas sete anos de idade já esbanja tanta sabedoria. Agradeço especialmente pelas palavras: “Papai, faça o seu melhor, tudo bem se não conseguir, só quero que você dê o seu máximo.”

Ao meu segundo filho Francisco Mamede Bellini Leite, por ser a pessoa mais capaz de demonstrar amor por meio de expressões e gestos em todo o mundo.

Aos amigos Lucas Nascimento Machado, Daniel De Luca, Daniel Silva, Alexandre Ferraz, João Moraes e Nathália Pantaleão pelas discussões, oportunidades e por não me deixarem perder meu espírito de filósofo durante os anos da pós-graduação. Também à Fernando Palácios e Marcelo Eiterer por correrem atrás de mim mesmo quando eu não merecia.

Finalmente, e especialmente, à minha esposa Marina Luiza dos Santos Mamede por me incentivar a buscar o trabalho que sonho. Agradeço Marina Mamede também por ter convivido por tantos anos com uma pessoa tão presa em sua própria mente e que por isso muitas vezes não demonstrou a sensibilidade que uma mulher espetacular como ela merecia.

Agradecemos o apoio da CAPES pelo financiamento dessa pesquisa.

ACKNOWLEDGEMENTS

First and foremost I thank Keith Frankish for reading my work, commenting, having online video conferences, hearing my ideas, and most of all for standing by me until the end even when he had little personal gain in such.

Richard Samuels for reading my work and challenging me on so many levels, and also for the patience to endure my misunderstandings.

Andy Clark for inspiring this work from beginning to end and for e-mails illuminating issues in predictive processing.

Also, Gualtiero Piccinini for material and hints for this work.

There used to be a discipline called speculative psychology. It wasn't quite philosophy because it was concerned with empirical theory construction. It wasn't quite psychology because it wasn't an experimental science. But it used the methods of both philosophy and psychology because it was dedicated to the notion that scientific theories should be both conceptually disciplined and empirically constrained. What speculative psychologists did was this: They thought about such data as were available about mental processes, and they thought about such first-order psychological theories as had been proposed to account for the data. They then tried to elucidate the general conception of the mind that was implicit in the data and the theories. Speculative psychology was, by and large, quite a good thing. William James and John Dewey were speculative psychologists and so, in certain of his moods, was Clark Hull. [...] This book, in any event, is unabashedly an essay in speculative psychology.

—Jerry Fodor in The Language of Thought (1975)

Resumo

A Teoria de Duplo Processo tem ganhado fama enquanto um arcabouço para explicar evidências em tarefas de raciocínio e tomada de decisão. Essa teoria propõe que deve haver uma distinção afiada no pensamento para explicar dois aglomerados correlacionais de características. Um aglomerado descreve um processo rápido e intuitivo (Tipo 1), enquanto o outro descreve um devagar e reflexivo (Tipo 2) (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011). Entretanto, como Samuels (2009) notou, existe um problema em determinar o motivo desse grupo de características formarem aglomerados, mais do que os rótulos Tipo (ou sistema) 1 e 2 conseguem capturar, 'o problema da união'. Entendemos que podem haver diferenças nas arquiteturas cognitivas que sustentam cada tipo de processo, assim exigindo arcabouços cognitivos distintos para cada. Argumentamos que a abordagem do processamento preditivo (como apresentada por Hohwy, 2013 e Clark, 2016) é um arcabouço mais adequado para processos do Tipo 1. Tal abordagem propõe que a cognição tem o trabalho de prever o que perturbará os inputs sensoriais em um próximo momento. Essas não são previsões pessoais mas múltiplas previsões subpessoais que até o sistema visual realiza em vários níveis em cada milissegundo que passa. Ao invés de ser baseado em representações simbólicas de cada aspecto do mundo, essas previsões são realizadas com base em informação estatística atualizada a cada momento. Kahneman (2011) vem argumentando que existe uma ligação entre a percepção e processamento Tipo 1. O que sustentamos é que tal ligação existe pois julgamentos do Tipo 1 na verdade são previsões derivadas de níveis altos de sistemas perceptivos que funcionam por meio do processamento preditivo. Por outro lado, argumentamos que tal arquitetura não funciona para processos do Tipo 2. Em vez disso, esses processos parecem estar baseados em sistemas simbólicos clássicos executando busca heurística como explicado por Newell (1980). Em conclusão, propomos que uma arquitetura dupla é necessária para explicar por qual motivo existem dois aglomerados de características. Esse arcabouço incluiria processamento preditivo para explicação de processos do Tipo 1 e computações em representações simbólicas para processos do Tipo 2.

Palavras-Chave: Teoria de Duplo Processo; Raciocínio e Tomada de decisão; Processamento Preditivo; Sistemas Simbólicos.

Abstract

Dual Process Theory has increasingly gained fame as a framework for explaining evidence in reasoning and decision making tasks. This theory proposes there must be a sharp distinction in thinking to explain two clusters of correlational features. One cluster describes a fast and intuitive process (Type 1), while the other describes a slow and reflective one (Type 2), (see Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011). However, as Samuels (2009) has noted, there is a problem of determining why these group of features form clusters, more than what the labels Type (or system) 1 and 2 can capture, the unity problem. We understand there might be differences in the processing architecture that grounds each type of process, thus requiring distinct cognitive frameworks for each. We argue that the predictive processing approach (as held by Hohwy, 2013 and Clark, 2016) is a more suitable framework for Type 1 processing. Such an approach proposes cognition is in the job of attempting to predict what will perturb sensory inputs next. These are not personal predictions but rather multiple sub-personal predictions that even the visual system makes at various layers at each millisecond that passes. Rather than being based on a symbolic representation of each aspect of the world, these predictions are made on the basis of statistical information updated moment by moment. This statistical content tracks previous sensory states and the causes of these previous sensory states. Kahneman (2011) has been arguing that there is a link between perception and Type 1 processing. What we hold is that such link obtains because Type 1 judgments actually are predictions stemming from higher layers of perceptual systems which work by means of predictive processing. On the other hand, we propose such architecture does not handle Type 2 processes. Rather, these seem to be based on classical symbol systems executing heuristic search as explained by Newell (1980). In conclusion, we propose a dual framework is necessary for explaining why there are two clusters of features. Such a framework would include predictive processing for explaining Type 1 processing and computations on symbolic representations for Type 2 processing.

Key-words: Dual Process Theory; Reasoning and Decision Making; Predictive Processing; Symbol Systems.

TABLE OF CONTENTS

Introduction

This introduction is twofold. The first part presents an outline of how the thesis will look like, introducing problems, our line of argument and hypotheses while providing a first inspection into each of the chapters. The second part presents an introduction to the Fodorian architecture of the mind which will serve as our theoretical starting point.

I Problems, line of argument and hypotheses

When solving math problems, it is interesting to note how some calculations, such as '2+2', almost seem to carry with it its own answer. Thus, when we hear or read such calculation, the number four rushes to mind inevitably. This observation is at odds with other calculations such as '74-37' which are also simple but no answer comes along. One applies a few mental steps in order to reach the answer, it is easy, but the answer does not come to mind as one reads or hears the numbers. Regardless of what the characteristics of the numbers are which evokes this difference in thought, it is notable that there seems to be a difference between springing to mind inevitably and applying mental steps.

This sort of observation has led psychologists working under reasoning, judgment and decision making to suppose there are two sorts of processing in thinking which are responsible in one side for faster and effortless judgments and on the other side slower and effortful reasoning.

Of course, the evidence for a distinction in two types of processing is not limited to such simple observation but rather has appeared repeatedly since the 60s when (mainly) the tradition known as heuristics and biases started studying human reasoning, judgment and decision making by crafting tasks in which conflicting responses were common. In Chapter one—*Research in Reasoning and Rationality*—we will then review some of the most famous tasks in such tradition along with possible interpretations given to the results. Including heuristics, evolutionary psychology, pragmatics and dual process theories.

Dual process theories also are not limited to such simple theorizing and characterizations as we introduced. Thus in Chapter two—*Basic Dual Process Theory*—we work on fundamental

theoretical questions[1] that have been raised for dual process theories, including: if dual process theories can be seen as a complete theory of cognition or only of reasoning; what features characterizes each type of processing and which are defining; how one process communicates with the other and even if there are different processes at all or if distinct features points to dual systems, dual minds or dual modes (the reference problem).

Samuels (2009) notes that even if one considers the evidence to be convincing and the dichotomy of processes (termed Type 1 and Type 2) and their property clusters (termed S1 and S2) well placed, we still have a basic research question open, which he calls the unity problem:

> "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. 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-exhibit-ing processes. Again, this does not strike me as a serious objection so much as a challenge for future research—one that requires a more detailed account of the systems responsible for type-1 and type-2 processes." (Samuels, 2009, p.141).

The unity problem is distinct from the reference problem. The reference problem is the problem of determining what these property clusters refer to in the mind, to which a possible answer would simply be 'distinct systems', for instance. The unity problem asks: even supposing there really are two main systems responsible for the way we reason, what are the characteristics of these systems that explain why they exhibit these distinct features, rationalities and responses? What kind of systems are these that explain why they bind these group of properties together?

The unity problem guides our endeavors as we strive to explain what is underlying our mental architecture that is responsible for computational differences which result in the cognitive and behavioural effects observed by dual process theorists.

Fodor (1983) shares a similar duality of mind, which will be our starting point in the next part of this introduction. He argues the mind can be divided into modular input systems (or vertical faculties more generally) and domain-general central systems (or horizontal faculties). Following the frame problem[2], Fodor (1983, 1987, 2001), one of the philosopher most responsible for

1 The curious mind may wish to consult appendix I for these questions.
2 A famous problem in Artificial Intelligence and Philosophy (see Pylyshyn, 1987 and Bellini-Leite, 2017).

developing the computational theory of mind in the first place, dismissed it as a complete picture of mind. He detects the problem originates because such theory cannot apply to central systems—which in his descriptions are abductive, based on semantics, and open to any aspect of the person's whole web of belief—but only to input systems which proceed syntactically and are informationally encapsulated.

One of the intuitions[3] that started our project is that Fodor (2001) diagnosed this inversely. We think rather that it is his characterization of input systems (inflexible) that renders central systems obsolete. If there is something like central systems in our cognition any sort of input to it must already be contextualized and relevant before even reaching it. We are with Fodor's starting chapters (in 1975) where he started to propose the language of thought as means of explaining our psychology of propositional attitudes for decision and choice. This is the main role for classical architectures in our proposal and not its application to input and perception (like in Fodor, 1983).

We hold that what he characterized as input systems is better understood under a Hierarchical Bidirectional Predictive Processing (HBPP) framework as proposed by Friston (2003, 2005, 2008, 2010), Hohwy (2013) and Clark (2016), which is also a computational theory of mind, but not *the* computational theory of mind of Fodor. The predictive processing account unifies aspects of different models (such as predictive coding, emulation theory, active inference) about a very similar story. Such story suggests the brain should be thought of as in a restless, active cycle of predicting what will perturb it in a proximal and distal future. Instead of being understood as reading input from the world, the predictive brain story suggests the brain uses statistics to anticipate input before they even arrive. These predictions are based on expectations (or a statistical generative model) which foresees the most likely outcome of stimuli and events and takes a bet that thus and so will obtain. By keeping and polishing expectations, the brain can poise itself ahead of the game when it comes to dealing with the world.

These models suggest the brain is formed by a hierarchy of processing (comprising higher and lower levels) where multiple layers of neurons are organized to compose a network with two major streams of information flow. The top-down flow is understood as conveying multiple

[3] See Bellini-Leite (2013) for a related intuition.

predictions, each higher layer attempts to predict the workings of the one underneath it. The bottom-up flow conveys error correction on previously made predictions to each higher layer. If predictions of a given event are on track than lower sensory stimulation will be attenuated, they will not even be considered. On the other hand, if predictions mislead, sensory stimulation will flag the difference between what was predicted and what was sensed and the system will try to overcome such gap. This, prediction error minimization, Clark (2016) claims, is the brain's major goal.

Clark (2016) notes a very strong and interesting shift the predictive processing approach suggests. It proposes that the forward flow consists not so much of all the features that were detected to be passed onwards to higher levels but only the error necessary to correct and update models. Instead of conveying all information from the environment, rather it provides a natural funnel which guarantees processing economy by focusing on newsworthy information in the form of error correction.

While Clark (2016) sees predictive processing as a unified framework for the mind, including perception, action, reasoning and language. We will defend in Chapter Three—*Intuitive Predictions*—that when it comes to thinking and reasoning, predictive processing is suitable to understand only Type 1 processing. We show this by demonstrating how predictive processing fits with each of Type 1 defining and correlational features, and does not with Type 2 features.

Finally, in Chapter Four—*Reflecting As Expression Search*—we propose that Newell's (1980)[4] architecture for cognition is suited for Type 2 features in reasoning. Newell and Simon (1976) exposition of heuristic search also plays a good role in this explanation and the GPS (Newell and Simon, 1963) model for human reasoning serves as an excellent example. We also understand Fodor and Pylyshyn (1987) make an interesting case when characterizing classical system as based on compositionality. We understand these authors make a good case for the powers of classical architecture which we see explaining Type 2 but not Type 1 features. We show how the reasons Type 1 features are not covered by classical architecture have been demonstrated by issues in artificial intelligence, mainly the frame problem.

[4] This paper is written retrospectively, that is why it has a late date in relation to other works in classical architectures.

We understand this thesis can answer the unity problem by explaining how computational differences in mechanisms used in each type of process make these property clusters come about. Thus, what explains S1 properties are the powers and limitations of predictive processing. On the other hand, what explains the S2 cluster is the powers and limitations of classical architectures.

Simply exposed, these are the main problems, our line of argument and hypotheses that this thesis follows[5].

II The Fodorian architecture as a starting point

We have now presented the path taken by thesis and the first step is to introduce the Fodorian architecture of mind. The Fodorian architecture is actually a dual process theory in its own, and one that is supposed to apply to all functions of the mind. The duality proposed by Fodor (1983) is one of input systems and general systems. What is interesting about this proposal is that it is very detailed and specific. Thus, Fodor (1983) lists very rigid criteria to explain each system. It is not exaggerated to say that any dual process theory in cognitive science was at least partially influenced by this proposal.

Fodor (1983) enumerates a few features which input systems are likely to have. He argues input systems are domain specific, that is, they respond to stimuli in one specific type of domain and ignore other modalities. Now, Fodor's domain-specificity is strong. That is because a domain specific system in his view is not that which deals with specific kinds of input, it must deal with only such inputs and must be specifically structured as to account for such input, rendering it blind to other domains. The property he mentions as one that helps identify domain-specificity is handling eccentric stimuli domains. These are domains that need specific handling and that differ from more general and usual tasks of the mind. For instance, a usual task is categorizing items according to their class, but to understand sentences one cannot classify tokens of language sentences, it is probable that some specific biases are necessary in order to constrain human

5 The reader is advised to return to this short explanation of the path of the thesis if at any moment he feels lost.

sentences understanding as Chomsky (1975) has noted. Thus, so the argument goes, language might need domain-specific modules that recognize important characteristics that makes it unique.

The second characteristic is that input systems are mandatory, they are obligatorily applied when they encounter their domain stimuli. Another example from language is that one cannot hear language as noise even if she tries. Also, Fodor argues there are few ways in which perception captures the world. In contrast, central processes are flexible and allow for a nearly unlimited amount of thoughts about the same object.

A third characteristic is that central processes (or at least conscious reports), have limited access to intermediate levels of inputs systems processing. That is, people have little or no conscious access and thus cannot report on how input systems might get the details of processing. That is why people have no intuitions about how the brain treats visual images. In contrast, the final consequences of input processing are perfectly understandable and reportable.

Input systems are fast and faster than thoughtful processes. This is a strange phenomenon since the sort of computations required to process the perceptual stimuli seem to be complex but input systems get them done in fractions of seconds. As Fodor (1983, p.63) puts is: “[...] the puzzle about input analysis is precisely that the computational complexity of the problem to be solved doesn't seem to predict the difficulty of solving it; or, rather, if it does, the difference between a 'hard' problem and an 'easy' one is measured not in months but in milliseconds”. In contrast, central processes can take months to decide on a chess problem.

A key feature of the Fodorian architecture is that input systems are informationally encapsulated. This means that there is a part of these systems which is blind to contextual information or knowledge information located elsewhere in cognition. Thus, there might be a module for color perception which does not take in suggestions from what subjects know about forests, that they are green. The famous case is the Muller-Lyer illusion, an illusion where two lines that are the same length look as if they are of different lengths. Measuring them and knowing that they are of different length does not help us see them as the same. Fodor (1983) argues some level of input analyzers are closed from other information in the system. He does concede that contextual information can bias perception, but he thinks this is because perception is not limited to input analyzers, so this bias can be applied by higher areas. Also, when it comes to priming language it

may be that lexical content is represented inside language analyzers, thus biases that come from related words may show up even if the system is blind to other contextual and knowledge information from the broader system. So there is a very specific sense in which information is encapsulated and it is not any sort of priming bias which can falsify Fodor's requirements. As he puts it:

> "[...] to demonstrate that sort of interaction between input analyses and background knowledge is not, in and of itself, tantamount to demonstrating the cognitive penetrability of the former; you need also to show that the locus of the top -down effect is internal to the input system. That is, you need to show that the information fed back interacts with interlevels of input-processing and not merely with the final results of such processing. The penetrability of a system is, by definition, its susceptibility to top-down effects at stages prior to its production of output." (FODOR, 1983, p.73).

Interestingly, information encapsulation is, in the Fodorian architecture, the essential feature of input systems and also that which answers the unity problem in Fodor's (1983) dual process theory. That is because other features can be explained as resulting from the suggestion that input systems are informationally encapsulated. This concept, however, does not work along with recent approaches in understanding modularity (see Anderson, 2014, Lupyan, 2015, Clark, 2016) and even dual process theorists themselves have been skeptical (see Stanovich, 2004). However, by letting go of this essential concept in Fodor's view, dual process theorists lose the answer to the unity problem and lack a philosophical base that guides theoretical coherence. This is the sort of base we attempt to provide for dual process theorists in this thesis. Although the focus is not in whether cognitive impenetrability has been demonstrated or not.

Moving onto Fodor's sixth feature we have that input systems have shallow outputs. Because input systems are said to be encapsulated, one can expect that their outputs will not be conceptually inflated. Thus, the output of a visual input system will not be theory-laden, that is, such output will not be in the form of "photons" for the physicist. Such conceptually inflated perception must be subsequent to input analyzers. Fodor (1983) explains in details what the output of such systems would be, he defines this class as basic categorization. Basic categorization is that which is in an intermediate level of categorization, that is, the output is not something like colors or shapes but a concept, although a concept in just the right level of abstraction. Thus, when one sees a dog, the basic categorization output is 'dog' rather than features such as 'skinny old' or more

abstract categorizations such as ‘mammal’. Fodor argues these basic categories are: more frequent in language use; are learned earlier; are the least abstract concept in the conceptual hierarchy it is placed; are used to teach children; will come faster to mind than other related concepts; are the favored ones when used to describe what we see or hear; appear to be phenomenologically salient and are concrete such that you could draw a dog but not draw something that is just an animal.

Seventh is the proposal that input systems are associated with fixed neural structures. “Neural architecture, I'm suggesting, is the natural concomitant of informational encapsulation” Fodor (1983, p.99). This seems to be a basic premise of neuropsychology, since for dissociations and double dissociations[6] to occur there must be some sort of regional-functional mapping. As we will briefly mention in chapter four, however, there are various ways of understanding the relations of functional and regional mapping (see Anderson 2014, for differences in functional differentiation and specialization). The eighth feature is also very related to the neuropsychological discussion, that these input systems are subject to very specific breakdown patterns. This is the claim that input systems are more prone, in lesions, to exhibit isolated function deficiencies. Thus, various neuropsychological patients might have specific lesions in the visual system but Fodor (1983) believes it is harder to find isolated lesions of central processes. The ninth and last characteristic of input systems is that they seem to develop in various children in a similar pace and following an innate path, sometimes present in infants, such as in vision, and in other cases following orderly steps of development in children, such as in language.

While the unity problem for input systems is solved by the postulation of information encapsulation, central systems are defined as global, in opposition to encapsulation. Such globality is better explained in terms of four very much related features described for central processes, namely, that central processes are domain general, horizontal, isotropic and Quinean.

Central processes are believed to be isotropic because elements relevant for a given thought might be drawn from any belief web in the system. Thus, when we think about atoms, we might relate them to solar systems. Furthermore, we might have used vision to see the image of an atom

6 These are the main successes of neuropsychology, where dissociation is when a lesion is found to disrupt function A but not B and double dissociation when another such lesion is found to disrupt B but not A.

in a paper and speak in words and sounds about solar system, thus these processes are domain general. That is, they can easily relate different modalities.

Central processes are also understood as horizontal because the same sort of computations might be applied for any sort of task, they were not tailored for any sort of task as were input systems and so they apply the same sort of principles to solve various sorts of tasks. Finally, they are Quinean because the value of one thought depends on the value of various other related thoughts in the web of belief. In this sense, how we judge the worthiness of the thought that animals are important depends or our views on morality, biology, culture, social status and so forth.

The two main features are that central processes are isotropic and Quinean and these are the ones that grant 'Fodor's first law of the nonexistence of cognitive science'. This 'law' states that the more processes are global, the less cognitive science will be able to understand them. In short, Fodor (1983, 2001) believes only encapsulated processes can be studied by cognitive science since the computational theory of mind cannot accommodate isotropic and Quinean features and the computational theory of mind, in his mind, is the only serious proposal cognitive science has.

Since Fodor's answer to the unity problem incurs in his 'first law' which abandons cognitive science (and puts nothing in its place), we will need to propose a different answer to the unity problem.

This answer will go roughly like this: If there is any sense in which input systems transfer its properties to Type 1 features and central processes transfers to Type 2 features, what Fodor took as computationally encapsulated input systems can be better accounted as the working of contextualized predictive processing and what Fodor took to be global central processes are actually capacity limited, classic computational, propositional processes that depend on (relevant) content from the latter processes. The globality phenomenon is not as global as Fodor has described but rather content limited by statistical correlations of previous occurrences.

We must, in advance, however, place an important caveat. We need to make it clear from start that, unlike Fodor, we will not propose a framework for the mind or any general solution to fundamental problems in cognitive science. What we propose is a framework for dual process theory of reasoning and decision making, and such is not a complete theory of the mind (and probably not even of reasoning). Thus, our proposal of an answer for the unity problem is for dual

process theory alone, it does not imply that the mind has a dual structure explained by a distinction between predictive and classical computations. In other words, we will not try to fix Fodor's dual process theory, rather, we will try to help dual process theories of reasoning by giving them (what Fodor did not) an answer to the unity problem compatible with the realm of cognitive science.

CHAPTER ONE:
RESEARCH IN REASONING AND RATIONALITY

1 Research in Reasoning and Rationality

In order to understand where dual process theories come from and how they are grounded on evidence of psychological tasks, in this chapter we will develop an analytic review of theories that attempt to account for empirical data on tasks of reasoning, judgment and decision making. We start by exposing the means by which we believe such theories should acknowledge evidence, and that is by accounting for data without the influence of evaluative normativism. Since rationality discussions have been developed in parallel, sometimes in confusion with, reasoning theories, we will also mention what sort of use we must make of rationality theories in reasoning. The experiments on reasoning, judgment and decision making are vast and hence we will concentrate in exposing three of the most famous tasks: the Wason selection task, the Linda conjunction problem, and the Harvard Medical School problem. Finally, we will review some of the most famous interpretations to these problems and theories developed for explaining them, by the heuristics and biases tradition, evolutionary psychologists, pragmatics and dual process theories. We agree, with what Samuels *et al*. (2002) has noted, that behind what seemed to be a polemic field, mostly because of what was termed 'the rationality wars', a gathering consensus in theories of reasoning and decision making was in fact emerging. We propose to show that the current form of this consensus is best captured by dual process theories.

1.1 Reasoning and rationality without evaluative normativism

The most famous methodology for evaluating a cognitive function in psychology is using tests, which are designed in the form of tasks that should have correct and false answers. Depending on which answers are given, and by the variation of tasks, psychologists believe they can determine some of the structure of cognitive processes. For example, they may use digit span to determine how many digits one can hold in working memory, and which processes they can execute while holding this information. In such case, the digits are given and psychologists expect the subject to repeat the given digits as requested. This is a clear case where there is one correct answer (the correct digits) while any other digit sequence would be incorrect. If correct, then psychologists

infer that working memory was successful, while if incorrect they can determine that working memory failed, and possibly why. This is the basic model of psychological research of cognitive processes. When it comes to reasoning and decision making however, determining which should be the correct answer for some task is not so simple.

Following Samuels' *et al*. (2004) classification, the reasoning and rationality agenda includes three projects. The descriptive[7] project, in which scientists mean to discover how people actually reason and to describe the cognitive processes that underlie human reasoning as it is. In contrast to what is researched in logic, the descriptive project should have little interest in which are the most valid forms of reasoning, or how to optimize argumentation. Rather, the project means to discover exactly what is happening in one's mind when he reasons in everyday situations or in laboratory tasks. The normative project has more of a philosophical aspect as it means to determine what it is to be rational. The goal of such project is to discuss on what grounds we should base our judgments of reasoning or rationality, to specify principles, standards and norms that people might follow in reason. Given such norms and principles, the evaluative project aims to empirically determine, on groups of people, which individual is reasoning correctly, or which is rational. Using tasks such as the descriptive project does and basing judgments on a norm from the normative project, the evaluative project then is supposed to have the tools necessary to determine the rationality of individuals.

For a psychologist to determine if a reasoning strategy is correct or false in his laboratory it is assumed he needs some normative standard from which to evaluate the value of the answer. There are various candidates for normative theories for rationality, but as we shall see, that is not an advantage, it is precisely the problem for evaluativism.

We argue that descriptive theories of reasoning need not worry about the evaluative project. The core idea is that by looking at the empirical results themselves, without determining which is correct or false, one can determine how people do in fact reason. The steps of the evaluative project should only be subsequent to an established empirical theory of reasoning. They are not necessary, they are only complementary to researchers who want to know more, or to determine more, or to

7 where descriptive contrasts with normative not with explanatory. Not to be confused with descriptive in a phenomenological sense.

make specific practical implications for their results. The normative project is necessary for planning tasks and as a possible guide as to how to read the results. As a guide it cannot determine which and how answers should be considered, only help indicate. Difference in task construal by subjects can also be a viewed as a result. If determining which individual gave right or wrong answers is unnecessary, then the need for a debate over difference in task construal is also unnecessary, as long as some people do give modal answers or some expected answer, then the fact that most did not give expected answers because of different task construal is a result that informs us about their cognitive mechanisms as well, not about the normative invalidity of tasks. Subject interpretation of the task is already a result about cognitive mechanisms.

Depending on what normative principle is chosen, on a classic task, different responses can be analyzed as rational, and there are good philosophical positions backing these principles. The main difficulty for an evaluative project is determining which, out of many candidates, should be the correct normative theory. One can use a general account for all tasks or different accounts depending on which task is analyzed. Either way, there seems to be no solid reason for choosing a normative principle over the other. Elqayam and Evans (2011) call this the arbitration problem. Some of the formal theories that are candidates for deriving normative principles are: classical logic, fuzzy logic, many values logic, other non-classical logic, decision theory, game theory, Bayesian probability, information theory or pragmatic accounts. Of course, some theories will not serve as norms for some tasks. But given that more than one theory can serve as a norm for one task, the arbitration problem seems to imply that there seems to be no simple way of determining which to choose. Furthermore, the answer to a task can be 'correct' or 'false' depending on which norm is applied. They are, therefore, essential if one needs an evaluative account of rationality.

After somehow deciding on which theory to base norms, one need also derive these norms in propositions that include 'should' or 'ought', since these formal theories by themselves are not normative principles. Stein (1996) for example, derives the conjunction principle from probability theory: One ought not to assign a lower degree of probability to the occurrence of event A than one does to the occurrence of A and some (distinct) event B. This is yet another difficulty, since as Samuels *et al*. (2004, p.37) note, "if normative principles of reasoning are not logically or

probabilistically derivable from formal theories, then in what sense are they derivable?". Which implies: what license is one using to decide which norms to derive from the chosen formal theories?

One strategy could be trying to determine which formal theories should be used on the basis of empirical evidence. Since people can have effective behaviors, and since these behaviors correspond to a given formal theory, then such formal theory could be the one through which we should evaluate rationality. Also, one can try to determine the most rational and correct response based on the answers intelligent individuals give. However, as Elqayam and Evans (2011) point out, by doing this, researchers are subject to the naturalistic fallacy, as conceived by Hume and Moore. The empirical data allows one to form propositions about how the world is, but there is no reason to believe one could use these propositions to derive conclusions about how the world ought to be. The normative ground is just as the ethical ground, empirical evidence might give us directions on these grounds, but the final word is always a philosophical commitment. And of course, to decide which person is more intelligent you already need some sort of previous evaluation.

A neutral approach for interpretation of evidence, that analyzes the frequency of responses and individual differences, is what we endorse. Instead of trying to determine, as is done in other psychological research, which is the correct or false response, what needs to be done is to explain why responses are given, and which processes are responsible for such responses. Not all possible responses are given in a uniform fashion. Suppose, for example, that in a task there are ten combinations of possible responses. What we tend to see is two or three more frequent combinations of responses. The job of the descriptive theorist is only to explain why these frequent responses happen. The fact that they are frequent means that there is some type of reasoning process behind it, or else these would simply be random choices like the seven other infrequent combinations in the supposed task. One could also try and understand the infrequent responses, but first one would need to show that they are not just random variations. In such attempt, normative theories could help. If few individuals are answering according to normative principles across various tasks, then the psychologist should want to consider their answers more than he would consider random or other infrequent answers. There is no point (for the description) to try and

determine one solution to a task as the best or correct choice, there is simply no descriptive gain for reasoning theories in doing so.

The usefulness in normativism for reasoning tasks is that it can help us outline the range of possible reasoning strategies our subjects could take without needing to evaluating them. The experimenter can contrast the answers in a given task with normative principles, as to help them understand what a subject was thinking during the execution of the task.

If rationality theories are used in a non-evaluative manner than the arbitration problem does not come as a burden. If we use these normative principle only to envision what subjects could be thinking by responding in a given way in a reasoning task than having multiple principles is actually positive. That is because none of these principles suggest the 'best way' to reason about the task at hand. They suggest possible ways to reason about a task and as so they serve as guides into understanding why a subject responded to the task in one or the other manner. It is important to be clear that we are not against evaluativism completely, we only stress that it cannot and should not be used in descriptive projects of human reasoning.

1.2 Classical reasoning, judgment and decision making tasks

The last half of the last century in reasoning, judgment and decision making research has been marked by a large list of puzzling experiments that started mainly with the works of Wason, Kahneman and Tversky. As examples, we will focus on three classic experiments and discuss their interpretations as well as some similar experiments which they have inspired: The Wason selection task; the Linda problem; and the base-rate Harvard Medical School problem.

1.2.1 Selection

The selection task is possibly the most famous experiment in this trend of reasoning research. In its original version, Wason (1966) presented students with four cards and told them that every card had a letter on one side and a number on the other side, and that either could be facing upwards. They were asked to decide which cards they would need to turn over to determine

if the experimenter was lying[8] in uttering this statement: If a card has a vowel on one side, then it has an even number on the other side. The four cards were: A vowel (P), a consonant (~P), an even number (Q) and an odd number (~Q). Following classical logic, the only two cards that need to be chosen are the vowel (P) and the odd number (~Q), since only these two can logically determine if the statement is true or false, this is seen as the normative response. However, most of the students choose the vowel (P) and the even number (Q). Interestingly, if they can see the two sides of the cards, most students accord with the normative response.

This task has been replicated various times and the modal response is always P and Q, typically less than 10% chose P and ~Q (Stanovich & West, 1998a). Wason's (1966) original interpretation of this was that subjects were irrationally choosing these cards because of a confirmation bias (which we will call a confirmation effect). A confirmation effect is a tendency to verify only the statement that one has in mind, instead of also trying to falsify it. When the evidence of falsification is ignored, one can believe the confirmation effect has occurred. However, Evans & Lynch (1973) have demonstrated that this does not always occur. They applied the selection task with a slight modification. Instead of having a positive consequent, the statement was presented with a negative consequent: "if there is a vowel on one side then there is not an even number on the other". If responses were following the confirmation effect, then the choices should be a vowel and a card which is not even. However, that does not occur, people choose cards that match the lexical content of the statement regardless of the negation, a vowel and the even card. This is evidence that in both cases (original task or negative modified task), they are not then trying to confirm the statement, but are rather responding to the lexical content itself. This was termed the matching bias by Evans & Lynch (1973) (which we will refer to as the matching effect).

Notice that while this is a good start as to why people choose these cards in this task, it by itself does not yet describe which cognitive mechanisms are responsible for these responses. Evans (1996), provided interesting evidence for processing in selection tasks. He tested subjects with a computerized version of the task and instructed subjects to keep the pointer on the card which they were inspecting. The computer recorded the amount of time each card was inspected before the

8 In recent versions experimenters have asked subjects to determine if the statement is 'true or false' or if 'the rule is being violated'.

choice. The results showed that people spent most of their time inspecting the cards they choose. They barely inspect the other non-selected cards. This is evidence that the cards which are not chosen are dismissed by tacit processes.

This original version of the task is called the abstract selection task. The results on it were quite puzzling by themselves; however, more similar tasks were conducted to further investigate interpretations to the selection task. These new results were not clarifying either. The most interesting modification to the task was the insertion of content, instead of using numbers and letters, a story was provided and the cards reflected actions the subjects should take. However, with the insertion of content to the selection task two distinct types of tasks were recognized. The indicative (or non-deontic) type and the deontic type. Cheng and Holyoak (1985) noticed that on the deontic type, a rule or regulation attempts to guide or govern action or behavior, it suggests what 'ought' or 'must' be done. On the other hand indicative conditionals have no rules to be obeyed or disobeyed, they indicate factual relationships that may be true of false.

The most famous deontic selection task problem was presented by Griggs and Cox (1982, p.415). It began with a story about the task: "imagine that you are a police officer on duty. It is your job to ensure that people conform to certain rules". The subjects were told the card presented information about people sitting at a table and that on one side of the card was a person's age and on the other side what they were drinking. The rule read "If a person is drinking beer, then the person must be over 19 years of age". So the subjects had to select the cards that they definitely needed to turn over in order to determine whether or not the people were violating the rule. The cards were: Drinking beer (P), Drinking coke (~P), 22 years of age (Q) and 16 years of age (~Q). This was termed the 'drinking-age problem'. The results were very interesting, 73% of the participants chose the P and ~Q cards. On their control group, none chose the P and ~Q cards on the original abstract selection task.

This led researchers to believe that content was providing a thematic facilitation of the task, and that subjects were now getting it 'right'. However, subsequent research showed that people did not choose the P and ~Q cards in various other tasks with content. In Stanovich and West (1998a) four of these indicatives tasks did not induce the P and ~Q response: "If 'Baltimore' is on one side of the ticket, then 'plane' is on the other side of the ticket." (destination problem); "If it's a 'USA'

postcard, then a ‘20c’ stamp is on the other side of the postcard”; “Whenever the menu has ‘fish’ on one side, ‘wine’ is on the other side”; “Every coin with ‘Madison’ on one side has ‘library’ on the other side”; but many of these individuals (31.5%) chose the P and ~Q on the drinking-age problem even without the introductory story. When adding an introductory story to the destination problem and to the drinking-age problem, choices of P and ~Q were of only 8% on the destination problem but of 85% on the drinking-age problem. These results show that content is not enough to make people choose the P and ~Q cards, it seems other peculiarities, or some specific type of content, determine results. We will discuss the interpretations of these results throughout the rest of the chapter, but let us first review results on the Linda problem and the base-rate Harvard Medical School problem.

1.2.2 Conjunction

The Linda problem was first developed by Kahneman, Slovic and Tversky (1982, p.91). These authors analyzed three groups of students, a statistically naive group with undergraduates from the University of British Columbia and Stanford with no background in probability or statistics; an intermediate group with students of psychology, education and medicine from Stanford who had basic training in statistics and probability; and a sophisticated group with graduate students in decision science of Stanford who had all taken advanced courses in probability and statistics. The original task consisted of two characters, Bill and Linda, however, the trends of study following this original research focused mostly on Linda, so we will follow tradition. The test consisted of a description of a woman and some choices about her work and hobbies, the subject had to decide the probability that each choice matched the description.

> “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. Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.
>
> ( ) Linda is a teacher in elementary school.
> ( ) Linda works in a bookstore and takes Yoga classes.
> ( ) Linda is active in the feminist movement. (F)
> ( ) Linda is a psychiatric social worker.
> ( ) Linda is a member of the League of Women Voters.
> ( ) Linda is a bank teller. (T)

( ) Linda is an insurance salesperson.
( ) Linda is a bank teller and is active in the feminist movement. (T & F)"

The description of Linda matched a stereotype of an active feminist (F) and not of a bank teller (T). But following formal probability, a conjunction of two events should not be more probable than any of the two individually. To affirm that one of these events individually is less probable than a conjunction of them has been termed the conjunction fallacy (but as Kahneman, Slovic and Tversky, 1982 originally did, we will refer to it as the conjunction effect). However, that rule does not correspond to the choice of individuals. 89% of the naive group, 90% of the intermediate group and 85% of the sophisticated group assigned 'T & F' as more probable than T or F individually. In all groups together, 85% of the subjects assigned F > (T & F) > T. Other (naive) subjects were also asked to compare only the two choices, F and 'T & F', even so, the conjunction effect did not disappear, 92% of subjects chose the compound target. Some from the intermediate group were interviewed about the task, and the majority stated they had selected the compound target because of a similarity or a typicality, but agreed, after some reflection, with the normative probabilistic response.

An interesting altered replication of the Linda problem was performed by Fiedler (1988) in three experiments. In sum, he wanted to test for alterations of two types: a) Simpler tasks first: if the conjunction effect would persist if first there was exposure of cases where everyone would agree that the normative conjunction principle applies; and b) Frequencies: if the conjunction effect would persist if the task was presented in terms of frequencies, as opposed to probabilities.

For the first type (a) of alteration, Fiedler (1988, p.125) presented simpler tasks first, such as: "Which event is more frequent: A massive flood somewhere in North America in 1987, in which more than 1000 people drown or an earthquake in California sometime in 1987, causing a flood in which more than 1000 people drown?". The original Linda problem was the last task. On another experiment, conjunction principles with Venn diagrams were exposed before the original Linda problem was introduced. For this 'simpler task first' alteration there was no significant change, the conjunction effect was preserved.

For the second type (b) of alteration, Fiedler (1988) presented the Linda problem first, but this time subjects had to consider 100 people that fit the description of Linda and to decide how

many of them should fit the eight occupation and hobbies from the original task. The alteration was: There are 100 persons who fit the description above (that is, Linda's). How many of them are: (followed by the original eight choices). In this case, the conjunction effect had a substantial decrease. Only 20% of subjects responded in a way that does not accord to the prescribed normativity, where in the original task this number varied from 70% to 80%. Thus, asking the question in frequency format has a very interesting effect.

Fiedler (1988) further investigated what would happen if the number of 100 subjects who fit Linda's descriptions changed to 168, an amount that is more uncommon and unnatural in our everyday thinking. Even with this last alteration the conjunction effect had no significant impact. The results seem to show that the difference in response comes not from the fact that the number 100 is easier to deal with, but rather from the exposure of the problem in terms of frequency.

1.2.3 Base-rate

The third set of classic experiments we will discuss has come in various forms, so we chose to highlight the one with the most impacting outcome. This test was administered by Casscells *et al*. (1978, p.999) to 60 people from Harvard Medical School. There were three groups of people, 20 were house officers, 20 were fourth-year medical students and 20 attending physicians. It consisted in answering a question about a simple statistics problem applied to a medical context:

"If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 per cent, 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?".

The results showed that only 11 participants (18%) chose the statistically normative response, which is '2%'. These 11 were almost equally distributed among the three groups. The modal answer was '95%', which results from considering only the last probability information given.

This is an example of the so-called base-rate neglect, since in this and various others studies (Kahneman *et al*., 1982; Stanovich, 1998d), people seem to be avoiding the first (general) statistical

information in favor of the second (specific). It seems to be a difficulty of reconciling and uniting two separate probabilistic pieces of information.

These results were replicated by Cosmides and Tooby (1996) when applying the exact same question to 25 undergraduate students of Stanford. However, they also wanted to know if one could eliminate the base-rate neglect if the problem was presented with several possible facilitation effects. They reworded the test in the following way:

> "1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive (i.e., the "true positive" rate is 100%). But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease (i.e., the "false positive" rate is 5%). Imagine that we have assembled a random sample of 1000 Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people. Given the information above: on average, How many people who test positive for the disease will actually have the disease? out of___" (Cosmides and Tooby, 1996, p.24).

This version included various attempts of facilitation: a) it shows the true positive rate, not just the false positive; b) the base-rates are given in frequencies as well as percentages; c) it explains about false positives; d) it includes the information that the sample was random; e) it specifies the size of the sample; and f) it asks for the answer in a frequency format, rather than that of a probability of a single-event. The results showed a large increase in the amount of answers that accorded to the normative standard. In their replication of the original problem 12% of the analyzed Stanford students accorded to the normative principle, whereas in the average of various reworded versions 76% did. The various facilitation attempts were controlled for, and interestingly, the most significant effect was presenting the information in frequencies only (not along with percentages) and asking for the answer in terms of frequencies. With frequency-aimed modifications they managed make responses accord to normative principles in up to 92%. They also found that the other facilitation effects without the frequency format only mildly produced the normative answer (36%).

Together, the results in the selection task, the Linda problem and on base-rate neglect show that in many cases most people do not accord to normative principles in solving simple problems, but with simple rewording of the same problems, most performances do accord with normative

principles. For the rest of this chapter, we will attempt to cover some of the interpretations that have been given for these results, bearing in mind that agents are minimal and that 'ought' accounts should not figure in our construction of descriptive theories of reasoning.

1.3 Interpretations of results

From here on we will try to cover most of the famous interpretations for these results and results alike. They have been taken as implying facts about our general rationality. But more importantly we are interested in what these experiments tell us about the cognitive processes of human beings. The main interpretations have been presented by the heuristics and biases tradition, evolutionary psychology, pragmatics and dual process theories. We will explain shortly what each proposes and we agree with Samuels *et al.* (2002) that the historic quarrel between heuristics and biases tradition and evolutionary psychology is mostly a communication problem, rather than a real dispute. We will propose further that, once the evaluative use of normativism is gone, it is possible to see that the dispute over the extent of human rationality really hides an emerging consensus in most interpretations about descriptive theories of reasoning, and that such consensus points towards dual process theories[9].

1.3.1 Heuristics and biases

The heuristics and bias tradition was mainly advocated by Daniel Kahneman and Amos Tversky. Kahneman *et al*. (1982) believed that the type of reasoning studies presented in the last section indicated that common people reason using basic heuristics that will sometimes get the needed results but will also make them prone to counter-normative responses. These heuristic principles reduce the complex tasks with which people are faced, and in general are quite useful, but since they process on data of limited validity, or because of this reduction done to the analysis of complex matters, they might lead to systematic errors, that we can understand as biases.

[9] We will review various interpretations of the evidence and how they *could* work together, but none of this implies they *must* work together, in fact some of these interpretations might be completely wrong. Thus we are not defending any specific claims or interpretations, rather showing how they hint at dual process theory.

Kahneman *et al*. (1982, pg.3) exemplifies this by means of a comparison of these types of judgments with the way in which our perceptual system processes visual information.

> "The apparent distance of an object is determined in part by its clarity. The more sharply the object is seen, the closer it appears to be. This rule has some validity, because in any given scene the more distant objects are seen less sharply than nearer objects. However, the reliance on this rule leads to systematic errors in the estimation of distance. Specifically, distances are often overestimated when visibility is poor because the contours of objects are blurred. On the other hand, distances are often underestimated when visibility is good because the objects are seen sharply. Thus, the reliance on clarity as an indication of distance leads to common biases. Such biases are also found in the intuitive judgment of probability."

Piattelli-Palmarini (1994), an heuristics and biases theorist who follows a pessimistic view on rationality, illustrates the meaning of a bias with the example of a lead and a feather. We seem to have a heuristic that associates the weight of an object with the impact it can cause when colliding. We can assume that one would rather have light objects over heavy objects falling on their heads. However, if the objects are a heavy feather and a less heavy lead than this heuristic would not work, since we would prefer that the feather fell in our heads because of its relational properties with the wind. Because of the bias that this heuristic might lead us to, we would even have the incorrect intuition that the feather must be lighter than the lead when in this case it in fact is not.

One heuristic which is repeatedly mentioned by these theorists is Representativeness. When subjects are asked to determine the likelihood that a given person fits a certain characteristic, or the probability that an object A belongs to the class B, they use the representativeness heuristic. This heuristic process works by assuming that A bears resemblance to B given some stereotype. If X is a stereotype, and the person is given tokens A, B and C, they will use a representativeness heuristic to determine which of these tokens best fit X based on resemblance. Kahneman *et al*. (1982) believe that this heuristic will work for various cases but will lead to biases since they do not incorporate more rigorous laws of probability.

The interpretation to the Linda problem is based on the representativeness heuristic. They believe the probability calculus implies the conjunction rule and therefore the correct answer to the Linda problem must be to give a smaller probability to the single event (Linda is a bank teller) than to the compound event (Linda is a bank teller and an active in feminist movements). However, given that people make these judgments based on the representativeness heuristic, and do not

follow all laws of probability, their answers will be biased. Since the similarity of an object to a stereotype can increase with the addition of more characteristics shared with the stereotype, then people will judge compound events as more probable than single events.

Because of cases such as the Harvard Medical School problem, Piattelli-Palmarini (1994) argues that humans are probability blind and suggests a general psychological principle: "Any probabilistic intuition by anyone not specifically tutored in probability calculus has a greater than 50 percent chance of being wrong." (p.132). According to the base neglect reading, it happens that we have a heuristic that makes us consider the last shown base-rate as the most important one to be considered. The bias is, as shown by modal responses in these types of tasks, ignoring previously mentioned base-rates.

Piattelli-Palmarini (1994) understands the modal response of the Wason selection task as an example of the confirmation effect. The bias is a consequence of having a confirmation heuristic, by which when people are convinced of a certain positive correlation they will attempt to find new confirmations of such correlation and shun evidence to the contrary. Although as mentioned earlier, the 'matching effect' seems to be a more favorable explanation and could be incorporated as a bias in their interpretation.

A stronger claim proponents of heuristics and biases tradition make is that given these various biases, the general rationality of human beings is inherently flawed. Piattelli-Palmarini (1994) believes these biases are cognitive illusions that human race is subjected to. In many cases we believe to be fully correct about the answer to a problem, however, our answers do not accord to normativity, so these are cases, in this view, where human beings clearly do not listen to reason, and such cases are natural and occur to everyone. In his book 'Inevitable Illusions' he sets out to try and teach us to cope with the "disorganized inventory of human nature" (p.2).

The general conclusion of the pessimistic heuristics and biases tradition is that the mind is composed of a number of simple heuristics which can work most of the time, but these experiments show how in fact our rationality is flawed, by showing how these heuristics are very prone to bias. So the goal of the project is to map these heuristics as to help learn about how to avoid (or to help others take advantage of) such biases.

Piattelli-Palmarini (1994) also responds to experiments that show how human reasoners can accord to normativity in altered conditions, such as the cases of certain content insertion in the Wason selection task and of describing probability problems in the form of frequencies rather than of percentages. He argues that those results change in nothing the fact that human rationality is inherently flawed, since in various situations we are prone to bias. He goes on to argue that since, of course, the heuristics are not flawed in all cases then the facilitation researchers are only discovering cases in which the said heuristics do work, while the biases are kept untouched in cases where heuristics do not work. The fact that heuristics do work in various situations is acknowledged by everyone including the pessimistic heuristics and bias tradition. Piattelli-Palmarini (1994, pg.192) argues that such a result "[…] in no way cancels, or reduces, the remarkable and worrisome fact that the illusions are present and vivid in other common situations […] previously tested [...]".

1.3.2 Evolutionary psychology

The main proponents of evolutionary psychology, Leda Cosmides and John Tooby, decided to rise up to the challenge of the claims that human irrationality is inherently flawed. Evolutionary psychology is a discipline in cognitive science that proposes a synthesis of such science with principles of evolutionary biology.

Evolutionary psychology (as summarized by Cosmides & Tooby, 2013) is a framework that proposes concepts for comprehending human cognition and motivation and adds new methodology to the discipline of psychology. They understand the brain as an organ with specific functions, just as any other evolved organ, whose main job is to process information as means of generating adaptive behavior. The brain was built by natural selection, which, according to their story, means that the design of its functions is linked with the goal of increasing the benefit of survival and reproduction. Although they recognize that our brains share the same history as various animals, for higher cognition, these scientists emphasize hunter-gatherer societies when humans had to deal with various environmental difficulties in order to survive. That is because, supposedly, relatively new brain functions evolved in response to the repeating elements of the 'Environment of Evolutionary Adaptedness' (EEA) (Tooby and Cosmides, 1990) and reflect its structure.

Tooby and Cosmides (1990, p.387) define the EEA not as a place or period in history but as "[…] a statistical composite of the adaptation-relevant properties of the ancestral environments encountered by members of ancestral populations, weighted by their frequency and fitness-consequences". Evolutionary psychologists (Cosmides and Tooby, 2013) inherit their ideas of the EEA from anthropology and biology and use evolutionary game theory as a method to discover adaptive information processing problems, as means of hypothesizing about what characteristics programs would need to have in order to solve them. Game theory is a method from economics for analyzing how agents behave in interactive situations. It models interactions between agents with defined decision rules that produce behavior. In evolutionary game theory decision rules that have successful outcomes will prevail in algorithmic selection. By looking at the consequences of various decision rules, one may hypothesize which is more likely to have been adopted in the EEA. With this in hand they believe they can have relevant testable hypothesis.

This part of the method shows that solutions depend greatly on which specific adaptive problem is faced. Because of this diverse problem solving structure the EEA had and since according to evolutionary principles the brain, being an organ, should reflect this structure, evolutionary psychologists believe the mind is composed of a collection of domain-specific problem solving mechanisms[10].

Cosmides and Tooby (2001) explain that adaptive problems have two defining characteristics: They are conditions that various individuals faced constantly in human evolutionary history and are problems whose solution increased reproduction of individual organisms or their relatives (because of a gene-centered view of evolution, see Dawkins, 1976). Some adaptive problems are computational. These can be defined as problems that can be solved by mechanisms that monitor some aspect of the environment and regulate the operation of other parts of the organism.

Cosmides and Tooby (2001) argue that to solve the variety of different adaptive problems humans must have developed at least as many domain-specific mechanisms as there are adaptive domains for successful behavior that make a difference. Evolutionary psychologists hope to

[10] This is one of the debatable proposals in evolutionary psychology, we do not mean to defend this but only expose it for discussion.

identify thousands of these domains, such as: parenting, food choice, mate choice, friendship maintenance, language acquisition, predator defense, sexual rivalry, status attainment and kin welfare[11].

Cosmides and Tooby (1994) intend that cognitive psychologists use the term function of a design in reference only to how it might have contributed to its own propagation in the EEA. They argue functions cannot refer to folk psychological ideas such as contributing to the individual's goals or to one´s well-being. Because natural selection is the designer of our brains, these other kinds of utilities exist as side-effects of other evolutionary functions. Such references of function cannot have explanatory utility as for why they have a specific organization.

According to Cosmides and Tooby (1994), domain-specific mechanisms are more likely to be selected in a natural process because of their intrinsic qualities. Since there is no need to have an engineered solution to competing tasks in the same mechanism, they display speed, reliability and efficiency. Different functions can be handled by different specialized mechanisms.

These mechanisms can have these intrinsic qualities also because they already know a great deal about the structure of their specific problems. In contrast to blank slate theories, evolutionary psychologists (Cosmides and Tooby, 2001) propose these mechanisms deal with problems already having solutions in advance. This would make them far more intelligent and efficient than systems that have no innate knowledge. As an example of evidence of such innate tendencies they refer to Johnson and Morton's (1991) research which shows how babies less than ten minutes old seem to have interest in face-like patterns but no interest in random versions of these patterns.

Evolutionary psychologists (Cosmides and Tooby 2013; Pinker 1997) argue that an architecture of multiple domain-specific mechanisms is both more consistent with an evolutionary realistic description and also more efficient at problem solving than a general purpose mind would be. Cosmides and Tooby (2001) explain that, in engineering, the same device is rarely effective at solving two different problems equally well. A saw is suited for trees but not for the kitchen, and a kitchen knife is useless in a tree. Therefore, it seems to be a basic phenomenon that a single

[11] Again, we do not mean to defend these strong thesis of evolutionary psychology, our purpose is only to explain how it works and in the end to see how it can relate to DPT. For different a perspective see Tomasello (2000).

general solution to two different adaptive problems will be inferior to specific solutions to each. A general purpose system, they argue, must necessarily sacrifice efficiency.

Cosmides and Tooby (1994) expose three main reasons for why a domain-general psychological architecture cannot guide behavior in ways that promote fitness. First, a problem is needed to define success, and adaptive problems have specific domains for success. So the first reason is that what counts as fit behavior differs from domain to domain, so there is no domain-general criterion of success or failure that could correlate with fitness. Second, general purpose mechanisms depend on perception and inference alone to reach goals. Cosmides and Tooby (1994) argue that adaptive problem solving cannot be deduced nor learned by general criteria, because they depend on statistical relationships between features of the environment, behavior, and fitness that emerge over many generations and are, therefore, not observable during a single lifetime. This is in line with Chomsky's (1975) poverty of stimulus argument which proposes our behaviors could not be accomplished if not accompanied by innate structures. Third, because they must evaluate all alternatives they can define, domain-general systems will run into combinatorial explosion when faced with real-world complexity.[12]

Pinker (1997) believes that concepts such as general intelligence and multipurpose learning strategies will become mere myths when scientists eventually map down all the domain-specific mechanisms that together were doing the actual work, making it seem like there was a general purpose system ready for use. Pinker (1997, pg. 27) believes such concepts "will surely go the way of protoplasm in biology and of earth, air, fire and water in physics".

General machines are dysfunctional and improbable to have been bought up by evolution, evolutionary psychologists argue. They admit domain specificity has limits for explaining all intelligent behavior, such as using novelty in a particular situation. For such reasons they theorize over how these many domain specific mechanisms could have limited information sharing, some form of network, or a scope delimiting marker (named scope syntax, see Cosmides and Tooby 2001) that would help on these problems. While they are careful enough to recognize such limits of domain-specificity, when it comes to explanation of phenomena, or understanding the evidence

[12] Of course, since we will be endorsing a domain-general mechanism, these are the sort of challenges we will have to face.

in reasoning tasks, they nonetheless evoke domain specific only processes, these are the processes that they emphasize, and that will feature in any more specific theory for explaining specific evidence. Concepts such as the scope syntax are more of a framework level explanation that do not feature in real research predictions.

Given this framework, evolutionary psychologists attempt to reinterpret data and develop new data to show that in fact the pessimism of the heuristics and biases tradition is false, that the mind is actually composed of "elegant designed mechanisms" (Cosmides and Tooby, 1996, p.18) built to solve problems as they were faced in our evolutionary history. A fundamental claim (Cosmides, 1989) of this interpretation is that reasoning procedures utilized in a task will vary according to the domain of the task. What is called content-effect for other theorists is evidence for evolutionary psychologist of such procedure use variation.

One of these hypothesized content domain is social contract. Social exchange is a very important evolutionary capacity for humans. Cosmides (1989) hypothesizes that there are innate social contract rules that follow a specific structured logic. An individual must usually pay a cost to receive a benefit. All participants in an exchange should follow the same cost-benefit rules in order for this to be a naturally selected attribute. So humans need to be able to avoid exchanges in which the cost exceeds the benefit.

Using evolutionary game theoretical procedures, Cosmides (1985) found that for social exchange to evolve in a specie one must be able to detect individuals who cheat, or fail to follow these established rules. So she defines a social contract as a relation of perceived benefits to perceived costs, where one needs to meet certain requirements to others in order to receive a benefit from them. Cheating is receiving the benefit without having achieved the agreed cost.

Following the social contract structure on a Wason selection task, irrespective of logical category, a cheater detection procedure should follow these choices: Choose the 'cost not paid' card and the 'benefit accepted' card and ignore the 'cost paid' card and the 'benefit not accepted card'. The reason is clear, only the first two could help detect potential cheating.

Cosmides (1989) analyzes two possible social contract rules that can be overlaid on the Wason selection task. On the standard social contract rule, participants need to investigate this rule

structure: If you take the benefit, then you pay the cost. The other possibility is the switched social contract rule which reads: If you pay the cost, then you take the benefit.

On the standard rule, the supposedly correct social contract behavior answer matches the normative logical answer: P (benefit accepted), ~P (benefit not accepted), Q (cost paid), ~Q (cost not paid). So the social contract correct answers would also be P & ~Q, since only they would help on cheater detection. The switched rule does not match normativity: P (cost paid), ~P (cost not paid), Q (benefit accepted), ~Q (benefit not accepted). While the logical answer would still be P & ~Q, the social contract answer would have to be ~P & Q.

Evolutionary psychologists understand that on the traditional abstract Wason selection task, people do not respond according to normativity since they have no special purpose mechanisms to derive such formal logic conclusion; that is not how people were designed to reason. They understand the drinking-age problem as a case of a Wason task with a standard social contract rule. That is, the rules for detecting the drinking cheater match exactly the normativity rule. Because we have the domain-specific cheating detection mechanism, the modal answers of subjects are P and ~Q on that task. Furthermore, they argue that not all content effect results in modal P and ~Q responses because not all content have this structure. For a content to induce the normative response, the domain-specific mechanism that deals with such content must have a structure that matches the abstract logical path. So they do not claim that social contract content boosts reasoning performance, rather they claim participants were just following a social rule that matches normativity by accident. They ground their position with evidence that in fact the modal response is not P & ~Q when following a switched social contract rule. On these cases participants favor the social contract response ~P & Q, rather than the normative one.

With that interpretation, Cosmides (1989) claims to have explained all previous research on the Wason selection task and to be able to predict when content will make participants match normativity.

Evolutionary psychologists (Cosmides and Tooby, 1996) claim also to have explained the data on the Linda problem and on base-rate neglect. The heuristics and biases conclusions were that people did not inherit a mental calculus of probability, and that they used heuristics (such as representativeness for the Linda problem) which often result in biased responses. Cosmides and

Tooby (1996) argue that there is not just one calculus of probability but various, depending on the chosen normativity, and whereas it was shown that subjects do not follow one of them, that does not imply they could not follow another.

Cosmides and Tooby (1996) explain how there are disagreements in the mathematical community about how to understand probability. As Cosmides and Tooby (1996) explain, Bayesians (the normativity used evaluatively by the heuristics and biases tradition) believe probability refers to a subjective degree of confidence that can be used to estimate the chances of a single event occurring. On the other hand, frequentists believe probability always refers to a history of events, so a single event, that is an event without history for comparison, cannot be attributed a probability.

Gigerenzer (1991) argued that the human intuitive mind represents probabilities in frequentist terms. He believes some of our innate mechanisms make us good intuitive statisticians of the frequentist school. The idea that we have mechanisms designed for dealing with frequencies is called the frequentist hypothesis. Cosmides and Tooby (1996) argue that our ancestors in the EEA did not have familiarity with data collection and information accumulation stored in a numerical fashion as we do today. Therefore, since we relied only on individual experience and at best other's opinions, the probability of a single event was impossible to be noticed. What was observable was not the percentage of chance of success in a hunt, but rather the amount of times the previous hunts were successful. Organisms could not evolve a cognitive mechanism which receives as input information that was not observable.

However, our ancestors could keep count of how many times one has been successful in a given endeavor considering all the attempts. One could use the information that 5 out of 20 hunts were successful, for instance, for the forthcoming hunt. So if this type of inductive reasoning was an adaptation advantage it is probable that it would have evolved in a frequency format.

Cosmides and Tooby (1996) argue that considering the probability of the next hunt as '25%' is not efficient because it needs an extra conversion which also would lose three benefiting properties of frequency format. First, '5 out of 20' says more than '25%' because it includes the number of events that the statistical information is based on, and therefore includes a degree of reliability. Second, the frequency format includes the possibility of increasing the data with the

addition of new observations, '5 out of 21'. And third, the frequency format is more flexible, it can easily adapt changes such as the specific success of each hunt, or the time when a few of them were executed in comparison to other ones[13].

As we have shown in the previous section, reconstructing the Linda problem or base-rate neglect problems in frequency formats facilitate responses that accord to normativity. Because of that, Cosmides and Tooby (1996) argue that we have special-purpose mechanisms for reasoning with frequencies and therefore do have a probability innate device, but fail in percentages and other formats because the input does not match the mechanism's requirements.

1.3.3 Pragmatics

Much of the effort of pragmatists (Cohen, 1981; Adler, 1984) studying these reasoning and decision making tasks have been of arguing that normative standards by which these tasks are measured are incorrect, therefore other theorists conclude wrongly that humans are irrational. While we agree that nothing in these studies show human irrationality or rationality, difference in task construals do not disqualify the evidence, they actually point to further evidence for descriptive theories as we will argue further in this chapter.

Despite the focus on normativism, there have also been some interesting explanations for specific tasks proposed by pragmatists (Adler, 1984; Sperber *et al.*, 1995). These are based on Gricean principles, mainly the maxim of relevance and the cooperation principles. We will first explain some of Sperber and Wilson's (1986) relevance theory, which is a version of Gricean pragmatics, and then we will show how these ideas propose interesting interpretations for the tasks.

Sperber and Wilson (1986) argue that on their model the important point of communication is that the audience perceives evidence of speakers intention through inferences. Through these inferences one can extract contextual assumptions called implicatures. These implicatures are based on the principle of cooperation which establishes that the communicator has the interest of

[13] You do not have to endorse the frequentist approach to understand or accept that frequencies are easier for humans to handle then percentages. One thing simply does not imply the other. It also says more about the types of inputs we handle better more than what it is claimed, that it suggests internal representations.

communicating his intentions in such fashion so that he uses the best possible words. Associated with this principle is the maxim of relevance, which will be soon discussed. These implicatures are not communicated through codes but by the fact that the speaker provides evidence for his intentions. For example, a son asks “Dad, can we play outside today?”, to which the father responds “It’s raining”. The father’s words do not directly answer the son’s question, but by following cooperation and relevance, the son is able to infer the proposition ‘we cannot play outside today’.

The communicator through stimuli tries to make manifest to his audience a set of assumptions. These assumptions become manifest depending on strengths. The strength is the confidence by which certain assumption is held, and depends on its processing history. For example, assumptions based on clear perceptions seem stronger than assumptions whose truth depends on the trust on the speaker. When an assumption is useful in a given situation it is strengthened, and when it is not it is weakened.

The strength of an assumption affects all three types of contextual effects mentioned by Sperber and Wilson (1986). The first is the derivations of new assumptions by means of contextual implications. Contextual implications are deductions that necessarily depend on new information present in some context and on older information. With only one of these elements no conclusion can be drawn, but with both of them, new information can be extracted. The second one is the strengthening of older assumptions, by means of new evidence that does so. The third is the elimination of an older assumption in favor of newer assumptions that contradict the previous.

Sperber and Wilson (1986) define relevance by its relation with contextual effects. For an assumption to be relevant in a given context, it must have contextual effects there. The strength of other assumptions in such context must be somehow affected. The irrelevant assumption might contribute with new information, but one that does not contribute to the present context. It might try to strengthen an assumption which is already held with certainty, and it might be too weak to affect other assumptions. They can however be relevant if they are part of a relevant behavior, such as changing the subject.

Sperber and Wilson (1986) believe the effort in processing also influences relevance. The effort increases with the increase in use of mental processes. The greater the effort to cause a contextual effect, the lesser the relevance will be. The authors posit a spectrum with various points

between the amount of contextual effects and the amount of effort required. On one end is the maximum contextual effect and minimum effort, and on the other, minimum contextual effect and maximum effort. Therefore, Sperber and Wilson (1986, p.125) define two conditions: 1) “an assumption is relevant in a context to the extent that its contextual effects in this context are large”; 2) “an assumption is relevant in a context to the extent that the effort required to process it in this context is small”.

Sperber and Wilson (1986, p.158) believe both the speaker and the listener to presuppose optimal relevance in communication. They define this presumption: 1) “The set of assumptions I which the communicator intends to make manifest to the addressee is relevant enough to make it worth the addressee's while to process the ostensive stimulus.” 2) “The ostensive stimulus is the most relevant one the communicator could have used to communicate I”. So Sperber and Wilson’s (1986) principle of relevance is the thesis that every act of ostensive communication informs the presumption of its own optimal relevance.

Sperber *et al*. (1995) have shown how relevance theory explains previous results on the selection task. They show how modal answering follows relevance judgments. They believe the choices of P and Q in selection tasks, which are modal, are more relevant than the choices of P and ~Q. In tasks where P and ~Q are the modal response, the focus of relevance is on such answers.

They explain what happens on the traditional abstract selection task based on the effort and effect side of relevance. On the effort side, the normative response to the task requires subjects to consider negations, while the P and Q response do not. Difficulty in negations have been previously demonstrated (Horn, 1989; Wason, 1959). So P and ~Q can be overlooked simply because it is tougher to process but also because its effort will lead to a conclusion of irrelevance.

On the effect side, general conditional statements make subjects expect that given P, Q will occur, so that they should search for instances of Q, in the absence of evidence pointing in other direction. They will search for the content that has been communicated to them through the rule and follow the cooperative principle. The reading of the rule leads to the search of P and Q cases. This is also in line with the ‘matching effect’ explanation.

Sperber *et al*. (1995), based on relevance theory, have managed to predict which selection tasks will make subjects mainly respond P and Q and which will make subjects respond P and ~Q.

They have, therefore, developed a recipe for elaborating ‘easy’ or ‘hard’ selection tasks. To build an easy selection task one must: 1) select a pair of simple features P and Q such that P and ~Q can be represented with little effort, or less effort than P and Q; 2) Create a context where knowing P and ~Q cases can have greater contextual implications than knowing P and Q. 3) Present the rule in a pragmatically felicitous manner, which basically means making it accessible, realistic and interesting.

Sperber *et al*. (1995), showed through a series of experiments on four types of selection tasks that they could control for answers of P/Q and P/~Q. Let us present at least one, the virgin-mothers problem. Such problem was contrasted with results in the traditional tasks in a within-subject setting. The introductory text reads:

> “Until recently, it was obvious that a woman who has children has had sex. With artificial insemination, it is now possible for a virgin to have children. The leader of the Hare Mantra (a very secret religious, Californian sect) has been accused of having had some of his sect's virgin girls artificially inseminated. His goal, it is claimed, is to create an elite of "Virgin-Mothers" alleged to play a particular role in his religion. The head of the Hare Mantra makes a joke out of these allegations. He claims that the women of his sect are, without exception, like any other women: if a woman has a child, she has had sex. Imagine that you are a journalist and that you are preparing an article on the Hare Mantra. You learn that a gynecological survey has been carried out among the Hare Mantra women. Some of them might be "Virgin Mothers". You go and visit the doctor who carried out the gynecological survey. Unfortunately, the doctor pleads professional secrecy and refuses to tell you what he discovered. You realize that, before you, on the doctor's desk, there are four individual information cards about Hare Mantra women examined in the gynecological survey. However, these four cards are partially concealed by other papers (as shown below). Of two cards, one can only see the top where you can read whether the woman has children or not. Of the other two cards, you can only see the bottom where you see whether the woman has had sex or not. You are determined to take advantage of a moment in which the doctor turns his back to uncover the papers and to learn more. Indicate (by circling them) those cards that you should uncover in order to find out whether what the leader of the Hare Mantra says ("if a woman has a child, she has had sex") is true, as far as these four women are concerned, indicate only those cards that it would be absolutely necessary to uncover.” (SPERBER *et al*, 1995, p.63).

The choices of card were: Children – yes (P), Children – no (~P), Sex – yes (Q), Sex – no (~Q). Since the focus of relevance was on women did not have sex and who had children, the choices of P and ~Q were made by 78% of the subjects. While only 26% of the same subjects on the abstract version of the task chose P and ~Q.

By finding various facilitation effects based on relevance, Sperber *et al.* (1995) forces evolutionary psychologists to find Darwinian algorithms for each of these ‘easy’ selection task they

can formulate. This scenario would make their case rather implausible. However, this does not completely exclude the possibility that there could be both relevance effects and Darwinian algorithms governing task performance (Cosmides and Tooby, 2000).

Adler (1994) also tries to reinterpret tasks by means of pragmatics, basically by analyzing how experimenters cue for relevance, how subjects expect them to be informative and to engage in communicative cooperation. He focuses on the Linda problem and a base-rate problem. The base-rate problem is the engineers and lawyers problem from Kahneman *et al*. (1982), which is much like the Harvard Medical School problem, but, like the Linda problem, is based on the description of a character and the chance that such character fits certain characteristics.

To understand these tasks, Adler (1994) proposes that when one engages in abstraction she loses sight of relevant properties that might be in a context. He argues that abstractions tend to be uncooperative, since there always will be a more pragmatic reading to a communication event than that of taking essential properties and ignoring others. He argues that in these tasks, to respond according to normativity, subjects must reject as irrelevant portions of the actual contribution that normally appear appropriate, so that abstraction sacrifices some of the informativeness of their contribution.

For instance, the description and the task of rating the probability of Linda having each property has to be ignored in order to give the normative response that it is less likely that Linda is a bank teller and a feminist. Because of that, Adler (1994) suggests that the actual task (respecting conjunction principles) is really hidden in the background, and subjects will follow what seems to be expected of them by means of relevance and cooperation, that is that they think of Linda as a feminist. Samuels *et al.* (2004) argue that "Linda is a bank teller who is not a feminist" has been controlled for, and hence there seem to be no linguistic factors of relevance involved. But Adler (1994) means to point out also that subjects are following what seems to be the exposed, rather than the hidden, interest of the experiments. If they were to abstract away from the exposed interest they would violate the cooperation principle.

The same type of explanation is given for the engineers and lawyers base-rate problem. To be cooperative subjects focus on the description of the character rather than dedicating their effort at calculating probabilities by means of base-rates. Adler (1994) argues that the description of the

character is the experimenter's most salient contribution, while the base-rate data are part of the background information. However, there is evidence these normative difficulties are not only because of an attempt to read the experimenter's focus. Since on the Harvard Medical School problem, the information is given in a relevant and direct manner to the physicians. Also, Adler's distinction between abstraction and relevance does not explain why frequencies facilitate normative responding, as Samuels et al (2004) already noted, and they facilitate both in the conjunction problems and base-rate problems. If Adler's (1994) reading was the complete story, relevant information should also inhibit normative responding in frequency tests, but that is not the case.

Anyhow, Adler (1994) seems to have a noble goal in mind, which is to argue that domain-independent abstractions are not always the ideal against which all cognitive success should be measured by. That the relative costs and benefits between being guided by relevance or abstracting away need to be taken in consideration. In a way he is also arguing against making an evaluative use of normativism for reasoning tasks. On the other hand, Sperber and Wilson (1995) really make a distinctive contribution, from the pragmatics camp, to descriptive theories of reasoning.

1.3.4 The hidden emerging consensus

The initial accepted understanding of the field was that the heuristics and biases and the evolutionary psychology ideas were really at odds with one another. However, Samuels *et al.* (2002) point out that behind the collection of rhetorical excesses of both traditions, a hidden consensus based on their real core claims was emerging. They argue that once one eliminates such rhetorical excesses on both sides, the challenge of evolutionary psychology is really no challenge at all. Moreover, that the claims of evolutionary psychology could not make sense if not by endorsing the core claims of heuristics and biases tradition. An example of a rhetorical excess of the heuristics and biases tradition is claiming our species is probability-blind. Whereas an excess by the evolutionary psychologists would be claiming that we have 'elegant mechanisms' designed to solve our problems.

Possibly one of the main reasons for the apparent dispute has emerged from the different focus of theses research projects. While the heuristics and biases tradition has focused on finding experiments in which subjects do not accord to normativity and on explaining these 'failures', evolutionary psychology has been concerned in showing the reasons behind our 'success' in various tasks.

Therefore, perhaps because of methodology, the heuristics and biases tradition were showing cases where our mechanisms do fail to accord to normativity. Evolutionary psychology must agree that these mechanisms do fail, in order to propose cases where they do not fail. Both traditions agree that we have mental processes built by evolution to solve adaptation problems and that when facing the problems of today, these mechanisms might fail in various tasks but succeed in many others.

For instance, because of the evidence, heuristics and biases agree that frequency format or specific content will result in the success of participants because that is when these mechanisms are truly working. On the other hand, evolutionary psychology will also contend that in various cases people do not accord to normativity because the mechanisms are not designed to deal with such. It is important to notice that in both cases the total system will yield many mistakes although it will also result in correct answers.

The heuristics and biases tradition also does seem to contend that the mind is composed of a group of different specialized mechanisms. Kahneman *et al*. (1982, p.88) compare problem solving "with the operation of a flexible computer program that incorporates a variety of potentially useful subroutines". They also argue that "the actual reasoning process is schema-bound or content-bound so that different operations or inferential rules are available in different contexts", and that "consequently, human reasoning cannot be adequately described in terms of content-independent formal rules." (Kahneman and Tversky 1982, p.499).

The same can be found in the other way around, evolutionary psychologists (Cosmides and Tooby, 2001, p. 162) can also be found endorsing heuristics and biases claims:

> "A male robin red breast may not look particularly intelligent when overcoming obstacles to attack a tuft of red features; nor does a human male when he spends time looking for pornographic pictures rather than courting actual women […] These mechanisms lead to such odd outcomes because there are things in the world other than rival robins and living women that satisfy the input conditions for the monitoring devices

> employed by the computational systems that (respectively) regulate aggression in robins and courtship in humans".

These apparent differences come because of making an evaluative use of normativism, not because of major differences in descriptive theories. Piattelli-Palmarini (1994) even admits a subscription to the modularity of mind for reasoning tasks, but argues (p.2) that:

> "This cognitive unconscious is a trait that [...] may have saved our ancient ancestors from wild beasts and famine, but even granted that such a simpleminded Darwinism should have actually been at work in their creation, for a very long time these illusions have been no more than a burden. Darwinism or not, we should all learn how to protect ourselves, individually and collectively, from the effects of our cognitive subconscious".

Perhaps not so surprisingly, of the three points Samuels *et al*. (2002) point out as the real disputes between these traditions, two are problems that stem from making an evaluative use of normativity.

The first is how to best apply probability in given tasks, which norm should be used in evaluating responses in specific probability tasks is not agreed on. However, if both agree that in some cases people fail to reason according to some norms and in others people do succeed, what difference does determining which is the best norm make for descriptive psychology? The answers to the problems stay the same regardless of the way you choose to evaluate correctness. Multiple normative standards can be used to understand what subjects were thinking.

The second problem noted in using normativism in an evaluative fashion is even further away from the realms of descriptive theories, it is a real dispute about which the correct interpretation of probability theory is, a Bayesian version or a frequentist version. While this is a real dispute, it is not for psychology to determine.

The third is a real descriptive theory incompatibility that points to others. Evolutionary psychology argues that their opponents propose limited explanations, they believe a better account of specific reasoning mechanisms need to be described, rather than just heuristics like representativeness or availability. In fact this seems like a good critique, but perhaps the main characteristic of heuristics and biases was in identifying the biases, not the reasons why.

However, we believe this leads to some further different incompatibilities that add to what Samuels *et al*. (2002) pointed out. For instance, it is not convincing that proponents of heuristics and biases would buy the story of social contract theory. While heuristics and biases and

evolutionary psychology as general level frameworks do not differ too much, many differences might lie in specific theories for explaining specific phenomenon. However, these are differences that theorists in the same general framework tend to encounter and dispute over. It is a type of difference that Cosmides and Tooby could have with other evolutionary psychologists, not a core difference between traditions.

It is also important to note that even though Piattelli-Palmarini did claim these heuristics were designed by natural selection, he was a critic of adaptationism. Recently, he (Fodor and Piattelli-Palmarini, 2010) even attacked the idea of natural selection directly. However, this should not count as differences from the two reasoning traditions since these recent ideas are not taken very seriously and have very few supporters, they are not tied the heuristics and biases tradition.

This does not however, change the fact that as Samuels *et al.* (2002) pointed out, there is an emerging consensus on what these reasoning, judgment and decision making tasks tell us, namely that the mind is, in great part, sensitive to differences in type of input and that themes relevant to adaptation and social practice are favored over normativity. Mind mechanisms are efficient in dealing with what they are trained to work on, but might be responsible for unexpected answers in some tasks which are not in the form which they were designed to deal with. For these reasons, sometimes modal answers will accord to the expected normativity, while in others, other problem structures will dominate their processing choices. This consensus also includes the pragmatist collaborations we mentioned.

However, even granted that such emerging consensus is real, it seems some descriptive part is still missing, namely the cognitive mechanisms behind the answers that accord to normativity. Although modal responses to the original Wason selection task, original Linda problem and original base-rate problems do not accord to normativity, what are the cognitive mechanisms that explain why at least some people do respond in accordance with normativity?

We believe that the next interpretation, given by dual process theories is a version of this emerging consensus, which synthesizes what each previous tradition best showed and also provides an explanation for the mechanisms behind normative responses.

1.3.5 Dual process theories

As we have previously argued, making a non-evaluative use of rationality can be an important part of interpretation of reasoning tasks because it offers the theorist a range of behaviors and reasoning steps that could be expected from the participant on a task. Evolutionary psychology focuses on ecological rationality psychologists and it is expected that individuals will answer accordingly. However, as is consensus in all camps, what is evolutionary valid might not be the best choice in the modern world (See Stanovich, 1999, 2004; Cosmides and Tooby, 2001, Piattelli-Palmarini, 1994). Choosing fat food over non-fat food was a good strategy for the hunter life, but not for our modern life. Evolutionary psychologists do notice this but they do not develop theories to account for it. While normative rationality cannot be used to solve all our problems, for instance, since it would be too irrelevant for evolutionary situations, using normative rationality might be important when making a life-long economic decision for instance.

Now if we expect both ecological rationality and normative rationality from the participants in a task, it is plausible that they also might have different task construals. However, as previously mentioned, difference in task construal is not only a normative matter, it tells us about mechanisms people are using in reasoning. Stanovich (1999) noticed that to have different task construals there might be different mechanisms which provide these construals. That is why we must take all modal answers and all consistent answers (by the same individual through various tasks) as evidence for reasoning processes, irrespectively of if any given task construal is the preferable according to the theorist's evaluative use of normativism.

The best way dual process theories can account for the evidence on reasoning is by considering all modal answers and consistent answers (by the same individual through various tasks) as answers that tell us about people's underlying cognitive mechanism. So, on the original Wason task, there seems to be a pragmatic choice (P & Q), and there is also a normative choice (P & ~Q). On the Linda problem, there is an ecological or pragmatic response (Linda as bank teller and feminist activist is more probable), and also a normative choice (Linda as a bank teller alone is more probable). On the Harvard Medical School problem we have a heuristic response (that ignores prior probability) and we have a normative response (that considers both probabilistic information). As Stanovich (1999) noticed, these different possibilities of task construal hint at the

mechanisms that support them. In general, in these types of tasks studied, we seem to have either an intuitive response or a normative response. In the cases of facilitation, the task fails to show us the conflict between these two different types of responses. So in such cases, what happens is that evolutionary or pragmatic rationality and normative rationality coincide in the same answer; the answer is the same, irrespective of which one an individual follows.

One could need different concepts of rationality to explain modal answers or consistent answers (by the same individual through various tasks), that do not fit any of these rationality concepts. However, it seems this has not been the case in the history of these studies. It must also be acknowledged, at least for descriptive reasons, that we have no need to consider evolutionary or pragmatic rationality or normative rationality as the optimal choice for any given task. Why should it matter, for descriptive reasons, if people should not have to consider the probability of single events, as Cosmides and Tooby (1996) argue, if at least some can and do attempt to think in such a way?

Of course, the number of individuals that give normative responses in these types of task is low. That is the main reason why these individuals have been ignored, and not considered part of the evidence a theory should explain. However, through individual differences Stanovich and West (1998a, 1998b, 1998c, 1998d; Frederick 2005 and also Toplak *et al*., 2011) show that normative responses are given not only as a random choice but as a normative choice, because they are given by the same individuals (who also score high at analytic thinking) across various tasks.

Stanovich and West (1998a) tested participants in various versions of the Wason selection task, in three studies. Some of these tasks were non-deontic (normally associated with non-normative responding) and one of them was the drinking-age problem (a deontic problem in which modal answers are linked to normative responding). They found that individuals that respond normatively on one non-deontic task tend to do so across the other tasks. Also, individuals tested with high cognitive ability tend to respond normatively across tasks. Interestingly, the individuals that did not answer normatively in the drinking-age problem had worse results in cognitive ability testing. They had, also, similar cognitive ability results to participants that displayed normative answers on the drinking-age problem but did not on any other. Those who exhibited the matching effect on other tasks also had worse results in cognitive ability testing.

Stanovich and West (1998b) replicated the original Linda problem, and 80.7% of their sample displayed the conjunction effect. However, the 29 subjects who responded normatively had also tested higher on cognitive ability. The correlation between cognitive ability and normatively responding had a large effect size.

Stanovich and West (1998d) got mixed results on the importance of cognitive ability in base-rate. When studying base-rates applied in a selection task, cognitive ability did not predict normative answers (considering base-rate as relevant). However, individuals who did not include prior probabilities in two tasks (David problem and the Mark problem) which are similar to the original base-rate problems (the engineers and lawyers problem in Kahneman *et al*.,1982 and Harvard Medical School problem) had lower cognitive ability than individuals responding normatively.

Toplak *et al*. (2011) tested 346 individuals in cognitive ability, syllogistic reasoning, working memory, cognitive reflection and 15 diverse heuristics and biases tasks among other measures. They found a .41 correlation of cognitive ability with normative responding in heuristics and biases and syllogistic reasoning tasks. When summing up working memory and cognitive ability the correlation was of .47. Again, the consistent coherence through tasks and cognitive ability in normative respondents show the necessity of explaining not just modal answers but the cognitive mechanisms used by normative responders. Normative responding is not just a random variance but a rare (but existent), justified response, which should serve as evidence for descriptive theories of reasoning. Thus, this use of cognitive ability is to determine that some people are answering normatively consistently (in contrast to a random choosing), not to determine that therefore the normative answer is more appropriate. Even more interesting was the correlation of .49 (that is, moderate to large) of cognitive reflection task with syllogistic and heuristics and biases tasks. Although the cognitive reflection task also correlates with cognitive ability and working memory, regression analysis showed that the cognitive reflection task was a unique predictor of normative responding. This is interesting for the Dual Process Theory (DPT) interpretation for the reasons will we argue for.

The cognitive reflection task was developed by Frederick (2005) as a task that would tap an intuitive answer that came quickly to mind, but one which is normatively awkward. While we

seem to have intuitive answers for some simple problems, there are others for which we have no intuition at all, such as the square root of 1897. To figure out the correct answer to the cognitive reflection task subjects would need to think a while longer. The task is very simple, it consists of three simple questions with relatively easy solutions, but to which our intuitive responses do not correspond. The introduction to the task (which would come along with other tasks) read:

> "Below are several problems that vary in difficulty. Try to answer as many as you can":
> "(1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
> How much does the ball cost? _____ cents
> (2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take
> 100 machines to make 100 widgets? _____ minutes
> (3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size.
> If it takes 48 days for the patch to cover the entire lake, how long would it
> take for the patch to cover half of the lake? _____ days" (FREDERICK, 2005, p.26)

The intuitive answers are 10 cents to the first problem, 100 widgets to the second and 24 days to the third. However, the careful solution of the problems easily shows that the answers are 5 cents, 5 minutes and 47 days. An amazing virtue of this task is that there can be no dispute over normativity and rationality, any researcher must agree that the task is posed as basic mathematics. This lessens the possibility of divergent interpretations by reducing the rationality standards that we suspect the subject to be using.

It is possible to verify that there is an intuitive answer and a careful answer because of a few factors. An intuitive answer is one which comes to people without much effort and tends to be modal, and in fact 10 cents, 100 widgets and 24 days are highly modal answers. Even those who respond normatively report having first reached the intuitive answers, this is confirmed by verbal report, notes taken while solving the problem and introspection. People who responded 10, 100 and 24 also believed that the task was easier than those who gave the normative answer, the former estimated that more than 90% would get the correct answer, while the latter estimated that only 62% would (both were overestimates). Also, similar tasks that do not tap intuitive answers forces people to try and respond normatively, there is no effortless and modal awkward response in these cases.

These differences between an intuitive quick responding and normative slow responding have been shown from all these reasoning, judgment and decision making tasks throughout the years. Because of that, theorists have been gradually developing dual process theories, theories

which posit a difference in two types of processing: an intuitive implicit form of processing and a thought-out explicit form. However, these theories come in various forms (Chaiken, 1980; Evans And Over, 1996; Fodor, 1983; Nisbett et al., 2001; Kahneman, 2011; Sloman, 1996; Stanovich, 2004). In the next chapter we will discuss if we can use a general DPT or if a specific one is always needed. Since we have not yet discussed this matter, for now we will just use a generic version. Evans (2008) reviewed several of these theories, even from different areas of research (like social cognition and evolution) to point out a common list of features that are usually ascribed to these two processing types.

According to this multi-theoretical cluster of attributes for each type of processing, the correlational features for Type 1 processes are: unconscious, implicit, automatic, low effort, rapid, high capacity, default, holistic, perceptual, evolutionary old, follows evolutionary rationality, shared with animals, nonverbal, modular, associative, domain-specific, contextualized, pragmatic, parallel, stereotypical, independent of general intelligence, independent of working memory. The correlational features for Type 2 processes are: conscious, explicit, controlled, high effort, slow, low capacity, inhibitory, analytic, reflective, evolutionarily recent, follows individual rationality, uniquely human, linked to language, fluid intelligence, rule based, domain general, abstract, logical, sequential, egalitarian, heritable, linked to general intelligence, limited by working memory capacity. Evans (2008) noted that positing all of these features as defining characteristics of these types of processing is troublesome, because these characteristics will not always stand. In the next chapter we will examine all of these theoretical details of dual process theories. Therefore, for now, let us consider a generic DPT with correlational features for each type, but with one fundamental difference: that Type 2 processes will aim at normative responding and Type 1 will aim at intuitive responses (be it evolutionary, heuristic or pragmatic).

This generic DPT can account for the hidden emerging consensus and for the individual differences literature. DPT accounts for the emerging consensus proposing that the mind is mostly composed of Type 1 processes which have helped humans respond to adaptive problems but that might find difficulties in the modern world. These rapid, contextually fine-tuned processes are efficient in dealing with their input, but might be prone to mistakes in tasks with different designs. For these reasons, when Type 1 processing is well suited, modal answers will accord to what

theorists expect the normative answer to be. However when these Type 1 processes are not able to solve a given task, Type 2 processes can override and aim at a normative answer. It goes further than the hidden consensus in explaining the evidence, because it also succeeds in explaining the cognitive mechanisms that govern choices that individuals who respond with normative answers. It is predicted that people with higher general intelligence, who are skilled at analytic intelligence, and who can use cognitive reflection will tend to override intuitive answers to aim at normative answer by using Type 2 processes. This does not mean that Type 1 processes will always be responsible for what the heuristics and biases tradition calls biases, Type 2 processes can result in non normative answers as well, but they will differ from Type 1 responses, in that they are not modal, intuitive, and are usually mistakes one would encounter when aiming at normative responding (such as making a calculation mistake in the widgets answer on the CRT).

We believe the CRT makes a pretty good case for DPT because in it, it seems clear that there are two types of possible responses, one intuitive Type 1 response and the other slow and reflective Type 2 response, the normative and intuitive answers are also easily agreed upon. The best part of this, for the DPT is that individuals who are able to give Type 2 responses in the CRT will tend to give what theorists believe to be Type 2 responses across several reasoning and heuristics and biases tasks. And individuals who respond with the intuitive Type 1 mechanisms to the CRT will tend to respond accordingly on the heuristics and biases task. Frederick (2005) and Toplak *et al. (*2011) argue that the CRT can therefore be seen as a task that measures the tendencies of individuals to override Type 1 answers and engage in Type 2 processing.

One could ask why different types of rationalities, such as pragmatic and evolutionary rationality, should not also be evidence of different types of processing themselves or how we one can know that they are not actually part of Type 2 processing. There are two ways to determine how a rationality aim will be fitting each type of processing. The first is by how they relate to correlational features of each Type of processing. So evolutionary, pragmatic and heuristics strategies all tend to have various Type 1 correlational characteristics, such as: being unconscious, intuitive, implicit, automatic, low effort, rapid, default, evolutionary old, modular, associative, domain-specific, contextualized, stereotypical, independent of general intelligence, independent of working memory. In that sense we can group those strategies as Type 1 strategies. The second way

is by investigating individual differences. One could suspect one of these three rationality strategies or other possible imaginable ones to be the focus of Type 2 processing if individuals who do well on the CRT tend to prefer them across various other tasks.

This is a general framework that neatly suggests differences in cognitive mechanisms exposed by the evidence (usually described as puzzling) in reasoning, judgment and decision making literature. As we have seen in this chapter, this can simply be done without the need to appeal to the use of evaluative normativism. We are not arguing that dual process theorists have been free of an evaluative use of rationality all these years, only that their framework *can* provide the most complete account of these tasks without making use of evaluative normativism.

Our claim that DPT can be seen as the result of the hidden emerging consensus is also backed by the fact that main theorist from the heuristics and biases tradition has admitted to have something like that in mind before and have recently proposed a DPT (Kahneman, 2011). Further, Sperber, the main advocate of pragmatics involved in the discussions of reasoning has also subscribed to a similar view (see Mercier and Sperber, 2009), it is a view that is continuously growing.

We can also point out how the evidence has been suggesting DPT throughout the years. Evans' (1996) computerized version of the selection task hints at DPT by showing how modal Type 1 responses are quick and intuitive, since subjects tend not to evaluate other cards, they spend time evaluating the cards they have quickly assured to be the relevant ones. If there are Type 2 processes involved in this type of answering, it is only after Type 1 responses have already determined relevant cards.

The Type 2 assumption for normative responding is supported by the empirically demonstrated connection between cognitive ability and the tendency to engage in analytic processing (Ackerman & Heggestad, 1997; Cacioppo *et al*., 1996).

Frederick (2005) also found that subjects that do well on the CRT, that therefore have tendencies to engage in Type 2 processes, tend to make more patient choices and to prefer receiving more money by waiting than less by receiving readily accessible money, when short periods are analyzed. They will also be more prone to engage in possible risk taking in economic decisions.

Evans and Over (1996) understand Nisbett and Wilson's (1977) evidence that there may be little or no direct introspective access to higher order cognitive processes as a result that shows conflict between Type 1 processes and Type 2 theorizing about them. They also point out an association between using Type 2 processes for problem solving and tiring out the individual, which shows its correlation with need for working memory and concentration. Also, there is a correlation of normative responding with thinking aloud, using storing strategies such as offloading, and use of verbal memory.

Now, let us use this generic DPT framework to account for the three major tasks in this chapter. On the Wason selection task individuals will tend to use Type 1 processes which will intuitively call for the pragmatic matching effect (answering P & Q). Some individuals however, will aim for a normative response that will take more effort and working memory use, by means of Type 2 processes (Answering P & ~Q). In the cases of content facilitation Type 1 and Type 2 processes will both end up in the same response, therefore there will be no way of distinguishing the two, that is why there will always be a very high modal answer on these tasks and no individual difference will account for variation responses.

On the Linda conjunction problem, by using Type 1 processing people will violate the normative conjunction rule by preferring a description that fits more with contextual knowledge (Linda is a bank teller and is active in the feminist movement). Individuals who tend to use Type 2 processes will prefer a normative response that does not violate the conjunction rule. Again, facilitations will make the conflict between these two type of process disappear. Stephen Jay Gould (1991, p.469) had a revealing intuition on this: "I know the [conjunction] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me—'but she can't be a bank teller; read the description'."

On the Harvard Medical School problem quick and effortless Type 1 processes will make participants ignore base-rate, while Type 2 processes will require that they work out all the possibilities by using working memory and re-thinking first impressions, which will aim at normative responding. Again, facilitations in forms of frequencies will make the conflict between these two type of process disappear.

DPT summarizes the hidden emerging consensus because it can account for changes in specific explanations. For instance, what can be behind an intuitive answer can be evolutionary algorithms as proposed by the evolutionary psychologists, the use of heuristics, or relevance effects as suggested by Sperber and Wilson (1986), all of these fit with the correlational features of Type 1 processes. Whichever one of these specific theories tend to be right, they will also point to Type 1 processes and therefore to DPT, they can also all be right and apply to the explanation of different tasks. For instance, DPT will still hold if evolutionary psychologists are correct about evolutionary processes for dealing with frequencies and pragmatists are correct about a relevance factor for choosing the P & Q cards in the Wason selection task. It is important to note again that these other proposals do not account well for the cognitive mechanisms behind normative responding across tasks, so these disputes are only over exactly which intuitive Type 1 processing is behind modal answer in given tasks.

It is possible that evolutionary psychologists will not accord to the Type 2 explanations since it evokes a general-type process. However, Cosmides and Tooby (1994, p.94) do indicate that there might be a few of these general purpose processes if dependable on domain-specific ones:

> "In short, although some mechanisms in the cognitive architecture may be domain-general, these could not have produced fit behavior under Pleistocene conditions (and therefore could not have been selected for) unless they were embedded in a constellation of specialized mechanisms that have domain specific procedures, or operate over domain-specific representations, or both."

Perhaps, this could, again, be a difference of focus of research, rather than that of different cognitive theories. Perhaps difficulted by rhetoric excesses such as (Cosmides and Tooby 1994, p.90): "it is in principle impossible for a human psychology that contained nothing but domain-general mechanisms to have evolved, because such a system cannot consistently behave adaptively", since no one today would argue for a cognitive architecture with "nothing but domain-general mechanisms".

Stanovich (2004) criticize evolutionary psychology mainly in three points: 1) that they do not account for how a hunter-gatherer brain can deal with scientific, social, technological advancements; 2) they do not explain the evidence for and the role of general intelligence in their

model; 3) they downplay individual differences, thinking styles and personality. Also evolutionary psychologists have not managed to dismiss the evidence for DPT in individual differences or give a plausible response for how normative responses in the original selection task or the first Linda problem are possible in their account. It is possible these two traditions still have core disagreements about the architecture of mind and how reasoning occurs. Although evolutionary psychologists might not agree, we see DPT as one that can accommodate results and evolutionary mechanisms proposed by evolutionary psychologists and go further on to explain how modern thinking is possible. Little of modern problems are linked in any direct sense to problems of the EEA, and yet we solve them efficiently, perhaps not all of us, but at least some of us, with the same human brain.

In fact, while trying to explain how relevance effects and their Darwinian algorithms might both exist in determining selection task responses, Cosmides and Tooby (2000) propose the principle of pre-emptive specificity, which states that human cognitive architecture will try to answer problems first with domain-specific mechanisms and if not successful more general ones will be called in. This seems to be in line with the emerging consensus towards DPT. As we have seen evolutionary psychology argues that general systems would not be naturally selected for. A dual process theorist might also agree which such claim, and propose, as Stanovich (2004) has, following Dennett (1991), that these mechanisms became possible by means of cultural evolution, by the emergence of memes or by the forces of the Baldwin Effect (see Baldwin, 1896; Dennett, 1991).

1.4 Concluding chapter remarks

We started this chapter endorsing a methodology for interpreting results in tasks of reasoning, judgment and decision making without making an evaluative use of normativism. We have seen how disputes about the boundaries of human rationality have been confused with disputes about the mechanisms of reasoning that humans use in these tasks. Such confusion hid an emerging consensus on the evidence of these tasks to descriptive theories of reasoning. This emerging consensus in short states that there are usually two types of possible responding to

reasoning, judgment and decision making tasks, that is in an intuitive manner (be it by relevance, heuristics or Darwinian algorithms) or in a normative (probabilistic, logic, or others) aimed manner. We argued that this consensus points to DPT, since it is the only one that accounts for such distinction in types of responses, and for the cognitive mechanisms behind normative responding.

Studies in individual differences point out how normative responding is not a random variation, but is in fact a different type of response based on different cognitive mechanisms. We do not mean to imply (as some might) that general intelligence, analytic thinking, or reflective attributes show which individuals are answering correctly. Rather, to us it only points out that there is a pattern between these cognitive mechanisms and normative responses on tasks. This does not imply that a Type 1 response (which might follow relevance or Darwinian algorithms) to any of these tasks is wrong and the normative one is correct, so there is no need to make an evaluative use of normativism in a descriptive DPT. In fact, the opposite can also seem highly plausible. That is, perhaps people who rely more on intuitive relevance judgments, for instance, will be of higher success in social challenges for instance. Success depends on context and on whoever evaluates it. It is more plausible to state that we have two basic types of reasoning processes both of which can be reliable (or malfunction) in various contexts.

It might be argued that the reason why dual process theories can be seen as accounting for such a hidden consensus is that they are too general and hence have little explanatory power. This is a challenge we will face in the next chapter, that is, one of applying rigor to dual process theories, to discover the best ways of defining and understanding each type of processing and to see how well different authors of this camp (such as Stanovich and Evans, 2013; Kahneman, 2011; Mercier and Sperber, 2009; Nisbett *et al*., 2001) can keep coherence with one another and therefore can be understood as being in a collaborative enterprise.

CHAPTER TWO:
BASIC DUAL PROCESS THEORY

2 Basic Dual Process Theory

We have seen in the last chapter how evidence in reasoning tasks points to a difference in two forms of reasoning, one intuitive and fast and the other slow and careful. However, for such difference to be explained in a theory, much more rigor is needed. In the current state of the art in dual process theories, there are various proposed theories with much in common but also with some crosstalk. In this chapter, therefore, we must choose a coherent theory to continue our endeavors, be it a specific one or a general one, and we must lay out the reasons for such choice. In choosing, we will face a dilemma. If our theory is too specific it might narrow our following results too much. However, if our theory is too general then it might be superficial. In order to make such decisions we must explore the following internal issues to dual process theories: (Q1) Can dual process theories that come from other evidence bases such as social cognition, learning, language and neuroscience be united with dual process theories of reasoning and decision making? (Q2) Do the clusters of features really divide perfectly into two groups? (Q3) If most features are only correlational, which are the defining ones? (Q4) Are defining features needed for a coherent theory? (Q5) Do these noticed distinctions point to two minds, systems or types of processing? (the reference problem) (Q6) Is a group of features older in evolution than the other? (Q7) Do these two processes share the same knowledge base and goal structure? (Q8) Do they operate in parallel and compete for control of behavior, or do they cooperate, with production of Type 1 default responses that are then assessed and sometimes overridden by Type 2 processes? (Q9) What processes determine if Type 2 processes are called in or not? (these questions can be followed in appendix I).

There are certainly other internal issues with the theory, but the ones mentioned above are central. If one were not to address a few of these issues in detail it would not even make sense to call this clustering of features a theory. Therefore, we start this chapter by reviewing and commenting on these major conflicts while also clarifying what each concept of these feature means. We then propose to form a Basic Dual Process Theory (Basic DPT), one which will be specific enough to be meaningful but general enough to gather some of the consensus. This Basic

DPT should be found consistently at the bottom of most dual process theories of reasoning and should allow for an easy access for further features to be implemented.

This chapter will focus on four major groups of conflicts that subscribers of dual process theories face with one another. First (section 2.1), we will review a few theories that come from other domains, such as social cognition, learning, language and neuroscience, to see how they add to theories of reasoning and decision making (mostly Q1). Second (section 2.2, 2.3 and 2.4), we will talk about correlational features and defining features of each process (mostly Q2, Q3, Q4). Third (section 2.5), we will see if the proposed distinctions apply to minds, types or systems (mostly Q5, Q6, Q7). Then, we will discuss differences in forms of conflict resolution between Type 1 and 2 responses (mostly Q8, Q9) (section 2.6). Finally, we sum up with a section (2.7) on Basic DPT.

2.1 Single or multiple domain?

One major conceptual matter is to determine if theories from various domains (such as social cognition, language and reasoning) can come together to form a coherent dual process theory of the mind (Q1). One strategy to find an answer is to compare work from diverse research fields in psychology. By the beginning of the 21rst century various similar dual process theories had emerged in different fields of psychology, such as learning and consciousness (Schiffrin and Schneider, 1977, Reber, 1993, Nisbet and Wilson, 1977, Baars 1988), language (Mercier and Sperber, 2009), social cognition (Chaiken, 1980, Chen and Chaiken, 1999), reasoning and decision making (Evans and Over, 1996, Stanovich, 1999, 2004; Kahneman, 2011) and neuroscience (Schneider and Chein, 2003; Lieberman, 2003, Goel, 2005, 2007, Kahneman and Frederick, 2007). It would be compelling to think that all of these theories are actually referring to the same kinds of process in the mind. If that were the case, then perhaps there could be a general dual process theory that referred to distinctive processes of the various functions in the mind, an architectural theory of the mind. If the evidence from all these fields merged to a single theory, such theory would be much stronger than any of the others alone. However, Evans (2008) already tried to unite these different theories, but without much success, since, despite the similarities there are various

intrinsic disparities. These disparities are related to the questions we asked at the very beginning, that is, theories propose varying features of each process, varying conflict resolution models[14] and have varying ideas about what these processes refer to in the mind (dual types, systems, modes or minds). That is why we need to examine these issues in detail to see how a coherent, basic and fairly consensual theory could look like. We start now by reviewing some dual process theories of different domains to see how they can be relevant to our goals.

We start by characterizing what we see as two major groups of theories: General level theories and specific domain theories. The first group includes theories that attempt to explain general processes of the mind that can apply to multiple domains. In this group are theories that propose dual distinctions such as automatic/controlled (Schiffrin and Schneider, 1977), explicit/implicit (Reber, 1993) and conscious/unconscious (Baars, 1988)[15]. Some of these theories are then used by specific domain theories such as Chaiken's (1980) theory of social cognition, Lieberman's (2003) neuroscientific approach, Mercier and Sperber's proposal in language and proposals in reasoning and decision making (Evans and Overs, 1996, Evans and Stanovich, 2013; Sloman, 1996; Stanovich, 1999, 2004; Kahneman, 2011). Therefore, specific domain theories tend to be dependent on general theories, but not vice versa.

The distinctions explicit/implicit, conscious/unconscious and automatic/controlled seem to be given in most dual process theories so we will discuss both how they are presented in their general level theory of origin and how to distinguish these concepts which are usually very related.

2.1.1 Learning

The distinction between explicit and implicit processes has as one of its main relevant bases the study of learning and consciousness. Reber (1993) presents a distinction between implicit and explicit learning where the former is a type of learning that can be complex and occurs without the awareness of the subject, while the latter is a declarative form of learning, which means the subject

14 discussed in section 2.6

15 This does not imply any of these theorists have a domain general view of the mind, only that their proposals are used to describe theories of various domains of the mind.

can describe what he has learned and also the rules governing the knowledge he obtained. Reber (1993) mostly explains what he precisely means by implicit learning through the description of various experiments in which subjects learn patterns by being exposed to stimuli while being unaware of such patterns. For instance, Reber (1967) asked subjects to memorize what seemed to be random strings of letters (such as: PVPXVPS, TSSXXVPS, TSXS). These strings, however, were in fact formed by an artificial grammar, a hidden rule-governed structure. The subjects were told it was a memory task. These strings varied in length from three to eight letters and were presented four at a time. Subjects had to reproduce these memorized strings. In the beginning, subjects were making various mistakes, but after some time they were able to make only few mistakes in reproducing these strings. The control group was asked to do the same task, but strings presented to them were truly random, so for them no advance was made, the high number of reproduction errors of strings continued. Both groups believed their strings were random, so a form of implicit learning occurred in the testing group. Evans and Over (1996) have related this form of learning to Type 1 processes, and explicit learning to Type 2 processes.

Reber (1993) shares with other dual process theorists the idea that these implicit processes are older in evolution and that declarative processes needed to be developed over these implicit ones at a later stage (Q6). He also shares with Stanovich (1999) the idea that individual differences can only be found in non-implicit processes, that is, implicit learning is species-specific and not specific to individuals.[16] The concept of implicitness is rather similar to that of the unconscious. Other researchers have long held that we may be largely unaware of how we think, but aware only of the results of what we think. "It is the result of thinking, not the process of thinking, that appears

16 Although, rather strangely, like some of the critics of dual process theories, Reber (1993) also notes that his distinction needs to be seen as in a continuum from implicit to explicit and not on two distinct poles of opposition. He has made very strong claims on this particular issue: "It is one thing to have an appreciation of the differences between the implicit and the explicit; it is another entirely to conclude that they are processes of altogether different kinds. We do not want to allow ourselves to be seduced by what we can call, for want of a better name, 'the polarity fallacy.' That is, we need to be careful not to treat implicit and explicit learning as though they were completely separate and independent processes; […] There is, so far as I am aware, no reason for presuming that there exists a clean boundary between conscious and unconscious processes or a sharp division between implicit and explicit epistemic systems—and no one from Sigmund Freud on has ever argued that there was." (Reber, 1993, p.23). "In fact, there are no reasons, empirical or theoretical, for assuming that there is any well-defined cut-point or threshold separating the two at some point along a continuum. [...] Our science has had some dreadful experiences with issues like this one. Virtually every time that a continuum of processing or performing is discovered we seem to fall into the fallacy (Reber, 1993, p.24).

spontaneously in consciousness" (Miller, 1962, p. 56). "The constructive processes [of encoding perceptual sensations] themselves never appear in consciousness, their products do" (Neisser, 1967, p. 301). Nisbett and Wilson (1977, p.33) argue that "people often cannot report accurately on the effects of particular stimuli on higher order inference-based responses". Also that "[...] people may not interrogate a memory of the cognitive processes that operated on the stimuli; instead, they may base their reports on implicit, a priori theories about the causal connection between stimulus and response".

It is possible that the concepts of 'implicitness' and 'unconsciousness' really just share the same reference in various uses, the implicit and explicit distinction being used by theorists afraid of using the 'dangerous' word consciousness. Furthermore, psychologists who do use such 'dangerous' word tend to define it the exact same way as those who speak of implicit and explicit processes. For instance, Baars (1988) defines consciousness (or access consciousness[17]) as that to which we (or subjects) have access in a given moment, and unconscious that to which we do not. Notice that this is the same definition of explicit processes. Unconsciousness or implicitness is postulated when the subject lacks awareness of some mental process. "Implicit learning is the acquisition of knowledge that takes place largely independently of conscious attempts to learn and largely in the absence of explicit knowledge about what was acquired" (Reber, 1993, pg 5). Where by explicit we believe he means reportable, and by reportable theorists of consciousness, such as Baars (1988), usually mean conscious.

Supposedly, explicit/implicit processes could be used to refer only to learning and memory functions, so they could be understood as conscious or unconscious learning, whereas consciousness could refer to any given mental process, be it language, vision, audition, decisions and hence refer to a global function. However, when specific domain theorists use the explicit/implicit distinction, such as Evans and Over (1996) and Lieberman (2003), they apply it to reasoning and judgment also. Evans and Over (1996, p.10) say implicit (or tacit) systems of

17 This definition of consciousness is at the level of the easy problem (Chalmers, 1995), one that describes how such process (consciousness) functions at the neural, informational, computational or behavioral level (whichever fits the theorist), because it is an inference about cognitive mechanisms constructed from the subject's report. It does not address the hard problem (Chalmers, 1995) of consciousness because it does not address the nature or ontology of subjectivity or qualia, nor explains how the link of the cognitive level with the level of qualities is realized or possible.

reasoning "operate in parallel, are computationally extremely powerful, and present only their end-products to consciousness" and the "explicit system is employed in sequential verbal reasoning, which people consciously engage in and can give some report about". Lieberman (2003, p.44) says "Controlled processes (sometimes referred to as explicit, conscious, or rational processes) typically involve some combination of effort, intention, and awareness, tend to interfere with one another, and are usually experienced as self-generated thoughts". So the difference in use of the explicit/implicit distinction does not refer only to learning.

At first there seems to be little or no difference between these definitions, but we can make some important points. We do not believe the concept of explicitness equates with consciousness, there is a way of defining it only *in relation* to consciousness[18]. We believe the unique meaning that can be given to the present distinction is that an explicit content possesses an accessible representational format, whereas the implicit content does not[19]. Being accessible will be understood as a content that could be recalled consciously. Given that consciousness refers only to what is focused at a given moment, an explicit content is all that is available in principle to be recalled to consciousness. This final way of characterizing it seems to get the uniqueness of these concepts that distinguishes them from others such as conscious/unconscious and automatic/controlled.

We must also note that explicit memory is not equated with declarative memory. This distinction is made because declarative memory is that which one can recall in a language format. But the language format is not the only representational structure possible that is available to consciousness[20]. We can have, for instance, explicit episodic memory. One could also rehearse a harmony consciously without being able to expose it in language. That is, an accessible format which can be recruited to consciousness and thought about.

18 This use also differs from the philosophical use (such as in Harman, 1986) in which implicit refers to beliefs which are not represented in the mind at all, but that follow from other beliefs, such as: 'the belief that you are not in the moon'. Such use does not capture the findings of Reber.

19 In chapter three and four we will propose exactly what these differences in representational format are and why they are not accessible.

20 Although we believe more than just language is accessible by consciousness, this definition of explicit and implicit would still be fine for those who think language is the only accessible format.

We have defined implicitness negatively, as content that is present in the mind but that does not have an accessible representational format. However, as some believe defining concepts negatively is not enough[21], we can give some examples of what they could be: the memory of the exact movements patterns one needs for walking, the computation behind how we determine a measure of quantity intuitively[22] and other content such as editing procedures done by the visual system; these are not just unconscious, they are implicit because in such a format they could never be conscious, we can know them because of third person science but not by introspection. One could argue that we could teach Reber's implicit rules to someone for conscious access. But then they would be represented explicitly. The format the content was in when in testing was implicit because it could not be accessible to consciousness as it was. In fact, teaching the rules to people would precisely imply that they learn to put it explicitly.

2.1.2 Consciousness

Baars (1988) developed the most common theory of (access) consciousness (at the level of cognitive processes), which also seems to be similar to all of these dual process theories, Global Workspace Theory (GWT)[23]. He understands the brain as a group of specialized processors working autonomously and in parallel. However, he believes there must be some form of communication between these diverse processors. This communication is enabled by the generation of a meeting that exposes the results of these processors, the global workspace; such meeting is broadcast across the brain to all other specialized processors. This meeting is not dominated by only a specific group of processors. Various specialized processors can compete or cooperate to obtain access to the global workspace. This workspace is a shared memory with global access and the criteria for selection of winners is by degree of activation, relevance and novelty. Therefore, global workspace enables global availability of internally consistent messages that are informative to the system. This information provides control and coherent global action. The

21 Since in such case nothing is said about what the concept in fact refers to.
22 Known as subtizing.
23 Of course one does not need to accept GWT in order to accept DPT or the rest of this thesis. The use we make of GWT is merely illustrative, in that it helps us separate the concept of consciousness from other related concepts.

degree of activation of stimuli alone cannot be the whole story for access to consciousness, since in habituation, activation is still high. Habituation occurs because of lack of relevance in the stimuli, and relevance is tightly guided by novelty. Because of that, Baars (1988) suggests that all conscious content depend on activation, relevance and novelty together. So even in the cases in which the global workspace is referring to a memory of past events, it is doing so with the function of bringing some novelty out of that old memory. In first-person view terms, when we are remembering a past experience consciously, we are doing it in a novel way, for instance feeling nostalgia.

Baars (1988) argues that a global workspace can be useful for some tasks while not for others. Such global communication will not be useful for simple problems, known problems, or when a fast response is necessary. However, this is an optimal solution for problems where the need for cooperation of various sources of knowledge and different processors arises, and when there is time to think possibilities through.

Baars (1988) characterizes conscious and unconscious processes by summarizing research in psychology. The theoretical conclusions are that: conscious processes are serial, slow, internally consistent, capacity limited, interfered by stressors and related to awareness, while unconscious processes are parallel, fast, inconsistent (specialized processors can produce contradictory content), have vast capacity, are autonomous (not interfered by stressors) and related to lack of awareness.

2.1.3 Automaticity

These characteristics listed by Baars mostly follow the characteristics of the distinction of automatic and controlled processors of Schiffrin and Schneider (1977). These researchers define automatic processors as those that nearly always become active if matching (external or internal) input is found and are activated without the necessity of active control or attention by the subject. Automatic processes need training to be developed, and once learned they are difficult to suppress, modify or ignore. In contrast, controlled processes are those which subjects are in control of and that have limited capacity. The advantage of controlled processing is dealing with novel problems. These differences have been noticed by the study of two different types of tasks in psychology. Tasks with consistent mapping are those in which target stimuli do not vary. In these, automatic

processes are developed and reaction time and mistakes are decreased. Tasks with varied mapping are those in which target stimuli do vary in each trial. Therefore, automatic processes are not developed and the characteristics of controlled processing can be studied.

Schneider and Chein (2003) summarize such evidence with their automatic and controlled theory. Automatic processes require extensive training to be learned, while controlled processes do not change significantly across extensive trials. When stimuli-response mappings are consistent, processing speed improves and effort reduces. Automatic processing is fast and parallel while controlled processing is slow and serial. Automatic processes are less bothered by other concurrent tasks than controlled processes are. They are also less affected by stressors such as alcohol, fatigue and stress. While controlled processes are easily manipulated, (hence the term controlled) after a learned automatic process has been established it takes longer to unlearn it then it takes to learn it in the first place. Most automatic processes are first trained by means of controlled processes.

We believe the implicit/explicit, consciousness/unconsciousness, and automatic/controlled distinctions, in the way they are usually poorly defined in relation to one another (and even used) can in some cases mean the exact same thing. That is because they mostly refer to whether or not the subject was aware of a certain process and share the same processing characteristics (such as serial/parallel, limits in capacity, fast/slow, autonomous/dependent). However, automaticity can be defined exclusively in a way that is neutral to subject awareness that distinguishes such concept from the concept of unconsciousness: automatic processes are overlearned and nearly always become active in response to a particular input configuration. Controlled processes are, again, defined with reference to subject awareness and they even have a homuncular use, since the subject is used again in the explanation of the subject's mind. But it could be better defined as processes that must work out in real-time (that do not invoke previous solutions) to produce output with some novelty for the system at that moment. Even so the term 'controlled' seems to imply an intentional aspect, and thus seems like an inflated concept (which could be good or bad depending on the need of the theorist, see Frankish, 2004, 2009). So despite these various similarities, we can use these terms carefully with a unique definition to each that distinguishes them one from the other: Consciousness/unconsciousness refers to the access one has to the content of the environment, of the body and mental processes in a given moment; the implicit/explicit distinction refers to the

content's format, whether it has an accessible representational format (explicit) or not (implicit); and automatic/controlled refers to the readiness of the process, where automatic processes are overlearned and ready for use and controlled processes are developed in real-time (that do not invoke previous solutions, and are possibly intentional). These concepts could be further developed, as they are in their original theory, but with these simple descriptions of their uniqueness we can guarantee coherence, distinction and precision of concepts to be used in specific domain theories.

Having discussed these concepts and their general theory of origin, we can move on to very briefly analyzing a few specific dual process theories which come from other evidence bases other than reasoning and decision making: social cognition, language and neuroscience.

2.1.4 Social cognition

Chaiken's (1980, Chen and Chaiken, 1999) dual process theory attempts to explain how we think socially. In any given judgmental context, such theory distinguishes between two basic modes of processing: systematic processing and heuristic processing. Systematic processing is analytic by nature and enables a comprehensive treatment of information. Systematic processing requires cognitive ability and capacity, and thus is influenced by knowledge and limited by time constraints. Clearly, by definition, such processing mode bears resemblance to Type 2 processing. The heuristic processing mode applies various learned judgmental rules to reach responses, such as: "consensus opinions are correct", "moderate opinions minimize disagreement", "go along to get along", "expert's statements can be trusted". These heuristics can be applied without much effort from the subject. They must, however, be available (known), accessible (recallable), and applicable (relevant to put in use in some given situation). Evans (2008) argues that this might not correspond to a Type 1 processing, but to some less effortful mode of Type 2 processing. However, the description perfectly fits Evans and Stanovich's (2013) main requirement of using little working memory resources, and also some correlational ones of being fast and heuristic in nature. Evans

(2008, p.268) claims that such heuristic mode does not resemble Type 1 quick contextualization[24]: "heuristic processing in this theory sounds more like the recognition-primed decision making of Klein (1999) than the contextualization process postulated by reasoning theorists (Evans 2006, Stanovich 1999)". Such heuristic processing really does not resemble quick contextualization, but perhaps it resembles the second kind of Type 1 processes, noted also by Evans (2009, pg 42), the autonomous processes that are distinct from other Type 1 pre-attentive processes: "Autonomous processes are those that can control behavior directly without need for any kind of controlled attention." Seen in such manner, the Type 1 and Type 2 distinction can be found in this social psychology theory.

The theory also proposes a principle for when each type of process should occur. Chen and Chaiken (1980) postulate the sufficiency principle, which proposes that for any given judgment there is a continuum of judgmental confidence with two critical points. One point refers to the subject's actual confidence in the judgment and the other refers to the subject's desired confidence. The subject tends to use heuristic judgments first and when these low-effort processes do not manage to make the actual confidence reach the desired confidence, and if the subject has the required time and cognitive ability, then he may start using systematic processes to reach the desired level.[25]

Chen and Chaiken's (1999) theory also includes motivational criteria that influence processing. In this theory, motivations and goals are prior to these dual thinking processes, so they differ from the 'two minds' hypothesis (explained in section 2.5, Q5 and Q7). Therefore, there is no such thing as a motivation or a goal of one of these two types of thinking, rather, prior motivations and goals constrain the use of both of these processes. There are three defined motivations that constrain possible use of the dual processes: accuracy motivation, defense motivation and impression motivation. These motivations are also a priori dependent on the individuals' unified goals. When an individual has accuracy motivation, he tunes his thinking skills to reach the most accurate attitudes and beliefs. With such motivation, individuals tend to act open-

24 This is the first kind of Type 1 processing mentioned in Evans (2009). It refers to pre-attentive processes that usually guide our decisions to relevant matters.

25 This is an interesting point for the discussion of conflict resolution in Dual Process Theory [Q8 and Q9]).

mindedly and evenhandedly. If heuristic processing does the job accurately, systematic thinking is not called into play. If, however, heuristic processing is not enough, and there is enough of motivation and cognitive ability, then a systematic process can provide a more analytic review of information.

The second possible broad motivation is defense motivation, with which subjects tend to hold attitudes and beliefs that are coherent with one's interests or one's concept of himself. Therefore, subjects will tend to choose attitudes and beliefs that accord to his values, identities or personal attributes. Such motivation constrains dual processing to selectively process information that preserves defensive beliefs and attitudes. So heuristic processing will apply only the learned heuristics that are relevant to achieving the defensive strategy. If motivation is high and cognitive resources are available, then the subject can also use systematic processes to endorse defensive needs. For instance, they might analytically favor information which reinforces their defensive needs, and shun or rationalize information that shows the contrary.

The third possible broad motivation is impression motivation, a desire to hold attitudes and beliefs that will satisfy the current goal. Like the former one, it also biases information to its needs. Therefore, if one's desire is to minimize conflict in a group, by using heuristic processing, he might choose to apply procedures such as "Go along to get along" or "Moderate opinions minimize disagreement". With cognitive resources and high motivation, systematic processes might also be used while constrained by impression. Chen & Chaiken (1999) illustrate such possibility with an interviewee which is armed to counterargue views in opposition to those of the interviewer.

2.1.5 Language

Dual process theories have also been proposed in studies of language cognition. Mercier and Sperber (2009) share the intuition that there can be a fundamental distinction between Type 1 and Type 2 processes in our thinking. They ground this distinction in a massive modularity architecture (along with Samuels, 2009, Carruthers, 2009 and Eraña, 2012). They propose that there are two different types of inferential processes: intuitive inferences (which might resemble Type 1 processes) and reflective inferences (which might resemble Type 2 processes). They

propose that intuitive inferences are the conclusions of inferential modules, characterized as those accepted without need to attend to the reasons why. They understand reflective inferences as the indirect output of a specific module, the argumentation module. In this architecture, intuitive inferences are not realized by one system but by multiple. However, reflective inferences are the indirect output of only one module. Mercier and Sperber (2009) believe that the multiplicity of intuitive inferences and limitedness of reflective inferences corresponds to the old mind/new mind hypothesis[26] (an answer to Q6), in which intuitive inferences are present in animals and in humans, and reflective inferences are perhaps uniquely human, but limited even in the latter. They argue that while in most dual process theories, Type 2 processes are seen as built to enhance individual cognition, on their view reflective inferences can only be understood as a collective process of communication, this means both that their function is communicative and that they are constituted socially.

How reflective inferences work depends on the argumentation module. The argumentation module is a special purpose module like any other one. Its function is to take as input a claim and produce as output reasons to accept or reject it. However, we may just accept these reasons as a claim without further investigating the reasons for such choice. In this sense, the direct output of the argumentation module is also intuitive. Now, when engaged in communication, we are able to provide a comparison of results between arguments, reasons and conclusion which enable reflective inferences. These are indirect outputs because they are meta-representational, they are about the results of the argumentation module.

Mercier and Sperber (2009) argue that this capability evolved for the necessity of epistemic vigilance in communication. This is a need for filtering reliable information from all the possible false or wrong information that could be communicated. For instance, if the speaker lies he may have an advantage, but if the audience learns to filter such information, the liar may be excluded and will then have a disadvantage. Various type of game-like competition between speaker and audience produce a selective environment for epistemic vigilance. Reflective inferences were made possible by means of such vigilance over others' argumentation modules, in their view.

26 For an interesting development of the old mind/new mind hypothesis see Stanovich (2004).

As evidence for these claims, Mercier and Sperber (2009) point to a possible boosting of awareness of reasoning in group and argumentative situations. In such contexts, if one of the participants has the correct answer, then the others will also arrive at it. Mercier and Sperber (2009) argue also that evidence shows that it is not only that the best participant gets the correct answer and others will follow them, but rather that conflict in arguments call in reflective inferences that boost performance.

Therefore, Mercier and Sperber (2009) argue for a massive modularity account of dual process with a very specific function and working of Type 2 processes in communication. Their considerations enable further discussions about the old mind/new mind hypothesis (Q6) and the multiplicity or singularity of systems of thinking (Q5)

2.1.6 Neuroscience

Finally, there are also studies that use neuroimaging to find brain differences that could pinpoint dual processing. Goel (2005) proposed that the frontal-temporal pathway corresponds to a heuristic system and the parietal pathway corresponds to a formal/universal system. It is proposed that reasoning about known problems utilizes situation-specific heuristics in the first system and when no such heuristics are available the latter system is called to solve the problem. Belief biased responses to syllogisms (usually related to Type 1 processes) activate the ventromedial prefrontal cortex, while noticing the conflict between logic and belief activated right lateral/dorsal lateral prefrontal cortex. Goel (2007) argues that the somewhat chaotic neuropsychological evidence base suggests that there might not be a unitary reasoning system in the brain, but rather a fractionated system that is dynamically configured in response to environmental cues[27].

Schneider and Chein (2003) claim that controlled processes can only be active when facing a novel problem. So they studied which areas stopped activating after overlearning a task. Theoretically, these areas should then correspond to controlled processing. Brain regions which stopped activating were: bilateral prefrontal, anterior cingulate, posterior parietal, occipital-

[27] Of course this could also be a result of methodological and conceptual difficulties.

temporal, and cerebellar areas. Schneider and Chein (2003) further propose to explain what cognitive activity some of these areas realize. Supposedly, the dorsolateral prefrontal cortex performs sequential goal-directed control; The posterior parietal cortex provides the mechanism for selective attention; The anterior cingulate cortex monitors activity from automatic processes and sets decision criteria; The medial temporal lobe is responsible for the resumption of a task after interruption through recall/reloading of inner loop modules. Although they suggest various details, they concede these are speculative and preliminary.

Lieberman (2003) proposes a neural dual process theory that postulates the reflexive system (also called X-system) and a reflective (also called the C-system) system. The X-system (which supposedly corresponds to Type 1 processing) is responsible for linking affect and social meaning to currently represented stimuli. Three neural areas are associated with such system: The amygdala, the basal ganglia, and the lateral temporal cortex. The C-system (which supposedly corresponds to Type 2 processing) is activated when handling exceptions is needed. For instance, when a social rule is broken and one needs to rethink behavior in a certain situation. Three areas compose this system: the anterior cingulate cortex, prefrontal cortex, and the medial temporal lobe. The first is responsible for detecting the need for top-down control, the second implements control, and the medial temporal lobe stores information about past episodes for use in controlled processing.

Kahneman and Frederick (2007) interpret results on framing effects[28]. They propose the Type 1 response might be dominated by an emotional evaluation following high activation in the amygdala, which can be suppressed and result in conflict given activation in the anteriorcingulate cortex and that the orbital and medial prefrontal cortex might resist the framing effect by inhibiting Type 1 responses or by integrating emotional responses with further information.

There is some resemblance in these four proposals. For example, the dorsolateral region is mostly assigned to control functions like goal managing and conflict resolution, the medial temporal lobe has a function of recalling information necessary for control, and the anterior cingulate has a monitoring and conflict resolution function. In Lieberman (2003) and Kahneman and Frederick (2007) the amygdala plays a role for emotional Type 1 responses. However, the data used to explain dual systems are much more divergent than convergent. Automatic processes are

28 The effect of varying a response to the same question if asked in two or more different ways

not localized specifically in Schneider and Chein's (2003) proposal, whereas for Lieberman (2003) they have three specific locations, and they correspond to a frontal-temporal pathway in Goel (2005, 2007). There are various mentioned areas that are not common to all the accounts, such as cerebellar areas, ventromedial prefontal and occipital areas. Although methodologies used are alike, different specific theories seem to form up rather than a convergent general data that could support a unified neuroscientific view of DPT. Also, DPT is not useful for discovering the functions of brain areas, most of the functions of these areas have already been studied under other domains of neuroscience and are just transferred in an attempt to explain DPT.

To conclude section 2.1, we note that we understand that all of these specific domain theories could be based on a basic fundamental one, but (as Evans 2008 notes) as of yet no one has managed to propose a coherent unified version. There are too many different ideas and concepts, be it motivational characteristics of Chen and Chaiken's (1999), specific modules of Mercier and Sperber (2009), brain differences in Liberman (2003), Schneider and Chein (2003) and Goel (2005, 2007). And this is only in the ones we mentioned, it is unlikely that all of or even most of these theories will stand coherently together without some major theoretical analysis which is beyond our current scope. Although they can contribute with suggestions and hints for dual process theory of reasoning and decision making, unfortunately these theories still need to be thought of separately as specific theories. Some of these theories, such as Chaiken's and Lieberman's might be seen as extensions to what we call Basic Dual Process Theory in section 2.7, while other such as Mercier's and Sperber's and Goel's might not. To find out which theory could be based on Basic DPT one would have to study all these alternative ones in detail to determine which follow the same defining definitions of each process and which do not.

As for theories that speak of general features such as explicitness, consciousness and automaticity, these cannot be separated or excluded from any of the specific ones because they are also concepts inside various specific theories, including dual process theories of reasoning and decision making.

2.2 Defining features and alignment criteria

In the last chapter we presented a generic version of dual process theory by contrasting a list of features of two types of processes. Such a list has two main problems (Q2). First, some of these concepts are not well defined in relation to one another, many are very intimately related or could be understood to mean the same thing. So we cannot understand precisely the meaning of these concepts together if not previously carefully distinguished. As an example, consider these two statements in the same paragraph: "theorists have tended to equate System 1 with implicit (unconscious, preconscious) processes and System 2 with explicit (conscious) processes." (Evans, 2009, p.37) and "Some authors specifically frame their dual-process accounts in terms of conscious and nonconscious processing […] and others rely heavily on the implicit/explicit distinction" (Evans, 2009, p.37). On the same paragraph, concepts of consciousness and explicitness are both equated and used as core differences. As a philosophical work that examines DPT we cannot let this type of language remain. Second, even if we identify each of these concepts precisely, it is not true, empirically, that all of those features align perfectly as always co-occurring features. This has been a point of criticism developed by Osman (2004) and Kruglanski and Gigerenzer (2011). It is quite improbable that such a strong co-occurring requirement meets reality. Because even if, say, only six dichotomies are advanced, there are still 64 (2^6) possible combinations of these features that need always co-occur. If DPT were proposing such an alignment assumption (see Stanovich & Toplak, 2012) then only one of these possible 64 combinations of features would be enough to falsify the theory. Suppose these dichotomies were: conscious/unconscious, explicit/implicit, controlled/automatic, serial/parallel, slow/fast, resource dependent/resource free. Each process that lacked one element of these aligned features would serve as evidence to falsify DPT. For example, a process that was conscious, explicit, controlled but parallel would be evidence for falsification, even considering that most features of such process were rather aligned than unaligned.

Evans (2009), Stanovich and Toplak (2012) and Evans and Stanovich (2013a) have pointed out that the claim about these clusters of features, even with a reduced amount of them, was never meant to be a claim of necessarily co-occurring features. These claims were pointed out in reviews (such as in Stanovich, 1999; and Evans, 2008) as an examined attempt to unite aspects of various dual process theories that seemed to be at least partially compatible. Anyhow, such criticism did point out the need to better define what precisely dual process theories of reasoning and decision

making want, instead, to claim. To do so, the attempts to better specify the theory have been to try and map precisely a difference between defining and correlational features of each type of process (Q3, Q4). Defining features are used to identify each type of process and correlational features are only typically present. Stanovich and Evans (2013a) have argued that there are two defining dichotomies in these processes, that Type 1 processes are autonomous and do not require working memory while Type 2 processes require working memory and are (or can be) decoupled. Evans and Stanovich (2013a; 2013b) work hard to defend these defining features but do not worry about correlational features. Defining features are important but they are not enough to sustain a theory, that is, they describe only a few characteristics of each process. The correlational features are also a big part of the theory and need to be well addressed.

We argue that correlational features are still not well addressed because in their latest work they still list a very large and fuzzy set of features. So the recent large list of correlational features of Type 1 processes are: fast, high capacity, parallel, nonconscious, biased responses, contextualized, automatic, associative, experience-based decision making, independent of cognitive ability, evolved early, similar to animal cognition, Implicit knowledge, basic emotions. The recent large list of correlational features of Type 2 processes are: slow, capacity limited, serial, conscious, normative responses, abstract, controlled, rule based, consequential decision making, correlated with cognitive ability, evolved late, distinctively human, explicit knowledge, complex emotions (Evans and Stanovich, 2013a, p. 225)

The problem with this large list is that it does not differentiate the importance of each item in these correlational features group. These theorists probably believe that some correlational feature might be added or deleted from this list (since they change the list in each new paper, both with inclusions and exclusions), but it cannot be the case that all these correlational features are completely replaceable, since the theory with only defining features says next to nothing. Take Evans' (2009) first attempt of defining the dichotomy by means of working memory use alone. If such were the only distinction that mattered, then the theory would be trivial, since no theorist would disagree that some processes load on working memory while others do not or load less. The problem, therefore, is that such list sets some very important properties, such as being fast or slow, at the same level of being related with basic or complex emotions (a distinction that might not make

it to the next list). So we propose that some correlational features should be marked as essential to the theory, we term this group ‘core correlational features’. For the theory to hold, these must, at least, have correlations among them, they must be aligned with their type more often than not. Notice that we mean these features must always be present in the theory, not that they must always be present in an instance of a thinking process, otherwise they would be defining features.

We will base our suggestion of these core correlational features on the use of these concepts to defending DPT. A good starting point is taking what Evans (2009) states as the core (most agreed upon features) of DPT: that Type 1 processes are automatic, low effort and have high processing capacity; and that Type 2 processes are controlled, high effort and have limited capacity.[29] We can also add the explicit/implicit distinction because Evans (2012) uses it often in his arguments, and to lose it would mean to lose various of his arguments for defending DPT, which would be a much greater loss than, say, losing the idea that Type 2 processes are uniquely human. And despite somewhat mixed results (Stanovich and West, 1998b; Stanovich and West, 2008), dependability on cognitive ability is a great factor, since individual differences has shaped how the theory has developed. We add one last item to this core which is tending toward normative responses or tending toward intuitive responses, because such is the analysis used in the evidence base to build the theory in the first place as we showed in the last chapter. This could replace the features ‘biased response’ and ‘normative response’ (in Evans and Stanovich, 2013a), because as they themselves argue extensively, this is a very simplistic reading of the evidence. The feature ‘tends towards intuitive responses’ includes the possibility of bias and success, and ‘tends towards normative responses’ includes the possibility of mistakes in the attempt to find a normative response. If other listed features turn out not to have strong correlations, that will not be a such a damaging blow to DPT as it would in case new evidence shows that some of these core correlational features are not *even* correlated at all.

We also do not have to exclude the rest of the features from this new list and we can even leave space open to include more elements under a group termed ‘other correlational features’. These are features still in testing, so excluding or not listing them does no damage to the theory. On the other hand, if more evidence is gathered for a strong correlation of these other features then

29 speed is also mentioned but we elaborate on it differently further on.

they can move up a category, becoming part of the core features group. Thus, we will use a list of our own but one which was elaborated not by testing new experiments or further theorizing but by carefully reading the works of dual process theorists of reasoning and decision making, mainly Evans, Stanovich and Kahneman, while trying to propose a common, agreed upon, core to DPT, which with more elaboration we will term Basic DPT (The list of features we use can be followed by the reader in appendix II).

A further worry (see Kruglanski, 2013 Osman, 2004 and Keren, 2013) is that if there is any fuzzy state in between the two extremes of a defining feature than a continuum theory is more suited. We advance the *comparison principle* for dealing with cases where empirical psychologists discover instances of such fuzzy states. The comparison principle has two strands. The first is comparing properties of different responses in the same task (e.g. working memory needed for processes of response 1 in comparison to working memory needed for processes of response 2) and the second is examining how the response stands in relation to other defining features (e.g. if working memory use is similar and therefore cannot help decide the identity of the process, then use other defining features to determine which type the process is)[30]. The two types of processes are not limited to a single defining feature, they are concepts composed of multiple defining features. So it is no problem if one defining feature alone cannot help determine, in some case, an instance of a thinking process. The alignment assumption must hold at least for defining features, that is what grants their status as defining, they must always be present. This guarantees DPT's falsifiability and at the same time enables the comparison principle, which grants the theorist the right to use all defining features together as means of identifying an instance of a Type 1 or 2 process. We will go into further details on the comparison principle in section 2.7 when we show how processes can be identified and how the theory can be falsified. We must state it now also because it will be used to explain defining features.

We have two further steps in this section. The first is to clarify all the concepts used as features to give them a unique meaning (Q3). The second is to discuss why defining features have such a privileged status and why core correlational features do not (Q4). We have already clarified

[30] This does not mean that it is okay to find the opposite pairing, this is for when evidence is fuzzy, not for when it is clear to the contrary of the thesis. We elaborate this in section 2.7.

the concepts of implicitness/explicitness, consciousness/unconsciousness and automatic/controlled (in 2.1.1, 2.1.2 and 2.1.3) while discussing the general theories that propose them so we will not go back to them in further detail[31].

2.2.1 Working memory

Let us begin with the first defining feature. We believe (unlike Keren, 2013) working memory is a very safe concept inside cognitive science (we think that if it is not, then no other one is); various models of it are possible, but generally Baddeley and Hitch's (1974) theory covers the general meaning of the term. Working memory is a short-term storage that holds and processes about seven items at a time. The effort increases as the number of processes it needs to realize increases; too much load will make it malfunction. It is said to have a verbal storage to deal with linguistic representations and a visual-spatial storage to deal with imagery. Later increments to the theory also add an episodic memory buffer. The usefulness of this concept for DPT is that working memory can be safely measured, it can help distinguish between effortful and non-effortful processing and it is also deeply related to other features such as seriality, low capacity, slowness and control. Also, it surely seems to tap an essential distinction between the two processes.

However, we have still identified a possible slippery edge on this part of the definition. We raise a doubt: do Type 2 processes have to load heavily on working memory or can minor loading also be considered? Could Type 1 processes load mildly on working memory? Evans (2009) seems to distinguish it in terms of absolute use or no use at all. By such means, to be considered Type 1, a process could not involve any use of working memory at all. That is a very strong claim and the advantage of working memory being an easy measure for distinguishing the two types may be lost. For example, how is one to discover that when exhibiting the matching effect on the selection task one is not engaged in use of working memory? Testing participants for individual differences in working memory is no help, since measuring span or executive functions[32] will not help determine

31 although they are discussed again in relation to predictive processing in chapter three and in relation to classical architectures in chapter four.

32 Span is the amount of items an individual can hold in short-term memory and the executive functions are those that apply changes (functions in the mathematical sense) to such held items.

if working memory was *minimally* used. If the working memory defining feature requires Type 1 processes to make no use at all of working memory, then such feature will hardly be testable, since various other processes could be using working memory (like reading and thinking about and understanding the test task at hand). It also seems to be implausible because multitasking tasks that require Type 1 thinking should generate some sort of inner conflict, less but some, so at least some load should be present. It follows one could multitask using Type 1 processes better than with Type 2, but clearly there is also a limit. Perhaps the complete lack of working memory use can only be seen in pre-attentional processes, which are quick and come first in processing, and therefore can be isolated from other processes. It seems to follow from definition that pre-attentional processes would not require any working memory at all. Anyhow, to eliminate working memory use completely from Type 1 processes results in a need for a reduction of which processes count as Type 1 (usually a multitude).

Stanovich and Toplak (2012), perhaps aware of this conceptual issue, speak of processes that are 'relatively undemanding of cognitive capacity' versus processes that are 'capacity demanding'. We believe it seems more realistic to speak of lesser use of working memory for Type 1 and loading strongly on working memory for Type 2. This does however raise some issue on the language used by Evans (2012) and Evans and Stanovich (2013a). In the latter paper (p.225), when listing the features, the authors distinguish them by a complete lack or use of working memory participation: "Does not require working memory" and "Requires working memory". We read this as 'does not require so will not use'. If we were to read it as 'does not require but could use', then we would need a statement about the amount of load that Type 2 processes need to produce to differentiate itself from a Type 1 process that loads minimally, or statements about how to analyze such fuzzy states[33]. On the same paper, on the abstract and when discussing the issues we get a different statement about the need of working memory: "What defines the difference is that Type 2 processing supports hypothetical thinking and load heavily on working memory" (Evans and Stanovich, 2013a, p.223) and "Type 2 processing is distinguished from autonomous Type 1 processing by its nature—involving cognitive decoupling and hypothetical thinking—and by its

[33] This is what we chose to do by developing the comparison principle.

strong loading on the working memory resources that this requires" (Evans and Stanovich, 2013a, p.226). So this issue is not very clear in their words since they are not differentiating these two possible uses: requiring working memory from loading heavily on working memory.

It is not that we are being picky. To show this is a critical issue, consider this claim: "The problem is that there are also fast type 2 judgments that are made on the basis of simple rules and heuristics, with minimal reflection. This kind of type 2 processing will also make minimal demands on working memory […]" (Evans, 2012, p.129). Of course the only way to make sense of this is if a total lack of working memory is necessary for Type 1 processes, else the statement is meaningless. He could also be unintentionally meaning that such process is Type 2 because it is explicit, but then he needs to make a separate defining feature for explicitness. Anyhow, we believe these mistakes need to be avoided and using the distinction between 'less or no use of working memory' versus 'a strong load on working memory' is a safe way to end this problem. One might want to ask 'well how much is little or strong load?' In cases where that is not obvious, the answer should be found by appeal to the comparison principle, a comparison from two or more answers in the same task or by examining the other defining features (see section 2.7).

2.2.2 Speed and Effort

Before moving on to the second defining feature, let us examine the reasons why the fast and slow distinction needs to be correlational and not defining, according to Evans (2012). One of the reasons uses the argument cited just above (Evans, 2012, p.129), by saying that there can be quick Type 2 processes that load minimally on working memory. If there are such cases than speed should not be a defining feature of Type 1 processing. But by getting the working memory case clear, we can already find that this cannot be so. That if there is such type of process that loads minimally on working memory than it probably is, on a more careful analysis, a Type 1 process[34].

The other two arguments to show that speed is not a defining feature are also flawed. Evans (2012, p.128) considers these examples:

> "(a) an art expert 'knows' that a statue is a fake but cannot prove it by any explicit reasoning or knowledge, and (b) a hospital employs a short checklist to decide whether to

34 he also says nothing about decoupled representations in this paper still so there is no way that could be what is saving the argument.

> treat patients with chest pain as suspected heart attack cases, with much better success than more complex procedures. Both are relatively quick ways of making a decision, but it is evident that judgment (a) has the implicit type 1 characteristics, whereas (b) type 2 is entirely explicit."

Evans (2012) argues that while example 'a' is a quick Type 1 process, 'b' is a quick Type 2 process. Again, instead of sticking to his own way of distinguishing the two processes, that is, by focusing on the use working memory, he bases his argument on other ideas, which he probably sees as an umbrella form of Type 2 processes, that is their explicitness. But this point is nonetheless fixable, in his point of view, he could argue that 'b' is a process that depends on working memory and is nonetheless pretty quick. Indeed. However, he is understanding speed in a very fuzzy way. We could ask if speed is measured by response time, by the time the answer comes to mind, by the time measured in a specific context or time in varying contexts. Of course, depending on which task one is asked to do, and on which context a given process is used, the speed of processing will vary. So it does not make sense to consider an example of quickness from one task for a Type 2 process and contrast it to a Type 1 example in another task. Of course, if there were a way for one to use a Type 1 process to solve the problem in example 'b' then it should be faster. Therefore, alluding to the comparison principle we have: 'When solving the same task, a Type 1 process will always work faster than a Type 2 process'. This 'faster' should be understood in the sense of coming first to mind. This can be verified empirically by examining measures of speed in tasks that use less or more of working memory and decoupling or by subject report. We believe it is an empirical challenge to show a case in which the response to the same task that needs more working memory and decoupling will come faster to mind than the one that does not. At first glance, it seems testable and highly plausible, almost necessary from definition that such DPT thesis will hold and speed will always be correlated with less working memory in the same task, since a process that relies strongly on working memory should always be slower than one that does not, in the same task. Therefore, we believe we can consider a third defining feature, switching the fast/slow distinction to a more precise 'faster in comparison' and 'slower in comparison', in the sense that we just explained.

Carruthers (2009, p.120) makes a similar remark: "Since System 2 is realized in cycles of operation of System 1 it will be slow by comparison". And even Evans (2007, p.330) "All dual

process theorists agree that heuristic (System 1) processes are much faster than analytic (System 2) processes". It also seems to be a defining feature for Kahneman (2011). Allowing speed to be one of the defining features also permits the theory to incorporate direct supporting evidence (Roberts & Newton, 2001; Evans & Curtis-Holmes, 2005; Frederick, 2005; De Neys, 2006; Kahneman, 2011).

The other argument against the fast and slow distinction draws upon Gigerenzer's view on heuristics and Betsch's (2008) concept of heuristic judgments. Evans (2012) argues that these types of heuristic judgments have a rule-based explicit format and therefore need to be understood as Type 2 despite being fast. Again, the explicitness of the content was never a Type 2 defining feature to begin with. So if that is the only reason, then that is not good enough from a methodological perspective. Anyhow, Evans (2012) could be right that responses that follow heuristics are not necessarily always Type 1. Perhaps working memory could be involved in figuring out, in real-time, which heuristics to apply. In that case, such process will still load strongly on working memory and therefore be a Type 2 process. However, the point is that a Type 1 solution to the same process would involve less working memory and thus should be faster than that of a Type 2.

A similar story applies to effort. Effort is a simple idea that relates to how tiresome using a process is to an individual. Clearly it bears relation to speed, since the more time you spend in something the more tiresome it becomes. Also, we are speaking of processing effort so we need to understand effort as in tiresome mental process, not as effort needed to complete a whole given experiment. This can be measured by competing tasks and can also be assessed by subject report (see Kahneman, 2003, 2011). We have a hard time understanding how a Type 2 process could be less difficult to execute and less tiresome than a Type 1 process in the same task. Evans and Stanovich (2013a, 2013b) do not explain why effort cannot be a defining feature, but if you dig a little deeper you can see that the reasons are similar to those proposed for speed. They seem to be thinking of the distinction in terms of high effort and low effort across different tasks. Again, of course, you will have Type 2 processes which are low effort if you contrast it with responses of a different task.

For instance, Evans (2011, p.92) states (using old mind and new mind terminology) that: "Whereas the old mind forms associations, the new mind acquires shortcut rules and heuristics that

are applied explicitly but with little effort. Such low effort Type 2 thinking can also be a cause of cognitive biases". Of course some instances of Type 2 processes will be less effortful than others. Also, there is no reason to think that all Type 2 processes will be extremely high effort. So again we need both strands of the comparison principle here. If it is the case that a low effort process seems to be an instance of a Type 2 process than one must contrast it with the effort in what seems to be a Type 1 process in the same task. If this comparison is fuzzy, so that effort is not able to discriminate to what class a given instance belongs to than the comparison must be taken to the other defining features. For example, given two processes, even if effort is not different enough to discriminate them, if it is the case that one of them is slower and loads strongly in working memory we have enough reasons to think it is a Type 2 process. The other process might be faster and load less on memory and thus Type 1 even if testing for effort was not enough to discriminate them. What would falsify the theory, for example, is if the clearly more effortful process in a task loaded less on working memory and was faster in comparison. By making use of the comparison principle, effort, like speed, could be a defining feature as Kahneman (2011, p.31) has proposed: "The defining feature of System 2, in this story, is that its operations are effortful". Thus, we can add that to Evans and Stanovich (2013a) that a defining feature of Type 1 processes is that they are prone to 'less effort in comparison' and Type 2 processes are prone to 'more effort in comparison'.

Evans and Stanovich (2013b) probably worry that having too many defining features might make the theory not stand to evidence. But at the same time they lose communication with Kahneman, or they pretend to have such communication by pretending both theories are the same. Proposing the alignment assumption for the defining features is the only way to make these differ from core correlational features. Having more defining features makes the theory stronger but it also makes it less probable of being true. However, by alluding to the comparison principle we believe even this stronger theory stands to current evidence. At the same time even a recipe can be given for falsifying the theory, as done in section 2.7.

2.2.3 Autonomy

The next defining feature is that Type 1 processes are autonomous. We will keep Stanovich and Toplak's (2012, p.7) definition of the concept. They are autonomous when "the execution of

T1 processes is mandatory when their triggering stimuli are encountered, and they are not dependent on input from high-level control systems"[35]. Being autonomous is different from being automatic. The word autonomy highlights more clearly its independence as a cognitive process, but more than that it includes processes other than automatic processes. Recall that automatic processes are overlearned and nearly always become active in response to a particular input configuration. Stanovich and Toplak (2012) argue that autonomous processes include emotional regulation processes, Darwinian algorithms proposed by evolutionary psychologists, implicit learning[36] and automatic firing of overlearned associations. Therefore, automaticity is but one aspect of Type 1 processes that need not always be present. For instance, judgments of frequency (as studied in the last chapter) are not overlearned, but rather could be something like Darwinian algorithms present irrespectively of learning. Judgments of frequency are computed autonomously, but if we follow the definition of automaticity as being overlearned, in the sense of extensive training, they cannot be considered automatic. Now this might sound puzzling as most people tend to think of these as automatic, but this is precisely the sort of effect you get by following definitions strictly.

One might want to argue that these innate mechanisms are overlearned in the sense that these processes were collected during evolution. However, here we see how the second part of the definition, that of not being influenced by higher level input is also important. Because there seems to be a difference between mechanisms that are trained and thus are very open to change by scrutiny of higher level input and those which are not. Thus, some autonomous processes might not be automatic because they are not overlearned and are not open to change by training.

Autonomy is more of a characteristic of Type 1 processing rather than a distinction with a counterpart for Type 2 processes. That is because we are not really sure what this counterpart would be, if not, the already mentioned use of working memory. We could add a 'dependent on global communication' for Type 2 processing but we are not sure if this would be a defining feature as

35 notice that 'not dependent' does not mean they could never communicate with higher-level control, which is also agreeing with the interpretation of less or no use of working memory rather than always none at all. Also, if they were completely independent, how could they possibly be overridden by Type 2?

36 If DPT is about thinking and reasoning, we are not sure that one should include learning processes. However, implicit processes, those that do not have a representational format in thinking and reasoning, would still be of the autonomous type.

autonomy is for Type 1 processing. Of course we could also add as a counterpart that Type 2 processes are controlled and thus that these processes must work out in real-time (not invoke previous solutions) to produce output with some novelty for the system at that moment. We could also add intentionally controlled as an opposite to autonomy but we are really unsure about how to work this out in a less inflated way. Being so, we believe that controlled could be kept as a correlated counterpart to automatic, rather than autonomous. Therefore, we keep only clear concepts for defining features.

Kruglanski (2013) questions if the characteristic of being mandatory could define a type of processing: "How immediate is immediate, however? […] as a vast number of conditioning studies demonstrate, the strength of the stimulus-response association varies along a continuum." The answer is to that is alluding to the comparison principle: it should be considered in comparison to a second response to the same task or by consulting other defining features.

2.2.4 Decoupling

The fourth and last defining characteristic is decoupling. Stanovich & Toplak (2012) derive the concept of decoupling from two other similar ideas. Leslie's (1987) idea of secondary representations and Nichols and Stich's (2003) possible world box. The goal of both of these ideas is to model pretense[37]. In pretending, according to these ideas, the individual correctly perceives the actual situation but mentally entertains another. Therefore, they conclude hypothetical thinking could be the key ability in pretense.

Leslie (1987) terms primary representations those that must reflect aspects of the world directly, accurately and faithfully. He argues that if the pretend lexical item were also a primary representation, then it would add on to the lexical item's real sense. For example, if a shell represents a cup, then shells, in the pretender's mind, would become a new form of cup. But in pretense, rather, it seems that there is actually a dissociation from the primary representation

[37] It is their goal to model pretense. Ours is not. We have no theory of pretense, and not even a favorite one among candidates. The point in speaking of pretense here is merely illustrative, to help understand the notion of decoupling we speak of in DPT.

meaning (different reference). Not just the lexical content, but the properties of the pretend world would mix with properties of the real world (different truths). If the pretend world was primarily represented, then if the cup was full of water in the real world and in the pretend world it was empty, then the property of being full would extend to include situations where there is no water. Finally, a pretend world can have imaginary objects which cannot have a real-world reference (different existence).

Leslie (1987) proposes that in pretense one is using a meta-representation which copies and decouples the primary representation from its normal input-output relations. While this secondary representation is copied and can be manipulated to form pretense, the original primary representation remains fixed to its input relations (its reference, truth value, and existence). Decoupled representations do not have a direct reference in the world, rather they relate to parts of the primary representation. These parts can become variables to be manipulated in pretense and hypothetical thinking. For instance, when pretending that a ball is the moon, the shape is kept but other parts are altered, such as the color or the fact that it is now a space object.

Nichols and Stich (2003) speak of a possible world box. They explain that the job of a possible world box is to represent different possible situations given some set of assumptions. These assumptions feed from real beliefs and desires but differ from them since they need not be true or one might not want them to be true. The concept of 'box' means only that there is a specific functional mechanism for this kind of hypothetical thinking. It is a workspace built temporarily to store representations of other possible worlds[38]. It also does not commit the theorist to localize such box in the brain[39]. Nichols and Stich (2003) argue that a possible word box realizes simpler functions than mindreading[40] which is their explanatory end. By representing the world as it could be, the possible world box can be used for (besides mindreading) planning, generation of empathy and hypothetical thinking in general. We are unsure if either of these theories advance the debate

38 Nichols and Stich (2003) note that this concept of possible world is larger than that currently used Philosophy. That is because possible worlds with no obvious contradictions are still possible worlds in this use. After all, if the contradiction is not noticed than it is still possible for the subject.

39 It can follow Dennett's (1978) homuncular functionalism strategy, that is, to progressively reduce the explanatory power of such boxes, so that in the end they are supposed to be realized by simple mechanisms.

40 Mindreading as in the ability of predicting others beliefs, desires and goals based on their behavior, it is not, therefore, a supernatural ability.

in pretense. But we are not worried about pretense, but rather with what these theories can help us understand about decoupled representations.

It seems sustaining decoupled representations is distinct from primarily representing abstract objects. For example, a task can include abstract objects (such as x+y) but that can still be a primary representation since one represents these objects directly, after all they are given in an abstract format in the real world by the task. To be decoupled, the representation needs to alter some aspects of the input, making hypothetical judgments or simulations of possibilities which are not given. Indeed, that is why (supposedly) working memory is required according to Stanovich and Toplak (2012); to sustain 'in mind' both aspects of the primary representation and aspects of a simulation of possibilities in decoupled representations.

Stanovich and Toplak (2012) claim that the key and defining feature of Type 2 processing is using decoupled representations. Now, this can be understood in two different more strict senses. The first is that all Type 2 processes must include some decoupled representation component. This is a very strong claim, because together with the claim that all Type 2 processes must include working memory (Evans and Stanovich, 2013a) we get as a result the claim that, at least for solving thinking tasks studied in this literature, working memory can work only with decoupled representations. If both are strictly defining features, then if a process has strong working memory use but makes no use of decoupled representations it would either falsify the theory or be a process not mentioned by the theory, since it cannot be Type 1 or Type 2 by definition. We do not believe there is enough evidence to claim that strong working memory use must *always* be associated with decoupled representations. Notice that this is a problem that started only after Evans and Stanovich (2013a) combined their defining features (working memory and decoupling). We believe we should understand this claim in a second sense: that all processes that include decoupling must be Type 2. This is probably implied[41] when Evans (2007, p.334) claims that "abstraction probably implies analytic thinking, but not vice versa". In this weaker sense, we can include Type 2 processes that load strongly on working memory but that are not decoupled. Notice that this does not make decoupling only a correlational feature, because Type 1 processes can never be decoupled.

41 It is implied because as we have seen abstraction is not the same as decoupling. Evans (2007) did not use the concept of decoupling, although he did use 'decontextualized' alongside the idea of abstraction.

Therefore, it is still a key feature of Type 2 processes that only they *can* use decoupled representations.

There is also some trimming needed with this defining feature. Osman (2013, p.251) has argued there are examples of effortless mental simulations:

> "If we look to what causal reasoning involves, mental simulation is often required to play out various scenarios and to imagine hypothetical consequences. Is this capacity associated with T1 or T2? Take our ability to follow complex story lines in soap operas […] We are able to mentally simulate multiple complex events and infer what would happen next with a great deal of ease".

Indeed, decoupled cannot mean any imaginative production; it must mean decoupling from context. In the situation of watching a soap opera, the possible pathways the story takes follow from the context. If someone were to hypothesize about things that do not follow from the plot then they probably would lose the storyline because of overuse of working memory. A decoupled representation needs to ignore the environment for a moment, not deduce direct consequences from it even if in image form. Osman's argument could also be dismissed by saying that there is not enough evidence that people do not use working memory at soap operas. For instance, although it seems to be done with ease, it could be the case that people with higher scores in working memory use are better at discovering plot twists. But since we do not have such evidence at hand, and since there are probably other various sorts of imaginative and hypothetical processes, it is best to define decoupled representations as only those that do not directly follow from (or ignore) context or what is given perceptually[42].

2.3 Core correlational features

Now we move on to discuss how best to understand core correlational features. In these cases the comparison principle does not apply, since it is conceded that some of these features will not be present in a given instance of a given type of process. In fact, even a given instance of a Type 1 process, as identified by defining features, might have an opposing, Type 2, core correlational feature. What is important is that most often than not these features will respect their

[42] More on this discussion in chapter three and four.

alignment. They differ from the group ‘other correlational features’ because they cannot simply be abandoned without damaging much of the arguments and evidence in favor of DPT.

2.3.1 Capacity

Capacity is an interesting core correlational feature because it is related to working memory use. It would not be too risky to say that these features follow as a consequence of working memory use, but we will keep previous skepticism, since processes could in principle be limited for other reasons, for using other type of resources. Baars (1988) suggests limited capacity can be understood from a combination of three sources of evidence: selective attention; competing tasks and short-term memory. Therefore, claiming that Type 2 processes have limited capacity would mean that they work with a limited amount of inputs from selective attention (Simon and Chabris, 1999), they are usually disturbed by competing tasks (Schiffrin and Schneider, 1977) and are limited by the capacity storage of short-term memory (Miller, 1956). The contrast is that Type 1 processes have high capacity, but that is no surprise since by Type 1 we are referring to processes executed by a massive collection of systems. Although some (see Samuels, 2009) argue that Type 2 processes do also depend on a massive collection of systems, the fact that they depend on working memory does not let them work all at the same time as Type 1 processes could. Anyhow, we do not have reasons to think that all Type 1 processing will be of higher capacity in comparison to all Type 2 processing, even in the same task, and also it may be the case that even the opposite pairing can show up in a given instance.

2.3.2 Cognitive ability and normativism

Cognitive ability has its origins in Spearman’s G (1904) (General Intelligence). It is related to success in learning, problem solving and professional life. Although it is hard to define what it refers to, it is measured with tests and is a generally accepted feature in the psychology community

for its predictive success[43]. Despite somewhat mixed results (Stanovich and West, 1998; Stanovich and West, 2008), cognitive ability tests have landed support to DPT for being correlated (more often than not) to what theorists take as Type 2 responding (Stanovich & West, 1998a, 1998b, 1998c, 1998d; Toplak *et al.* 2011). Perhaps what is most important is that it is consistently found that cognitive ability cannot predict Type 1 responding, while, in comparison, it can predict Type 2 responding more often than not. The Cognitive Reflection Task has produced more interesting results, but it does not go in as a cognitive ability or as related to an independent feature, because supposedly it correlates to dual processing as a whole (Frederick, 2005; Toplak et al., 2011; Toplak *et al.* 2014). It is easy to see why this cannot be a defining feature, since there are subjects who will score low on cognitive ability but which will solve decision making tasks better than subjects who scored high.

Type 2 processes also tend to aim at normative-like responses, which means that they try to follow some agreed upon normative principle as in logic, probability or general mathematics. Even the mistakes in these processes can be seen as mistakes in execution of the procedures necessary for achieving a normative response. Type 1 processes tend to be aimed at intuitive responses, be it by pragmatics, Darwinian algorithms or heuristics. These are only correlational because nothing stops a fast and autonomous process with limited working memory use from being aimed at a normative response, for instance. Although these were essential characteristics used to develop DPT from the evidence base, now we can use them as core correlational because other safer defining features of the process have been found.

[43] This is hardly a decent definition of intelligence. But there is no use in getting into theories of intelligence since we will make no progress without major disagreements. And debating the nature of general intelligence is not what we have as a goal here. This methodological use of intelligence implies that whatever is tested in intelligence tests is what matters for us. Of course, this is not nearly a good enough explanation. However, we must deal with the fact of the correlation of intelligence tests with reasoning abilities somehow. And the simplest way we have right now is not theorizing about intelligence, and rather only stating that results in cognitive ability tests predict results in reasoning. It is also not necessarily related to intelligence per se but to academic success, since the tests used were like college exams such as 'SAT'.

2.3.3 Other distinctions

The implicit/explicit, consciousness/unconsciousness and automatic/controlled distinctions are also core correlational features, but we have already discussed them in sections 2.1.1, 2.1.2, and 2.1.3. It might be important to remember that as core correlational features these also might not be present in a given instance of a specific Type of process or even the opposite pairing can obtain.[44]

2.4 Other correlational features

The group properly termed 'other correlational features' refers to features that are only correlational and that might be abandoned by DPT without internal damage to the theory and thus we will not spend too much space worrying about them. They include the distinction between serial and parallel processes[45], abstract and contextualized and older or recent in evolution. These are features which have not been given much attention by recent work in DPT[46].

We have now covered each relevant feature proposed in DPT in some detail. The definition of each feature can be found in appendix III. We will now move on to the question of what these clusters of features refer to in cognition.

2.5 The reference problem

Another point in need of conceptual discussion is the reference of these two clusters of features. Even for those who agree that these different defining and correlational clusters of features exist, the question (Q5), the reference problem, becomes: what are they evidence of? Dual systems? Dual types? Two minds? Dual modes? Or something else? These questions are not so simple

44 Evans probably has not noticed that by making the explicit/implicit distinction only correlational he is agreeing with cases where Type 2 processes are not explicit and Type 1 processes are not implicit.

45 although this is important to the account we are developing further in this thesis we do not see it as central to Basic DPT.

46 We will, anyhow, come back to them in chapters three and four

because the evidence base does not seem enough to tell what exactly is being accessed. Therefore, one needs to analyze the best philosophical and theoretical reasons for choosing one reference over the other.

The best way to start off is by distinguishing the meaning of systems and processes. Schacter and Tulving (1994) have attempted this before in detail, in order to clarify their concept of memory systems. They have also been the reference used by critics of DPT (Keren and Schul, 2009), so, to be fair, we will refer to them: "Thus a memory system is defined in terms of its brain mechanisms, the kind of information it processes, and the principles of its operation" (Schacter and Tulving, 1994, p.13). Now Keren and Schul (2009) use the definition of a *memory* system. They even focus on the 'kind of information it processes' criteria. However, this is an important criteria for memory, but not necessarily for reasoning and decision making processing, since memory is distinguished by modalities or kinds. But although some processes may be domain-specific, reasoning and decision processing use all modalities and kinds of information in order to figure out the best response, or at least are not neatly individuated like memory. Thus, the main point of this definition which is of use to DPT is brain mechanisms involved and the principles of a system's operation. As we have seen brain mechanisms have been hinted at but there is no consensus on brain mechanisms underlying a System 1 or System 2. Thus it seems different 'principles of operation' is what mainly DPT has been aiming at.

Our assumption is that DPT has progressed, although also not to completion, in describing principles of reasoning operation. Schacter and Tulving (1994, p.16) describe different criteria used to identify memory systems: "A memory system must be described in terms of a property list, that is, an enumeration of its features and aspects by which its identity can be determined and its relation to other systems can be specified." This has obviously been where dual process theorists have spent most of their effort. A second point is that "as long as the system is intact, it operates class inclusively, in the sense that it can process any particular input or information of the specified kind" (Schacter and Tulving, 1994, p.15), and finally "[...] converging dissociations: dissociations of different kinds, observed with different tasks, in different populations, and using different techniques" (Schacter and Tulving. 1994, p.18). This last criteria for identifying systems has been

pursued successfully throughout the history of experiments (see chapter one). So thus it might seem that DPT is aiming at systems of reasoning.

2.5.1 Dual systems

Although independent system theories with specific distinctions had already been proposed earlier, such as, heuristic/analytic (Evans, 1989), implicit/explicit (Reber, 1993) and associative/rule-based (Sloman, 1996), the first consensual collective explanation for why these two groups of properties form clusters was attributing them to the processing of two distinct systems of cognition: System 1 and System 2 (Stanovich, 1999). However, over the years some problems with this system terminology were found (Evans 2009, 2012; Evans and Stanovich 2013a, Samuels, 2009). To explain the details of these problems, Samuels (2009) distinguishes two forms of these system theories: The token thesis and the type thesis. The token thesis states that the mind has a rigorous division into two distinct systems. Therefore, it is understood that processes exhibiting Type 1 properties will be executed by System 1 and processes exhibiting Type 2 properties will be executed by System 2. The type thesis asserts that there are multiple systems in cognition that can fit into the category of a Type 1 or Type 2 system.

Theorists noted that the token thesis is flawed (Samuels, 2009; Evans 2009, 2012; Evans and Stanovich 2013a). Samuels (2009) argues that even because of the methodology of cognitive science of functionally decomposing systems into smaller parts there are bound to be far more further divisions than just two particular systems. One could respond that there are not two systems of cognition but rather just a two systems division in a smaller region of cognition, such as reasoning (such as Sloman, 1996). So the token thesis for reasoning would be claiming reasoning is realized by exactly two systems. However, as Samuels (2009) points out, there is no simple way to map an isolated reasoning system into cognition (also shown by Goel, 2007). He considers five suggestions of the token thesis for reasoning: (1) Any inferential device is a reasoning system; (2) Any system that subserves conscious deliberative inference is a reasoning system; (3) Any device involved in paradigmatic reasoning tasks is a reasoning system; (4) Reasoning systems are to be identified with so-called ‘central’ systems; (5) ‘If you’ve got to ask, you’re never going to know’.

He dismisses suggestions 1 and 3 because they are too inclusive: for instance, perception and language are probably needed for inference. Therefore, there is no reason to think that two systems of reasoning will be found at such level. Suggestions 2 and 4 are dismissed on the basis of excluding too much. Most of the attention DPT gets is based on its claim that Type 1 processes do not depend on central processes, and could be unconscious and automatic, so defining all reasoning as related to consciousness or central processes alone misses the point. What is left is suggestion 5, which more precisely means ignoring a detailed definition and using the already known distinct research areas as defining of what reasoning processes are: arithmetic inference, probabilistic reasoning, decision making, planning, spatial reasoning, reasoning about social phenomena, ethical judgment, reasoning about the minds of others, enumerative induction, and abductive inference. Again, even at this level this is too inclusive; there is no reason to suppose that there are only two systems for all these functions. The token thesis seems implausible[47].

The multiplicity of Type 1 systems was acknowledged early, Stanovich (1990, p.118) argued that ‘‘there are no strong theoretical reasons to believe in the unity of automaticity […]”. However, as Samuels (2009) notes, the token thesis has been taken seriously for a longer time in the case of System 2. Theorists spoke of it as a unique system that overrides the multiplicity of Type 1 results. Perhaps this has to do with the fact that Type 2 processes seem to be related to our personal-level identification (Frankish, 2009). But as Frankish (2009) notes, even if Type 2 processes are more identified with our idea of self, which seems to be a unity, it could still be functionally decomposed into smaller procedures (as Dennett, 1991, has explained). Seriality also makes the case for the Type 2 token thesis, since one process at a time suggests one processor. Another point is that of modularity. As we have seen, DPT was initially influenced by the Fodorian architecture which posits a unity of central processes (which might correspond to System 2) and a multiplicity of modules (which might correspond to System 1, see introduction). However, Evans and Stanovich (2013a) argue that there are processes referred to as Type 1 that do not fit Fodorian

[47] We must be clear on a detail about our use of reasoning in this thesis. We will speak of DPT as a theory of reasoning. However, we do not share these sorts of trouble because we are not attempting to pinpoint specific reasoning systems (as in the token thesis) in the brain-mind. Thus, when we say DPT is a theory of reasoning, we are saying that it postulates cognitive mechanisms for explaining something like a folk psychology idea of reasoning, not that it postulates specific brain regions for reasoning.

modules criteria (Fodor, 1983) or Darwinian module criteria (Cosmides, 1989). Also, Evans (2012) finds it implausible for there to be only one System 2 which is responsible for hypothesis testing, rational decision making, overriding biases, belief revision; there are too many functions ascribed to just one system. In fact this started to come out as consensual by 2009. For instance, Evans (2009) and Stanovich (2009) mention that the overriding is clearly a different function from the actual processing of another result. Also Evans (2009) starts thinking about a Type 3 process to take some of Type 2 functions and Stanovich (2009) distinguishes the reflective level from the algorithmic level of Type 2 processing. As the multiplicity of computational processes (what is done) and specific functions (how it is done) increases the plausibility of the token thesis for a System 2 decreases.

The result of the awareness of this multiplicity can be noted by a shift to the Cognitive Kinds Thesis (CKT). CKT proposes cognition is composed of two *kinds* of systems, those that exhibit the Type 1 property cluster and those that exhibit the Type 2 property cluster. Samuels (2009, pg.132) grounds this distinction with three properties of a *natural kind*[48]:

> "(1) "It is associated with a range of characteristics or symptoms, which tend to be co-instantiated by instances of the kind, but need not be genuine necessary conditions for membership."; (2) "There is some set of underlying causal mechanisms and constraints—a 'causal essence', if you will—whose operation explains the co-instantiation of these various symptoms".; and (3) "To the extent that there is any real definition of what it is for something to be a member of the kind, it is not symptoms but causal essence that defines membership."

Samuels (2009) argues the CKT is the most plausible proposal for the reference problem. By considering a possibility for various system 2s one could easily include theories from other domains. An example we could suggest is Mercier and Sperber's (2009) argumentation module, as it could be thought of a System 2 for language processing only. Samuels (2009) notes that by including this multiple domain possibility it becomes predictable that various features might not always correlate. This difference of features by domains could actually point to different characteristics of these different systems 1s and 2s. So, in this view, a System 2 for social cognition might have different properties from the argumentation module for language while sharing some central System 2 characteristics. Both would still be systems of the same kind, they would, for

48 Developed by Boyd (1991) following the tradition of Quine (1969).

instance, share the same working memory resources and have various correlational features. As Samuels argues (2009, p.140) "This is because it is quite possible for distinct, individual mechanisms to differ in some respects even if they are members of the same broad class of mechanisms."

Although CKT is the only working system proposal, we believe it might not be a safe base for Basic DPT. It is very unlikely that different cognitive scientists would have produced theories from various domains that were precisely describing real similarities and distinctions of cognitive kinds in various domains. If all these scientists were finding converging evidence of the same things, it would be rather convincing, but since differences are also very robust, it seems much more realistic that these differences in properties are because of incompatible theoretical hypothesis rather than differences in real physical systems. It seems the CKT solution is one that attempts to unite dual process theories of different domains, but as we have seen, there are problems with different feature use in different theories as well. We think DPT should have good reasons for excluding and including theories and such good reasons need to consider the questions about features (Q2, Q3, Q4).

The different characteristics of features is what motivated research and the theory itself since its beginning, it is a core aspect. The different clusters of features are directly extracted from the evidence, it is what responses in psychological tests tells psychologists about their object of study. They cannot be freely altered by theoretical needs. There must be rigor when speaking of the clusters of features. For instance, if one takes Evans and Stanovich's (2013a) proposal as a base, then theories that do not take working memory use as a central feature for distinguishing the two types of processing cannot be part of DPT. Such restriction allows DPT to make clear, coherent and testable claims. While CKT is an interesting thesis, it might lead to a license for including theories which have different features. Furthermore, critics might use this broadness to find troubling counterexamples to the theory, such as: "simulations are central to processes typically considered to be T1, such as perception, categorization, and motor movements" (Keren, 2013, p.258). "Two-system theories have also been proposed to explain phenomena in other areas, such as perception [...] memory [...] self-control [...]." (Keren and Schul, 2009, p.534). Pure perception, memory, and motor movements are not among the skills which are studied by DPT.

The CKT can also be troublesome if we consider the neuroscientific data we briefly reviewed in section 2.1.6. As we have seen, the data is not convergent. Following Samuel's (2009) permission that difference in theories might be related to different system 1s and 2s, one could conclude that these indicate different characteristics of differing systems 1 or 2. The main problem with such thesis is that it does not allow us to distinguish between different results from different brain systems. How is one to tell if these results are helping individuate systems if they could in fact be explained by the use of different tasks, environments, individual differences, differences in use of concepts, and sample? Much more convergent evidence would be needed. However, since divergent evidence is taken in the CKT as evidence of differing S1s and S2s, the scientific process misses its aim. Having various research groups would help only to find more confirming evidence and not to contrast and eliminate wrong results.

2.5.2 Dual types

Samuels (2009, p.132) argues that “any version of dual-process theory worthy of the name is committed to these generic claims”: that “there is a division in our cognitive architecture—a division between cognitive systems—that explains this clustering effect”. The other claim being that “cognitive processes tend to exhibit either the S1 or S2 property clusters (C2)”[49]. We are not sure if the system proposal is the only one available, but if Samuels (2009) means that further specifications to the reference problem are needed beyond the 'type' distinction he is right. What Evans and Stanovich (2013a) propose in our mind is actually to ignore the reference problem for now. When proposing to speak of types and not of systems they free themselves from possible trouble in system terminology, but at the same time ‘types’ does not actually answer the reference problem. It could be thought that types are a defense of the reference as processes and not as systems. However, we doubt that dual process theorists would be worried in defending a more detailed identity of a process, such as the one provided by Schacter and Tulving (1994, p.12) "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

49 S1 and S2 meaning the correlational properties of each type.

systems but are not identical with them". Evans and Stanovich (2013a, p.226) actually argue for neutrality: "These terms indicate qualitatively distinct forms of processing but allow that multiple cognitive or neural systems may underlie them" they do not go on to speak of the difference of systems and processes or how each type of process.

It could be argued that features describe characteristics of systems, such as being 'limited capacity' and not of processes. In fact this is the reason why Keren and Schul (2009) think DPT must describe features of systems and representations. But this is actually no worry at all. All that would need to be done (if it were necessary to speak rigorously of processes) is a rephrasing of how features read, such as instead of 'limited capacity' one could read 'processes limited items', and instead of 'autonomous' one could read 'does not need to wait for communication with other procedures to function'. Features were characterized as pointing to systems because that was how the theory was usually read, but there is nothing in the evidence that makes us describe it so. Evidence was gathered by responses on tasks but the theorizing is just an inference made to explain pattern in responding, therefore, we can feel free to rephrase it without changing how DPT explains the evidence.

2.5.3 Dual modes

Although 'type' does not allude to a more specific characterization of DPT into systems or processes, Evans (2009) and Stanovich (1999) argue that DPT needs to be a theory of cognitive architecture. They contrast an architectural theory with those that propose different modes or a continuum between the two types. For instance, Nisbett at al. (2001) propose a distinction between holistic and analytic thinking styles in eastern and western culture. This is understood in DPT as a difference in Type 2 processing, it is a difference of modes or styles, influenced by culture, not of architecture and therefore should not be considered as a DPT claim. As we understand this, this meaning of architecture is in the sense of hardwired[50]. It also distinguishes from continuum theories. Keren and Schul (2009) who favor a single system theory argue that such a single system could apply different classes of functions which vary across a continuum. Further, single system

50 A hardwired function is that which is universal among humans.

theorists deny there are two dichotomous non-continuous clusters of features which correspond to a hardwired qualitative difference in the mind. The 'type' terminology, rather, is used precisely to classify these qualitative differences.

2.5.4 Two minds

Let us also consider the possibility of two minds. Mind in cognitive science is usually equated with the functional whole. In this sense, all cognitive functions together would compose one's mind. If dual process theorists then decide to claim that there are two minds, a different idea of mind needs to be proposed. Evans (2009, p.35) defines mind "as a high-level cognitive system capable of representing the external world and acting upon it in order to serve the goals of the organism". This definition of mind includes that of a system, a mind seems to be a kind of system. However, Evans (2009) uses it in a specific sense. A mind has access to various forms of processing, to other modules and the same mind can have access to distinct Type 1 or 2 processes. Frankish (2004) adds that each mind should have its own belief structure (Q7), and therefore makes a distinction between two types of beliefs[51]. Stanovich (2009) speaks of the algorithmic mind and the reflective mind, which are aspects of System 2, but these could never use Type 1 processes. Minds could also have different evolutionary histories (Q6), possibly with Mind 1 being shared with animals and Mind 2 being more recent in evolution. Stanovich (2004) also claims that these minds have different goal structures, because Mind 1 was developed by and is sustained by genes (following a gene-centered view of evolution) while Mind 2 was developed and is sustained by memes[52], supposedly these two units of selection could cause conflicting goal structures (Q7).

While this difference in evolutionary history and difference in belief structures are interesting, there are problems with the dual mind proposal. The concept of mind is ill defined. If a mind is a system then the dual mind hypothesis is just a token thesis of dual-systems theory. However, since the difference in evolutionary hypothesis and belief structure are interesting, this

51 He calls them 'basic beliefs' and 'superbeliefs' see (Frankish, 2004).

52 A Meme is an informational unit which could be the unit of selection in the development of ideas and cultural evolution (see Dawkins, 1976).

sort of development could have some alternative formulation in an attempt of distinguishing it from the token thesis of dual systems while saving the previous hypothesis. A possibility is that what characterizes the distinction is that Mind 2 is realized in cycles of operation of Mind 1 and both are functionally decomposable (Carruthers, 2009, Frankish, 2009). In that sense, mind is just a name used to refer to a certain level of processing rather than a real sharp division in cognitive architecture.

We believe none of these proposals are unproblematic and thus the reference problem remains open. At the end of this thesis, in our final conclusion, we will work a bit more on reference problem. For Basic DPT, which is a consensual reading we are attempting to make of dual process theories of reasoning and decision making, we believe the reference problem should remain open, since, as we have argued, what should unite DPT is its focus on well define features. Thus we will keep using the neutral 'type' terminology.

2.6 Conflict resolution models

In section 2.1 we reviewed some dual process theories from other domains and found some differences. However, there are different possible ways of explaining the same results inside a dual process theory of reasoning and decision making. One difference is that of the structure and relation of each type of processing to one another (Q8, Q9). What is needed is a model of conflict resolution, a model to explain how each type of process interacts.

Evans (2007) describes three possible cognitive models to explain competition of control that propose different procedures. First is the pre-emptive conflict resolution model. The main idea behind this model is that the decision of whether to apply a Type 1 or 2 process is made at the outset, that is, before any real problem solution has been computed. Supposedly some superficial minor aspect of the problem could cue what Type to use beforehand. This model implies that conflict is avoided rather than solved, because the conflict can only happen after at least two responses are ready. The second possible model suggests that the two types of processing occur in parallel and the decision to apply a Type 1 or 2 process occurs only after both solutions have been proposed. In the end they will have either the same response or conflicting responses. Sloman

(1996), could subscribe to this view[53]. Finally, the third one is called the default-interventionist model. It proposes that Type 1 responses are always default, that is, they will be given unless an overriding occurs by a Type 2 process. According to Kahneman and Frederick (2002) Type 2 processes monitor these default responses and may or may not override depending on their quality.

Evans (2007) argues that although the pre-emptive model is undermined because people seem to have access to both types of responses when in conflict, it cannot be completely ruled out. He also argues that evidence cannot rule out a parallel model either, but he has a tendency to gravitate towards the default-interventionist model.

Evans (2009, p.48) undermines the parallel model by using the horse race argument:

> "Here the problem is worse because the parallel model describes a horse race between a very fast horse (type 1) and a much slower horse (type 2). Not only does the fast horse have to wait for the slow horse to arrive, the slow horse also gets to decide who has won!"

The only advantage of having a faster process is if it could, at least in some situations, not need to wait on slower ones. Also, the evidence shows that in time-pressured tasks, Type 1 responses are much more frequent, indicating that they do not need to wait, if time is not available (Roberts & Newton, 2001; Evans & Curtis-Holmes, 2005).

There is some actual data that can be found for the primacy of Type 1 processes by introspection, or by the report of other's introspection when describing how they came to a conclusion during a task. The CRT is famous precisely because it is able to generate an intuitive response that will come to mind first[54] (Frederick, 2005).

Now, if we assume there is some waiting needed to be done and that Type 2 can only come with time, and that subjects report having found Type 1 response first, then in what sense is this process parallel? If Type 1 processes are bound to be ready first then they are not parallel simply because they will always precede the following response. The whole point in being parallel is that both are executed simultaneously. So if a process is much faster than the other, this already undermines the parallel model[55]. If each process supposedly had independent control of behavior,

53 We say 'could' because what Sloman (1996) says is also partially compatible with the default-interventionist model. Simultaneously believing two contradictory responses is also feasible by such model.
54 As explained in chapter 1.
55 Note that there is a second sense of parallel, which means being executed in a neural network, however that is not the only sense necessary for the parallel model of conflict resolution in DPT, but rather that the two processes are occurring at the same time.

then why would one need to wait on the other? The problem is that if one claims that the first does not need to wait on the second then what they would have as result is actually something like a default-interventionist structure, one answer always coming first, but having a chance of being overridden by follow-up processes if time is given.

Another difference is that the default-interventionist model assumes Type 1 responses are default, while the parallel model seems to assume the execution of a parallel procedure for figuring out a response. However, Type 1 responses can be shown to be default because they are recurring and uniform among a variety of people. That is why programs like evolutionary psychology and heuristics and biases were empirically successful. These researchers were able to find very strong modal responses in a variety of tasks, which are now grouped together because of sharing Type 1 features. Type 1 responses are frequently used and by discovering their mechanisms we know which default procedure has been recurrently used, be it a Darwinian algorithm or a heuristic. Therefore, having a default response seems to be a true characteristic of the conflict resolution process that is not emphasized by the other two models.

However, these differences themselves still do not show why Type 1 processes should have to wait in *any* model. Our best argument is that there is a pre-wired command (or a hyperprior, see chapter three) that states that if time is given, then the response should go through further scrutiny, be it in the form of actual changes or just rationalizations; or alternatively, that the further scrutiny command is constantly active, being only deactivated by external cues for quick answers. Also, perhaps it is of empirical interest to evaluate subject's actual and desired confidence in response in traditional reasoning tasks to see if Chaiken's (1980) suggestion (in 2.1.4) could be a determinant factor for conflict resolution.

2.7 Basic Dual Process Theory

We have now discussed major internal issues to dual process theories and we want to suggest what we see as a Basic Dual Process Theory. This theory is not any type of new proposal; this is rather a clarification of what this basic aspect of consensus in reasoning, judgment and

decision making can look like. Some theorists that we believe could agree with this Basic DPT are: Carruthers, Evans, Frankish, Frederick, Kahneman, Samuels, Stanovich and perhaps Lieberman and Chaiken. Theorists will want to say more than Basic DPT does and should feel free to add new proposals onto this base. These proposals will be understood as extensions to Basic DPT (a good example is Stanovich, 2011). However, it would be interesting if this basic theory could serve as a common ground which is consensual mostly because it is clear, theoretically coherent, predictive, supported by evidence and has not yet been falsified.

Since our following work (other chapters) will mostly be a meta-theoretical enterprise[56], for means of coherence, we will assume the truth of this Basic DPT, as it will be the starting point of our endeavors. Our further suggestions will also count as extensions to this basic version. We might contradict various other different extensions to Basic DPT. This Basic DPT does have, however, rigorous established features and a default-interventionist conflict resolution model. What we leave open is the reference and the unity problem. Theories from other domains (as long as they are related to thinking and reasoning, as is the case of social cognition) could adopt Basic DPT if they agree with the requirements that we have discussed and that we summarize in what follows. Therefore, it should be at least an umbrella theory for dual process theories of reasoning and decision making and possibly other thinking skills. That is, one which various theorists of this research field or similar research fields could agree on. In this section, we will show what this Basic DPT could look like by uniting all the best answers to the internal problems we have reviewed.

We start with the list of features as reviewed (also appendix II), this time having reviewed the definition of features and having three groups following the importance of features. As explained earlier, we have defining features, core correlational features and other correlational features. In order to subscribe to the Basic DPT one must agree on at least the eight cells of defining features and on some of the core correlational features. Theorists may disagree on which are some other possible correlational features, and even add more defining or core correlational ones, since theories are always changing and should keep being polished. However, by subscribing to at least to this Basic DPT, theorists can have a solid consensual base on which to work on.

56 Meaning that we will work with this theory of reasoning and not reasoning itself.

Critics (Keren, 2013; Kruglanski, 2013) of DPT have mentioned that Evans and Stanovich's (2013a) new formulation of the theory is less interesting and not falsifiable. Evans and Stanovich (2013b) respond that DPT is a meta-theory, that is, a general framework (such as evolutionary psychology) that serves to provide more specific directly testable explanations and predictions of tasks. For instance, DPT can answer why responses show the matching effect and which individual probably will not respond in such manner. It is true. Evolutionary psychology also has a not falsifiable general understanding of human psychology. They use a priori arguments for showing why our mental tools need to be Darwinian algorithms or why our mind is massively modular. At least for now, the framework itself is not decisively testable. It does, however, prove to be an interesting framework for proposing testable specific theories, such as social contract theory. Although, if terms are clarified, the meta-theory proposal for DPT is fine, we agree with critics that it makes the theory less interesting. In the same sense, it would be better for evolutionary psychology, for example, if their a priori hypothesis were actually testable hypothesis. So perhaps DPT could be more than a meta-theory.

Therefore, we believe it is less interesting only because we can formulate DPT more precisely in a way that it is directly testable and falsifiable while still holding to evidence and making strong predictions and explanations.

Type 1	Type 2
Defining Features	
Less or no use of WM	Loads strongly on WM
Autonomous	Can use decoupled representations
Faster in comparison	Slower in comparison
Less effort in comparison	More effort in comparison
Core Correlational Features	
Automatic	Controlled
High Capacity	Low Capacity
Implicit	Explicit
Unconscious	Conscious
Uncorrelated with cognitive ability	Correlated with cognitive ability
Tends towards intuitive responses	Tends towards normative responses
Other Correlational Features	
Parallel	Serial
Contextualized	Abstract
Older in evolution	Recent in evolution

Table 1 – Defining features, core and other correlational features

Image 1 (not table 1) shows how theorists can identify a process as Type 1 or Type 2 and also serves as a guide as to how to falsify the core of the theory. So, given any thinking task with at least two possible responses, a theorist can use the five[57] defining features as criteria to decide which type of processing was primarily used. Therefore, when analyzing a second response in comparison to the first, theorists should ask themselves five questions: 1) whether this second response used a process which was faster than the former; 2) how much use of working memory was necessary; 3) if the process was autonomous; 4) if the process required the use of decoupled representations and 5) whether the second response required more or less effort. This flowchart is also a guide for applying both strands of the comparison principle. In the case of the first strand, it

57 five because autonomous does not contrast with decoupled.

shows that questions are always posed in relation to two (or more) conflicting responses in the same task.

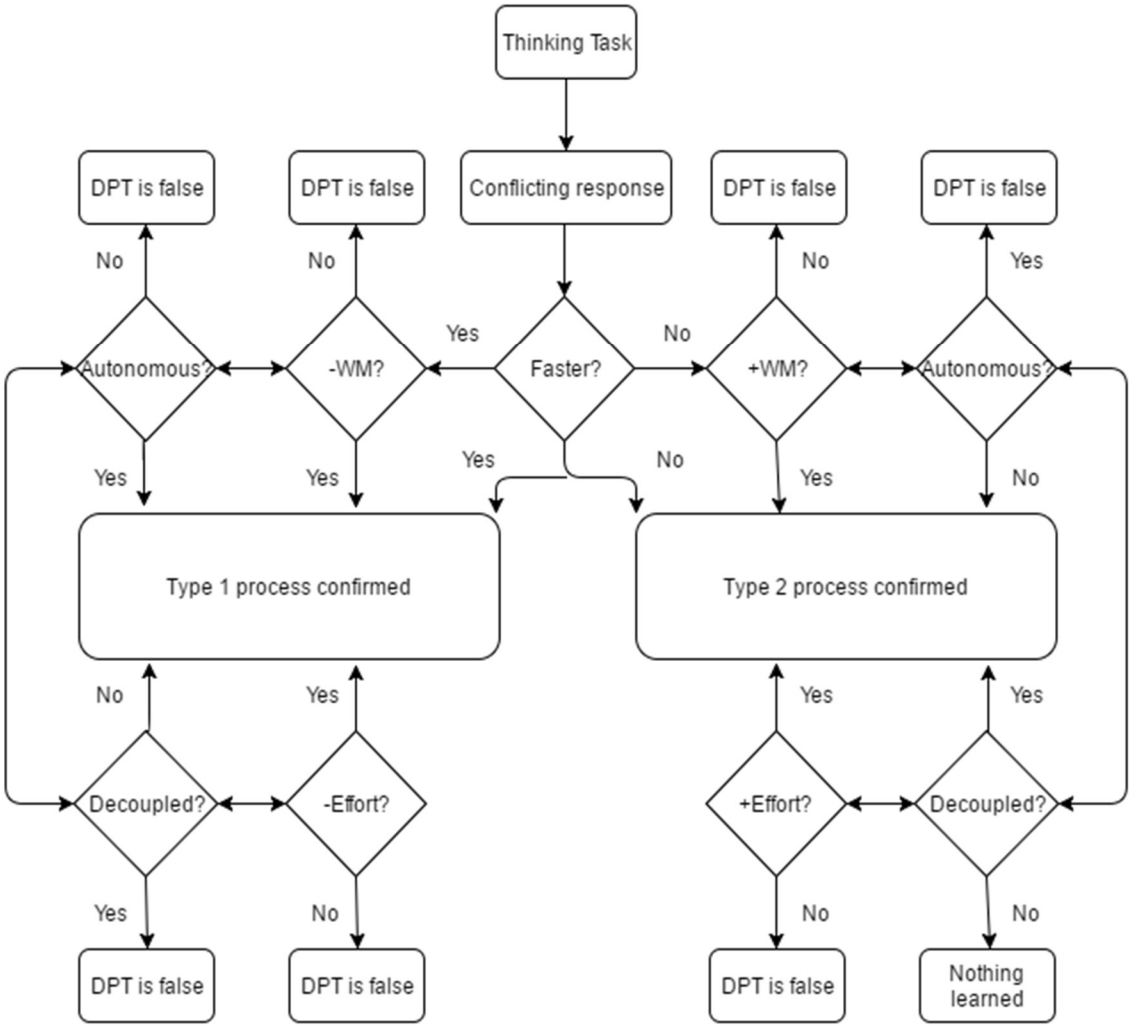

Image 1: Flowchart of how to identify a Type 1 or 2 process when comparing a conflicting response in a classic thinking task to another response in the same task. This is also a guide to falsifying Basic DPT.

For the second strand, it explains how to compare each feature. If some feature does not help individuate processes, the theorist can skip it and follow on to the next question. However, at least one feature standing to the contrary can falsify Basic DPT.

Although there is a certain order of analysis of features in this flowchart (speed comes first, then working memory, etc.), this is but one route. Any order can be taken and we suppose that each task might be done in a specific order. Also, in some tasks some features might not be verifiable.

If we are to take these defining features seriously, and if we want a stronger and falsifiable theory, then a few consequences follow. The first consequence is that these features must be the only criteria by which to truly identify a Type of processing. Other correlational features may be found and used as a guide, but only the eight defining cells can be decisive in determining the identity. Again, Evans (2012) seems to use the property of being implicit and explicit as one capable of determining identity. We understand that if he should want that, then he must move such feature to the level of defining features and explain the reasons for doing so (this would still be Basic DPT, except with extensions)[58]. If not, then listing defining features misses the point.

The second consequence is that Basic DPT predicts, and therefore only allows a few defining features to combine. That is, the alignment assumption holds for the defining features. Note that the use of the second step of the comparison principle is only allowed if identifying the type of processing by a single defining feature is fuzzy. If the combinations of defining features that are not allowed are clearly found empirically, then the theory must be false. Furthermore, some of these combinations are easily deduced from the flowchart, they are: if the process primarily responsible for the second response is 1) faster by comparison but loads more on working memory; 2) faster by comparison, loads less on working memory but not autonomous; 3) faster by comparison, loads less on working memory but makes use of decoupled representations; 4) slower by comparison but loads on less on working memory; 5) slower by comparison, loads strongly on working memory but is autonomous. 6) slower but less effortful and 7) faster but more effortful. Clearly there are more possible combinations, but these serve as examples. Notice that not being decoupled does not exclude the possibility of a slower and heavier process of being Type 2 and does not falsify the theory, since it is only a possible ability of Type 2 processes not a necessary requirement. By assuming the alignment assumption at least for defining features, the theory gains in predictive power and rigor. Therefore, the more defining features one assumes, the stronger are the empirical consequences; it will predict more but also be more easily false. At least for defining features, predetermined prediction must be possible, or else these features are not truly defining.

[58] The explicit and implicit features are central to the model we discuss in the resto f this thesis. But it could not become a defining feature without further discussion and clarifications, so it is not a defining feature of Basic DPT.

For instance, strategies that use decoupling in a task must be slower than ones that do not load on WM[59].

We must add that if falsification is achieved, such does not imply no DPT can work, only that this Basic DPT formulation does not hold. One can reformulate the theory to account for new evidence. However, if this happens repeatedly we should start losing our interest in DPT (see Lakatos, 1978). A final way of seriously damaging the theory is showing that the correlations of core correlational features do not hold or are so weak that they should not be grouped together.

When reviewing the reference problem we noticed that it bears little consensus. Understanding Basic DPT as a theory focused on rigorous cluster of features leaves theorists open to use the 'type' terminology without worrying about losing the point in the dichotomy. Again, since Basic DPT is to be built on, the consensus should be at the base and extensions could be proposed even for answers to the reference problem (such as Samuels, 2009; Lieberman, 2003).

Finally, these dual types resolve conflict in a default-interventionist fashion. As we reviewed it, it is the most believable architecture of conflict resolution. Various extensions are required in this point however. We need to know how they interact, when overriding will occur, if overriding is really a function of Type 2 processes or if a new mechanism needs to be explained, how monitoring of default responses occurs, among others. The general default-interventionist idea is also a very basic concept which really begs for further investigation.

We understand Basic DPT is still a very specific theory, it is no general DPT for all domains, however, we believe our following philosophical analysis are also extendable to more general ideas such as the difference between two evolutionary distinct minds. Therefore, our results will not be limited to Basic DPT for two reasons. First, because this Basic DPT is extendable (Evans, 2009, Stanovich, 2010, Kahneman, 2011) and second because there are more related general ideas to which our results could apply. Basic DPT, however, will be our reference and starting point for further ideas because it is now well defined.

[59] Indirect predictions in the sense of the meta-theory proposal are also kept (Evans & Stanovich, 2013b).

CHAPTER THREE:
INTUITIVE PREDICTIONS

3 Intuitive Predictions

From here on, our efforts will be focused on elaborating a framework, a broader story of the mechanisms that explain why there are Type 1 and Type 2 processes. Thus, we will present an answer to the unity problem as proposed by Samuels (2009). Fundamentally, we will propose that what distinguishes these two type of processes is that they are based on intrinsically different computational principles. The computational principles which underlie the features of Type 1 processing is predictive processing. While the computational principles that underlie Type 2 processing is classical symbolic processing. Therefore, rather than attempting to add new features to DPT (a burden that is laid mostly on empirical professionals), the job of this philosophical work is to provide an explanatory framework for each type of process, further developing on why they need to be seen as distinct, thus providing a richer story of how thinking works.

We will start by reviewing what predictive processing models are like (section 3.1), then we will make progress towards our goal by showing how predictive processing accounts for Type 1 features (section 3.2). Finally, we will proceed to show how predictive processing cannot account for Type 2 processes (section 3.3).

3.1 Predictive Processing

Marr (1982) and Gibson (1979) were researchers of very different views, but their endeavors shared at least one similarity. Although their work had the study of perception as a primary focus (mostly concentrated on vision) they also incited broader approaches of how to understand the mind. Marr (1982) understood the mind as a system that reads input from external stimuli in order to construct a rich inner representation of various aspects of the world. His model emphasized a bottom-up flow of information to communicate feature detection which culminated in higher representations. Gibson (1979) proposed there was no sense in reconstructing the world and rather that animals (human and non-human) pick up information directly from the environment for action, although the internal mechanisms for this was not much explored.

Predictive processing, as Clark (2013a, 2016) reads it, proposes a new shift to these approaches which also ends up generating a possibility for a broader framework to understand the mind. Predictive processing is a term that groups a strand of similar models developed by computational neuroscientists. The specific formulation of most interest to us is Hierarchical and Bidirectional Predictive Processing (HBPP) (Friston, 2003, 2005, 2008; Hohwy, 2013).

This mathematically formulated model proposes a form of processing where multiple layers of neurons are organized hierarchically (comprising higher and lower levels) that form a network with two major streams of information flow. The top-down flow is understood as conveying multiple predictions, each higher layer attempts to predict the workings of the one underneath it. The bottom-up flow conveys error correction on previously made predictions to each higher layer. Thus, despite having emphasis on top-down processing, the model proposes a bidirectional flow of information processing where the forward and backward flow bare unique and specified functions.

The upshot of such model is that the mind, instead of just capturing information from the world for action or representing the world in order to plan, is actually in the job of attempting to predict what will perturb it next. For an illustration, consider a person that is surprised by the absence of coffee of the table. That person's surprise is not caused by a perceived object but rather by a failure of prediction. His cognitive systems predicted coffee would be on the table, but no match was found, causing a surprise that resulted in further processing.

These are not necessarily personal predictions but rather multiple sub-personal predictions that even the visual system makes at various layers at each millisecond that passes. Rather than being based on a symbolic representation of each aspect of the world, these predictions are made on the basis of statistical information updated moment by moment. This statistical content is based on previous sensory states and the causes of these previous sensory states. This is possible since the system reads its own inner states and since it can establish causes[60] by proximal occurrence. Very roughly, It predicts that its own inner structure will probably be in a state Y given that it is in state X and previously inner state X was mostly followed by state Y.

60 Friston (2005, p.819) defines causes as "quantities or states that are necessary to specify the products of a process generating sensory information".

3.1.1 Generative models

Predictions are made on the basis of generative models. These models are not a picture-like representation of the environment or even a symbolic model, they are encoded in the form of probability density functions. A probability density function is a function that describes the chances of a variable having a determinate value based on the chances of occurrence of a range of similar values. Therefore, it is implied by definition that there is nothing absolutely determinate in the way the model is sustained. For an illustration, consider how Clark (2016, p.41) explains the difference of such encoding from traditional symbolic representations with an example:

> "Instead of simply representing "CAT ON MAT," the probabilistic Bayesian brain will encode a conditional probability density function, reflecting the relative probability of this state of affairs (and any somewhat-supported alternatives) given the available information. [...] At first, the system may avoid committing itself to any single interpretation, while confronting an initial flurry of error signals [...] This is typically followed by rapid convergence upon a dominant theme (CAT, MAT), with further details (STRIPEY MAT, TABBY CAT) subsequently negotiated."

Each level uses generative models computing the given likelihood of events to generate predictions about the level underneath it. These predictions are constrained by 'priors', which are previous probability distributions fixed before considering the evidence. The source of such previous probability distribution comes from previous learning which can range from evolutionary selection, developmental influences and other varying models at a higher level of the hierarchy. Whereas the idea of a prior is usually problematic for traditional Bayesian statisticians since it is hard to see what prior information we could have on data we study, in the case of predictive processing the idea of a prior fits naturally in the form of previous information the brain has gathered on some issue (following empirical Bayes, see Friston, 2003, 2005).

One can notice that there is not one generative model for the whole brain at any given time or brain state because models are updated constantly and because each level holds (at a given time *t*) a unique model. This update is primarily possible because of the forward flow (bottom-up) which communicates prediction error. This is a very strong and interesting shift that the predictive processing approach suggests. It proposes that the forward flow consists not of all the features that were detected and would be passed onwards to higher levels but only the error necessary to correct

and update models. Instead of conveying all information from the environment, rather it provides a natural funnel which guarantees processing economy by focusing on newsworthy information in the form of error correction.

3.1.2 Prediction error

Predictions flow downwards at each level and error correction flows upwards showing what exactly is faulty and needs to be corrected for future models. Therefore, lower levels will bring news flash since they detect the most recent error correction to propagate upwards, but the higher levels will have error correction coming from various other strands of the network. Thus, the higher layers will have models corrected from various sources and the lower levels will have tokens of newest corrections to be made, that is why at any given time there is not one generative model but various co-evolving models and also why there is a bidirectional flow of information. The upward flow has specific destinations while the downward flow has more ramification between areas, as found anatomically (see Friston, 2005).

Prediction error is also related to an important concept in this approach which is that of surprisal. Predictions are based on models which are a form of sub-personal expectation. When these expectations are not met, prediction error flags them with surprisal. Therefore, the system with its two directions of flow is always attempting to find a match from higher[61] expectations to the next information reported from the bottom. Surprisal occurs, therefore, when there is a mismatch between expectation and the information conveyed by error signaling. The goal of the system at every second is to minimize surprisal. To reach such goal it must constantly update its models in order to correspond to novelty. Having environmentally-tuned predictions enables the system to keep surprisal at the lowest level possible.

61 higher here does not mean cognitive (philosophical 'cognitive'), or conscious; in examples of visual processing they are higher areas in the visual cortex itself.

3.1.3 Free energy

One can have different ranges of mismatch, a minor mismatch might not alert the system in surprisal while major mismatches certainly will. To explain this variation (among other issues[62]) Friston (2010) formulated the free energy principle. Free energy is related to thermodynamics and information theory, but instead of entering in these more complex realms, let us understand the definition of free energy, in general, simply as energy that can be converted to work and as it is applied to predictive processing. So, according to Friston (2010), in Bayesian predictive processing: "[...] free energy is[63] just the difference between the models predictions and the sensations or representations that are predicted. Minimizing free energy then corresponds to explaining away prediction errors." Less free energy corresponds to having more stable models and less mismatch while having more free energy corresponds to requiring further processing[64] to explain away higher rates of surprisal. Explaining away means hypothesizing distal causes that accommodate the error signal.

The free energy principle states that "all the quantities that can change; i.e. that are part of the system, will change to minimize free-energy." (Friston & Stephan, 2007, p.427). This characterizes cognition as a process of minimizing its free energy. This means that cognition will always attempt to explain away surprisal, always polishing its generative models. Clark (2013a) reads this relation to free energy as putting more of the system's resources to effective work in order to minimize surprisal. However, a caveat must be added. When free-energy is at minimum, when predictions are working correctly is when effort is also at minimum, little effort is needed to navigate in an easily predicted world. And "using more of the systems resource" is usually an expression related to effort. Perhaps a better way of reading what this means is that when the system knows what to do with incoming information, when entropy[65] is low, that is, when prediction error is not too high, the systems will be put to effective work. However, when too much unexpected material follows from information, then systems will not know what to do with it, entropy will rise

62 It mostly aims at unifying different cognitive and even physical-biological theories under the same mathematical formulation.
63 This 'is' is how he understands being 'equal to', which to us would probably mean 'corresponds'.
64 This measure will be especially useful for us to suggest when Type 1 processes are being used and when Type 2 processes are called up.
65 A measure of chaos.

and free energy (as energy that can be put to work) will accumulate, new strategies will be needed to use such free energy and to solve prediction error. Thus, it must be noted that effort is higher when entropy is higher, not when entropy is lower. Not that Clark necessarily claims anything different, but the way he mentioned the relation to free energy might make the concept of effort confusing.

3.1.4 Precision

An interesting philosophical moral of the predictive processing model is that "What we perceive depends heavily upon the set of priors [...] that the brain brings to bear in its best attempt to predict the current sensory signal." (CLARK, 2013a, p.7). However, proponents of the model use the concept of precision to indicate that the system will not be doomed to be locked-in by its own expectations. That is because the value of bottom-up versus a top-down is decided in real-time. Responsible for the forward flow are error units, which are units specialized in error detection. For backward flow there are representational units which are specialized in certain features in order to predict lower level activity. The system is flexible in the sense that at any given situation, error units might be given more weight than representational units. In such case then, inner models would influence further processing and behavior less. Precision is the weighting of units. Precision enables flexibility which allows the system to focus on error when the model does not seem reliable. In contrast, when models are getting everything correct they are most reliable and more weight can be given to representational units. In such scenario the system can be guided directly by expectations which will make it faster and more successful. Precision is not calculated by some central system but distributed among layers.

Precision is the inverse of a signals variance, it measures how much such the signal fluctuates around its mean. Such measure can be used to decide whether the current information should propose a revision of prior hypotheses. It applies to the hierarchical economy which results in the power of strengthening what seems reliable for a given moment and attenuating what is misleading. By measuring uncertainty, it can boost the value given either to error signaling or to predictions. This is a major bet of the model because it explains (in Hohwy, 2013 and Clark, 2013a,

2013c, 2016): the mechanism underlying attention; how the top-down and bottom-up information flow is to be controlled; how interaction of multiple areas of brain including multimodal ones are to be governed; and even determine if, in a given moment we will act or just imagine. Further, various pathological conditions are explained as disturbances in precision, such as autism, schizophrenia, bodily illusions and more. We will not have space to cover all of these applications, but it is important to note that precision works as statistics of reliability, or inverse variance, that is, it is not a homunculi, but it might be questioned if it is effective for all these tasks (see Clark, 2013c). Variance is how much values (that are considered in contrast to one another) are spread apart, or differ. The more variance, the further values are placed from an expected mean, and less reliable they will be. Predictions (and errors) are weighted as a function of the inverse variance. So, more precise predictions will be weighted higher because their results are considered to be more reliable or less error prone.

An interesting example of how precising affects processing is in its relation to imagery. According to Grush (2004), when the system places a high level of weight in its model[66] it enables the use of imagery. Clark (2016) agrees with such idea and explains that each level has access to a statistical model of the level underneath, thus it is plausible that it can simulate properties without there being stimuli. Thus, when imagery is used is when response to stimuli is attenuated by precision and weight is given to what generative models more properly generate, a virtual version of objects based on what the layer has learned by previous occurrences.

3.1.5 Context-sensitivity

Another very important feature is that these systems are biased in order to be context-sensitive. As Clark (2013a, p.9) explains:

> [...] the best overall fit between driving signal and expectations will often be found by (in effect) inferring noise in the driving signal and thus recognizing a stimulus as, for example, the letter m (say, in the context of the word “mother”) even though the same bare stimulus, presented out of context or in most other contexts, would have been a better fit with the letter n. A unit normally responsive to the letter m might, under such circumstances, be successfully driven by an n-like stimulus.

66 Or raises its Kalman gain, in his terminology.

Since hierarchical prediction models work on the basis of probabilistic information they can manifest approximate predictions. This enables flexibility to context because although a unit might be responsive to the letter ‘n’ like in the example, if predictions suggest something related to the word ‘mother’ should come up next then predictions will bias this unit's normal conditions. Recall that information is encoded in probability density functions, which means that what is stored is a range of possibilities and not objects, thus, in a given process, token predictions (from other areas) are able to bias the system in order to favor a given side of the range of possibilities over the other. Since this bias is being emanated by top-down processes then large parts of the brain will bias each of its specialized systems with some coherence, tuning them to the appropriate token condition. This context sensitivity is achieved because in higher levels of the hierarchy communication between specialized areas are commonplace, and ambiguous objects are challenged with multimodal hypothesis. However, given that this is based on approximate information, it will not be uncommon for these systems to misbehave. In our view, this will be related to Type 1 awkward responses.

3.1.6 Active inference

One last important concept in predictive processing is ‘active inference’. Clark (2016) argues that even motor commands can be reduced to the prediction and error regime. Motor areas, as areas of proprioceptive perception, are always busy in expecting and predicting the sensory stimuli, but in this case the proprioceptive stimuli. Thus, action is treated as an outcome of perception of proprioceptive states. In action, the brain is attempting to minimize proprioceptive prediction error. A prediction concerning our future trajectory and position is generated and the highly weighted error relative to that prediction makes the action come about. Now, it is not that this triggers the motor command, motor commands are replaced. These highly weighted errors relative to future proprioceptive states already directly cause action. There is a tricky point in this proposal because both prediction (the goal) and error (the cause of action) are important and they need to be attenuated and increased in different levels of the hierarchy for action to come about. If errors relative to our future trajectories were attenuated we would only imagine our action.

However, errors relative to our current proprioceptive states must be attenuated for actions to come about, else we would stand still in checking our own place. Action comes about with highly weighted error relative to our future predictions and attenuated error relative to predictions on our current place. This treatment of action shows just how dynamical and complex the predictive processing account can become. Simply distinguishing between weighing errors or predictions is not enough, the level of the hierarchy where weighing occurs and relative to which prediction it is poised also determines function.

3.1.7 Caveats in our use of predictive processing

The evidence for HBPP does not come in a simple way in which we can satisfactorily discuss in two or three examples. Rather it can be pictured after analysis of various findings from diverse methods and over diverse range of tasks. Therefore, we will not have space to propose an elaborate discussion of evidence for predictive processing here (for some examination see Hohwy, 2013, Clark, 2013a, and Clark, 2016), but we will mention some of interest.

Consider Egner's *et al*. (2010) work on face recognition. They tested activity on the fusiform face area (FFA, known to process a low-level visual face recognition mechanism) over stimuli of houses and faces. Researchers induced controlled expectations (by matching colors) of house and faces stimuli. They found that although, as known, the FFA shows high activity when the subject is presented with face stimuli but not with house stimuli, the FFA had the same activation strength with the house stimuli as with the face stimuli when face expectations were high, but not when low. This is an interesting evidence and an example of how prior expectations influence even low-level systems given tokens of information recognition.

Hosoya *et al.* (2005, p.71) studying ganglion cells, which convey visual image to the brain, found that such cells signal not the raw visual image, but "departures from predictable structure, under the assumption of spatial and temporal uniformity" and that "They generally encode local differences in space and changes in time rather than the raw image intensity."

Knill and Pouget (2004) explain how evidence from cue integration function as evidence for representations in the form of Bayeasian probability distributions. Psychophysical experiments

show that integration (measured by the impact of each stimuli) from cues in different distances in vision and also in different sensory modalities (Ernst & Banks, 2002) tend to follow the same parameters predicted from Bayesian theory. Not only from distance cues but also by understanding motion. Weiss *et al*. (2002, p.598) explain that "psychophysical experiments show that humans also make some puzzling mistakes, misjudging speed or direction of very simple stimuli". They show how such mistakes of human motion perception are predicted by using probability distributions and Bayes' rule.

Finally, as for details of the predictive processing model, Markov *et al*. (2014) found that brains of macaques are organized in very distinct feed-forward and feed-backwards hierarchies with enumerable specific functions which correspond to some of those mentioned in HBPP. Also, modelling of electrophysiological data recorded during perceptual tasks has shown that precision could supposedly be encoded by a gain of error signaling superficial pyramidal cells (Brown and Friston, 2012). This has served as a useful measure of precision in other tasks such as perceptual evidence accumulation (for instance FitzGerald *et al*. 2015).

It is unlikely that we were able to give a rich enough explanation of the predictive processing approach, but we hope this brief summary will enable us to move forward in our goals. We also hope that readers will discover more about how we see its merit as we go through other topics of this chapter. Of course, mathematical details on how these Bayesian probabilities are encoded and processed in hierarchical models are not the topic for our sort of philosophy (further reading is suggested, see Friston, 2003, 2005, 2008, 2010).

We have no interest in holding a complete defense of a model for perception in this work. As far as we can tell, it still could be true that predictions can be only a part of perception and not the whole story as the proponents of predictive processing would take it. What matters, for us, is that at least a good part of perception involves predictions in the sense just explained and is it such part that is related to Type 1 processing in reasoning and judgment.

3.2 How predictive processing can account for Type 1 features

On the paper about his nobel prize for his works in heuristics and biases, Daniel Kahneman (2002, p.450) wrote that "From its earliest days, the research that Tversky and I conducted was guided by the idea that intuitive judgments occupy a position [...] between the automatic operations of perception and the deliberate operations of reasoning." Kahneman & Frederick (2002, p.50) claimed that intuitive thinking is "perception-like" and that "intuitive prediction is an operation of System 1". Further, that "The boundary between perception and judgment is fuzzy and permeable: the perception of a stranger as menacing is inseparable from a prediction of future harm".

Kahneman and Tversky (1982) have been speaking of 'intuitive predictions' for a long time. Kahneman and Frederick since 2002 have been explaining how there is a link between perception and Type 1 processing. What we hold is that link obtains because Type 1 judgments actually are perceptual predictions or hypothesis which work by means of HBPP. These authors have been noticing that intuition is somewhat like perception and have used the term prediction as what intuition does, but they failed to see how there is a unifying general computational model (from the work in predictive processing) of Type 1 processes that stems from the fact that they are predictions. They have been scratching the surface but no real application of the expectation-prediction mechanism, such as the one developed by the predictive processing approach, has been used as general explanatory framework for unifying Type 1 processing. This is what we propose to do now.

We will start with a brief general explanation of what it means that Type 1 processing occurs by means of HBPP. Then we will further show how this seems to be true by seeing how predictive processing respects and explains Type 1 features but not Type 2 ones.

Type 1 processes have sometimes been given input functions, say, following the early Fodorian model and the influence it has had on DPT. Perception also must have some input functions as well and traditionally has been understood as bringing precepts to mind, or as feeding input to systems in the brain. However, what really is interesting for DPT of reasoning and decision making is that Type 1 processes have an output function, in the sense that they generate an answer to problems. The predictive processing approach gives a clear output form to perception, by emphasizing its generative character. Thus, our strongest claim is that all Type 1 processing

answers (or its output function) are predictions based on probabilistic representations functioning as the HBPP explains. Predictive processing nicely accommodates the duality of Type 1 input and output functions. The representational units would be on the output side in the sense they function in order to fuel the generative model, while the error units would take the input functions usually ascribed to Type 1 processes. In this model then input and output are coupled, that is because input is really correction on generative modeling which formulate predictions (Type 1 answers).

The pivotal role of expectations for determining Type 1 predictions have gone mostly unnoticed despite the fact that task construal in the reasoning and judgment paradigm has been mostly a task of manipulating subject's expectations. Our model claims Type 1 processes take information over prior occurrences and over the current set of states (likelihood) and yields a prediction as fast as possible (posterior). If the time constraint is rigid, these predictions will generate actions (inner mental responses or, if too rigid, movements). If the system has time, then these predictions will be available for Type 2 evaluation. Thus, even what will be passed over to Type 2 processes already comes in the form of predictions. That is why manipulating subject's expectations in a task causes their Type 1 answers to vary accordingly and requires Type 2 effort to override them.

Type 1 processes are probabilistic. Both their computational strategy and the representations used are probabilistic. Content is encoded in the form of probability density functions, which means there is no symbol and no definite content, but values, means and standard deviation. That is why manipulating prior information alters its functionality, since it uses such to bias its workings into one or another direction, closer to or further from a certain value. A probability density function is a range of possibilities, and prior information will bias these possibilities so that those that are related to it will have higher chances of influencing answers. But it is not as if these values represent objects directly and discretely, they refer to distinct aspects of the input when perceptual systems are dealing with such objects. Not only is information encoded probabilistically but calculations, or inferences, follow Bayesian principles, this contrasts with following formal logic principles in computation, for instance. Of course all computation at a more basic level derive from logical principles but what they apply need not be further logical principle into content.

Finally, the sort of computation where Type 1 processes are executed are sub-personal (see Frankish, 2004, 2009) and their predictions are made by the same systems which process perception. A clear example is that a judgment (a prediction) about facial expressions is related to the FFA. The idea is that perception is not passive but already comes with predictions, and when in problem solving such prediction is precisely the Type 1 answer. We do not want to claim that Type 1 processes are purely perceptual (if in contrast to cognitive), only that such predictions stem from perceptual processes. So when perceiving a face which starts to show characteristics of anger, error correction makes the generative models adapt, and such models suggest Type 1 predictions which are quite like what we take to be judgments, such as (if put in words): 'this man is angry and might become violent'[67],

Kahneman (2003) and Kahneman & Frederick (2002) have argued that the list of features of Type 1 processing is shared with perception mechanisms. What we propose to do now is examine such list of features[68] showing that it is shared because both (or at least part of) perception and Type 1 processes work in the manner described by predictive processing.

3.2.1 The 'other correlational features' lines

Starting from the bottom we have that Type 1 processes are **older in evolution**. Predictive processing models are proposed to apply to animals just as much as to humans. Since we are claiming that Type 1 processes are predictions made by the same mechanisms of perception it should be obvious that they are indeed very old in the evolutionary timeline. Regardless of where you place the probable start of Type 2 processes in such timeline, it seems fair to say that meta-cognitive, explicit reviewing, recapitulation, symbolic and linguistic-like thought surely are more recent than predictions stemming from perception.

Friston (2010) wants to claim that predicting to minimize surprisal is actually a mark of life in general. Whether or not such claim applies, the point is that if it can even be taken as a serious

67 Of course, it only takes such an explicit form if it is reviewed by Type 2 processes, since Type 1 content is encoded probabilistically.
68 Using the chapter two list, the reader is advised to follow the list in Appendix II.

hypothesis to life in general that should at least suggest that Type 1 processes cannot be recent. Finally, Gowaty and Hubbell (2013, p.35) also want to make similar, more direct, suggestions: "[...] animals predict their futures and act as though they are indeed perceiving and responding to intertwined set[s] of probability density distributions".

The job of Type 1 processes under the current reading is to make sense of the world in 'perception time'. Making sense of the world requires context.. Unless in very specific situations where abstract content are relevant, say perhaps a math test, generative models will take any cue they can in order to propose the most **contextualized** information. Working with content is how Type 1 processes usually derive their predictions. If the system has been in a given state, which relates to given contents recently, then constraints from such prior information should bias predictions. This contextualized bias in Type 1 processing probably stems from the fact that we live in a world which needs to be made sense of in terms of its concrete objects and how one relates to the other. Of course, this in turn is figured out by inner statistics. We will return to the issue of how Type 1 processing can deal with contexts in the end of the next chapter. For now, what matters is that the predictive processing approach attempts to explain the mind as being mostly contextualized, as Type 1 processes are claimed to be.

Being **parallel** fits really nicely with the framework we are developing since predictive processing is a model of computational neuroscience which studies neurons and their networks as they exist in the brain, which are parallel by a matter of fact. If our framework proves to be of use, then it suggests moving the parallel feature to at least the core ones, if not defining. HBPP would make no sense at all if not parallel. Not only is it supposed to be structured in parallel architecture but its computations are essentially parallel in the sense that error correction and generative models across the system do not wait for any input of other systems, they are constantly generating predictions in multiple systems, in multiple layers, they are always attempting to be faster than the world, using serial strategies would seem pointless. Recall this was already a central feature of Sloman's (1996) dual process theory.

3.2.2 The 'core correlational features' lines

In chapter one we saw that what is termed **intuitive responses** is explained either by Darwinian algorithms, relevance (in the sense of Sperber and Wilson, 1986) or heuristics. It is DPT's merit that of being a framework that (somewhat) unites or permits these various types of explanations. We will show how predictive processing can accommodate these senses and how it accommodates intuitive thinking.

A central theoretical criticism on predictive processing and the free energy formulation is the dark room problem (see Clark, 2013a, 2016, for a debate). The problem can be formulated as a simple question: "if organisms seek surprisal minimization why are they not attracted to ultimately stable states such a dark room?". The common answer is that evolutionary constraints on processing guide the organism to seek needs (such as food) which will not be found in a dark room. For those like Friston (2010) who want to reduce the functioning of the brain to simple principles these extra constraints do not sound attractive[69]. After all, these constraints are varied and not mathematically reducible to a single account. For our goals on the other hand, evolutionary constraints on predictive processing is how evidence for Darwinian algorithms can be used in our favor. Theoretically, it makes perfect sense to understand these Darwinian constraints on predictions as hyperpriors, that is, as general evolutionary principles that establish limits on the network's freedom in order to guarantee the basic tuning with the world's requests and gene interest (in the sense of Dawkins, 1976). For instance, that objects cannot be located in the same place at the same time (Example in Clark, 2013a). We can add (to accommodate evolutionary psychology[70]): to store input in frequency format to make judgments based on them; to take

69 Now an interesting philosophical response might be that in a dark room error is already minimized, the agent does not seek a state where there are no errors to minimize, it seeks the active state of error minimization. There seems to be a difference between seeking minimization and seeking minimal surprisal states. If the organism seeks error minimization then the dark room scenario is actually repugnant simply because there is no error to minimize, to be locked in the dark room the system would need to seek minimal surprisal states. Prediction minimization is found neither in extremely chaotic nor extremely ordered environments. See goldilocks effect on Clark (p.266, 2016).

70 of course this is a change in important theoretical commitments such as massive modularity, however, using hyperpriors has very similar effects, and explains the same phenomena. Also hyperpriors can also be learned, but the discussion here if they were innate or learned very early is far less important compared to the significant (agreed upon) constraining and impermeability effects. Also, Clark (2016) believes hyperpriors should be kept at minimum and that they are about general world occurrences and not specific ones. So adopting evolutionary psychology would imply pushing the hyperprior to do much more constraining that Clark would want.

phenomena including food and mating as high precision candidates; to bias predictions of social contracts in order to spot cheaters. These hyperpriors are fixed constraints which take certainty (even if prone to error) over probabilistic revisions. The system is set so having these hyperpriors is a priori safer than establishing novelty on them based on world's contingencies, therefore they will constrain predictions in the same way.

A challenger, influenced by Cosmides (1996), could probably argue that evidence shows Type 1 processes works with frequentist probability and not Bayesian probability. But what evidence shows is that content represented in frequency formats are easier to process then content represented in, say percentage formats. But there is no reason to suppose that a neural network that works with Bayesian probability could not be trained to recognized frequency formats better than percentages. The actual underlying form of processing of the system has nothing to do with what type of input format it prefers. The frequency hypothesis understood with good sense only says evolution trained brains to look for input in frequencies, it says nothing about principles used in its programs. In fact, the frequency hypothesis is actually accommodated because we should expect that formats of occurrence are encoded by perception (by occurrences) helping it yield its predictions better than formats which require decompositions and abstract knowledge (as Kahneman, 2003, 2011 shows with his examples from perception). Another way to put it, is that these Bayesian calculations are over the statistics of perception, the rules that govern such statistical analysis are not available for problem solving. That is, after all, what the massive research of Kahneman and Tversky has shown all these years. In the same way, we cannot use the rules used in image coding in the visual system to help us in our math exam. Even our sub-personal judgments cannot use all of the rules that neurons do. However, this does not mean there is no cognitive consequence or differences depending on which rules neurons use to process information. Effects of predictive processing being probabilistic will be felt in all domains, but this does not mean that judgments (even those that stem from probabilistic processing) will have the power to use all of the rules applied by neurons in predictive processing to help in problem solving, neither that optimal Bayesian inference will be available for problem solving.

The heuristic nature of predictive processing is something Clark (2016) attempted to defend in his book. Consider this example:

> "[...] rather than attempt a full examination of reviews and menus for every restaurant within a five-mile radius, we might very well choose one that a trustworthy friend mentioned yesterday instead. We do so reasonably confident that it will be good enough, and thereby save the temporal and energetic costs of taking further information into account." (CLARK, 2016, p.245).

Heuristics both in Kahneman's sense as well as in Gigerenzer's sense can be understood in predictive processing as simplifications the system use in order to act more feasibly and quickly. That is because, as Clark (2016, p.1) argues "the brain [...] is an engagement-machine [...] that is perfectly positioned to select frugal, action-based routines that reduce the demands on neural processing and deliver fast, fluent forms of adaptive success." Brighton and Gigerenzer (2008) are concerned that the probabilistic mind might go against their heuristic approach. But the worries they state are optimality and tractability. However, what Clark (2016) proposes is precisely that predictive processing does not need to be understood as processing optimally, because sample number, time constraint, feasibility and physical limitations do not allow for such. But it really depends on the notion of optimality and in what level it is being applied. Also, the predictive processing approach is becoming famous in computational neuroscience precisely because its mathematical formulation is tractable, and its computational powers seem to allow for feasibility in previous intractable domains, such as context sensitive problem solving (see McClelland, 2013, Phillips *et al.*, 2015). Thus understood in bounded rationality terms (especially in Clark, 2016), predictive processing is perfectly compatible with the use of heuristics to facilitate problem solving.

The relevance principle (Sperber and Wilson, 1986) surprisingly also fits in a very interesting way. Recall from chapter one how the principle states that relevance will come from the correct mix of effects and effort: speakers will attempt to achieve maximum contextual effects with minimum effort. Now thinking of what this means in predictive processing we have that if the speaker informs the hearer of something it would already predict, no newsworthy information is transmitted, no contextual effect, irrelevant content. If the speaker passes information in a way which will require effort (too much error) for the hearer this means it will not be predicted directly, prediction systems will not know what to do with it (free energy will accumulate), and further processing will be necessary. The speaker must, therefore, find the set of information which will be both newsworthy and easily accommodated, it must inform something like 'non-entropic error', where entropic error is that which will require too much further processing (effort) to accommodate

(increase in free energy). This is what the relevance principle says in predictive processing terms: relevant speech is that which causes prediction error but not so much error as to accumulate free energy.

Finding solutions in conflicting tasks to reach normativity would result in too much free energy, and following the relevance principle suggest doing otherwise. Also, most of the pragmatic explanations (see Adler, 1994 and Sperber, *et al*. 1995) are related to the expectations people have when reading the reasoning tasks. Thus, they believe task construals will determine difference in results. In predictive processing this is accommodated because different expectations directly generate different outcomes.

We can see that various explanations that were embraced by DPT, such as Darwinian algorithms, relevance and heuristics, that is, mechanisms that were attributed to Type 1 processing can be interestingly accommodated in predictive processing. Obviously a few theoretical commitments must be reviewed in order for this to work, especially in evolutionary psychology. We still need, anyhow, to choose one explanation over the other when they are concurrent in specific explanations of evidence. However, this is the same burden DPT already had to carry by being an integrative framework. Now we have a higher-level framework to understand how and why these explanations are coherently linked, the label "Type 1" by itself did next to nothing for such unification and this might even suggest (since we have a hypothesis for their mechanism) ways of knowing when each explanation is best.

Next is '**uncorrelated with cognitive ability**'. We think the evidence that shows Type 1 processes are less prone to individual cognitive ability differences is very convincing. In various tasks, the non-normative but intuitive answer is ubiquitous among people. If most people seem to give the same intuitive answers, then cognitive ability and individual differences cannot be playing a fundamental role in these answers. Also, of course, cognitive ability is correlated with the overriding of these intuitive answers by normative ones (Stanovich and West, 1998a, 1998b, 1998c, 1999d). In our framework this means that cognitive ability tasks are tracking something other than predictive processing. Further, consider perceptual illusions. Are they not mostly independent of cognitive ability? Precisely. Type 1 answers that do not accord to normativity are like perceptual illusions because they are processed similarly, by hierarchical prediction mechanisms. The task for

Type 2 processing would be to override priors whenever they are not fit. Only this latter task would correlate with cognitive ability measures.

One might ask, does not the predictive processing framework predict that since contingencies will mold expectations that there should be individual level expectations? Yes. However, we might ask, what reasoning task would be able to identify individually varying expectations? In various tasks there is an insignificant variation of answers which are not intuitive or normative, these could be varied by individual expectations. Because of the fact that they are individual no general principle can be generated to account for them. Individual differences that are identified are in the sense of capacity, of being more or less able. Further, it is occurrences which guide inner statistics and the same mean should be available for everyone, since we live in the same structured world, such mean would be responsible for the great similarity among Type 1 responses in tasks.

Type 1 processing is said to be usually **unconscious**. This seems in line with the basic story of predictive processing: "[...] conscious guessing, is not the kind that lies at the heart of the story I shall present. [...] It is the kind of automatically deployed, deeply probabilistic, non-conscious guessing that occurs as part of the complex neural processing routines that underpin and unify perception and action." (CLARK, p.2).

We know we do not have conscious access to the multiple alterations that the visual system makes to visual input. But how could we? Predictions are occurring at every millisecond in various layers in the most diverse systems of the brain, conscious access is limited to a few important items. We are mostly unconscious of the results of predictions our systems make because they are the standard non-newsworthy material. We do have conscious access to some predictions but only when they are informative, when they need altering or whenever they need Type 2 reasoning to act on them. It might even be that some hyperpriors are set to never be altered and those are the ones we can never be conscious of. Thus, predictive processing will occur mostly without notice from the conscious agent.

We mentioned in chapter two that despite the implicit and explicit distinction being used vastly in the sense of access, that is actually the also use of 'conscious'. When it comes to the **implicit** and explicit distinction what is unique and coherent (even with the word) is the

representation format. We have seen that predictive processing has a very unique representation format. This difference in format is the way content is encoded in predictive processing, which is by means of probability density functions. As we have seen these functions do not even disambiguate items discretely, rather, they gather multiple occurrences of events and possibilities, along with information from various areas of the cortex together to generate probability. We propose this is actually the (usually unexplained) meaning of an implicit format, one that encodes probability of occurrences and not representations by means of symbols. This implicit format is not the type of format Type 2 reasoning can work with, Type 2 processes need symbolic, unit-like objects to reason over, and that is the meaning of an explicit representation format: disambiguated, stand-ins for a unified object[71].

High capacity is another correlated feature of Type 1 processes. It means there could be no countable limit on the number of items that predictive processing mechanism work on. And it is true. Predictions that co-occur in the same time-scale as changes in the environment cannot be so limited if they are to be of any use. This much should tell us that predictive processing cannot occur by classic symbolic mechanisms, as it would generate the frame problem (discussed in the end of the next chapter). We have a special section for pondering over how HBPP could be tractable. What matters now is that, however such feat is accomplished, it *must* deal with a huge flow of information in order to keep pace with changes in the environment. “Such systems exhibit powerful forms of learning [...] deliver rich forms of context-sensitive processing and are able flexibly to combine top-down and bottom-up flows of information within the multilayer cascade.” (CLARK, 2016, p.26).

Automaticity as was covered in chapter two concerns overlearned skills. Overlearned skills here can be understood as predictable ones. Let us use the classic example of learning how to drive a non-automatic car to see how predictive processing relates with automaticity. When we first sit behind the driver's wheel, even if we have knowledge on what must be done, our systems cannot coordinate all such knowledge in order to be useful (and safe). When we train ourselves the correct order of using gears, wheel turning and pedals we are tuning our predictive processing systems to

[71] We will come back to this in more detail in the next chapter.

the usual occurrences of car handling. Of course, before driving our systems cannot have useful priors on how to do it. So by letting our predictive processing engage with the stimuli necessary for driving we tune it to that particular context. For instance, here in Brazil, we are put to learn how to control non-automatic cars in cliffs. Our systems need to predict the moment when the clutch is ready to push the car at the correct strength to manage the cliff. But not only this, our systems need to predict more precisely when another car is stopping in front of us. They need to predict the order of gears and when they will be necessary, they need to predict that the car is being wrongly used just by clues on the sound it is making. There are various cues which need to be used to predict near-future occurrences. The system needs to know for various states that if it is in a given state, another given state is the most probable to follow. Once the system learns various important cues that lead to efficient predictions, it can handle most driving abilities automatically. Thus, an experient driver will incur in far less surprisal instances than a novice. In fact, the higher surprisals which will come by are in the form of unpredictable changes in the environment, such as an animal crossing the road. In contrast, the surprisal which will mostly concern a novice is in terms of his actions to handle the machine so an animal can go by unnoticed. Since precision is given to error correction in terms of body movements, precision on error concerning the road ahead might be attuned. If our systems have no useful priors for driving, they need to rely on effortful controlled skills to train predictions systems. As we shall see further on these effortful controlled skills cannot be predictive processing skills themselves.

Again Clark's (2016, p.257) version of predictive processing is described accordingly:

> "Fast, automatic, over-learnt behaviours are especially good candidates for control by models taking a more heuristic form. The role of context-reflecting precision assignments is then to select and enable the low-cost procedural model that has proven able to support the target behaviour. Such low-cost models [...] will in many cases rely upon the self-structuring of our own information flows, exploiting patterns of circular causal commerce (between perceptual inputs and motor actions) to deliver task-relevant information 'just in time' for use."

Another way to put it, which fits neatly with the framework we are developing is: "we need only note that very low-precision prediction errors will have little or no influence upon ongoing processing and will fail to recruit or nuance higher level representations." (CLARK, 2016, p.148) That is, if the task is overlearned and errors are weighted as low, systems will act without further

recruiting. This can be understood as hypothesis for automaticity, which has been used so much in psychology but without an explanation for why it differed from controlled processing.

3.2.3 The 'defining features' lines

Low effort is associated with Type 1 processes. Folk psychologically, are we in a situation of effort if everything is going according to plans? No. By contrast, if nothing is going as expected we need to alter ourselves to think and act accordingly. Likewise, in predictive processing, when prediction error is at minimum, little effort is needed. As Clark (2016, p.116) explains: "Forward models provide a powerful and elegant solution to such problems, enabling us to live in the present and to control our bodies [...] without much sense of ongoing struggle or effort." In our reading of Friston (2010), in cases of low effort there is little free energy to be put to work, since systems know what need to be done and are proceeding as they usually do, feasibly minimizing free energy. In a situation of little effort default predictions are enough to accomplish the tasks at hand. In contrast, when systems are constantly failing to predict, more effort is needed to put energy to work, to achieve a new solution to the problem. This new solution will not follow what is predictable, it would now need precisely the opposite, to work on the unpredictable, some different mechanism is necessary, and it is this latter one which is effortful. When the system can control most of its tasks by means of predictions, which corresponds phenomenologically to when subjects 'already know', those are the cases when little effort is needed.

Clark (2016, p.250) claims "Cheap, fast, world-exploiting action, rather than the pursuit of truth, optimality, or deductive inference, is now the key organizing principle." A cognitive architecture that proposes cognition attempts to predict the incoming information surely must have a recipe for being **faster** than other proposals. A predictive processing architecture can act faster because any cue captured from the world is readily met with predictions (even if bets) concerning a lot more than the cue itself shows. The predictive processor is always taking certain bets about what the current state of the world implies, losing accuracy in compensation for speed. So it fits really nicely with the idea that Type 1 processing needs to abandon certainty and accuracy for speed, an idea previously developed as quick and dirty heuristics. Predictions are also quick and

dirty and perhaps in a way that makes these properties even more ubiquitous since it spans even perceptual details and not only judgments. Thus, when watching a white scene in a movie, there might be guesses that there are no black and brown pixels in some areas of the screen, even if there are. The quick and dirty guessing thus extends far beyond what traditional frugality theorists had been considering.

Another property that allows for fast processing is predictive coding (Rao and Ballard, 1999). By predictive coding we mean specifically the property of these system to consider, from the world, only stimuli which result in greater prediction error. Thus, some stimuli are considered in real-time perception already as irrelevant for the adaptive use of the organism. Precision weighting quickly investigates the size or effect of the prediction error determining if it is eliminated or if it needs to further propagate to other areas. Focusing on prediction-relevant stimuli only permits the agent to quickly decide courses of action and to select amongst possible 'affordances'[72]. Type 1 processing can thus be understood as quick predictions emerging from the system's first considerations of these errors.

As Clark reads it, predictive processing proposes the agent is always tuned to environmental cues which can quickly help the system decide between affordances. The predictive architecture provides means for quicker selection, "allowing time-pressed animals to partially 'pre-compute' multiple possible actions, any one of which can then be selected and deployed at short notice and with minimal further processing." (CLARK, 2016, p.180). We understand that in the cases studied by DPT, mostly of people taking reasoning and decision making tests, this quickness of action comes in the form not of body movements but of simplistic hypothesis quickly springing to mind. Such hypothesis come to mind quickly because of the statistical relations they bear with the input. We can thus start to ponder about the basis of accessibility, which worries Kahneman (2002, p.456) "much is known about the determinants of accessibility, but there is no general theoretical account of accessibility and no prospect of one emerging soon." Accessible content could now be understood as the higher values in probability density distributions related to the range of possible responses to a given task. The more certain values have been used to reduce prediction error in the

72 Or possibilities for action. See Gibson, 1979.

(evolutionary and developmental) past the more the content will be accessible. This can be measured in terms of precision of prediction for given tasks[73].

Next we have **autonomy**. Autonomy is a property which stems from DPT's inheritance of Fodorian mind duality. As reviewed in the introduction, Fodor proposes a division between a central executive (which is much related to the working memory proposal in DPT) and input modules which are autonomous, they act independently, and in Fodor's term do not receive input from other modules. In this framework, we are pursuing a new reading of Type 1 processing instead of that of symbolic processing input modules. So do we lose the concept of autonomy? No, although we need some reworking on how it fits the predictive processing architecture. To be fair we will keep Stanovich and Toplak's (2012) definition of it: we can say that Type 1 processing is autonomous because "the execution of Type 1 processes is mandatory when their triggering stimuli are encountered, and they are not dependent on input from high-level control systems". So despite Fodor's influence, there are only two properties that interests dual process theorists when it comes to autonomy, that autonomous processes respond in mandatory fashion and that they can act independently from higher input from other areas.

We need to understand four concepts in order to see how Fodorian modularity has been shaped, and see how predictive processing treats interaction between systems: that of functional differentiation, soft modularity, hyperpriors and safe penetration.

Anderson (2014, p.52) notes that replacing rigid modularity does not necessarily imply "that regions of the cortex — much less subcortical structures — remain functionally undifferentiated". Although he worries the concept of specialization can limit a certain neural region to function only in a given range, say that of emotional recognition, he proposes to continue speaking of functional differentiation. Functional differentiation suggests regions will be locally biased towards some response profile and "a region's response profile will certainly reflect its underlying functional capacities and determine the role(s) it can play in various functional coalitions". He understands these as interactive differentiations because although they are biased to a certain response profile they might be used in networks and coalitions for other related

[73] Of course biologically realistic measures of precision are available only for (as far as we are aware) visual perception (Brown and Friston, 2012), so further measures would need to be developed.

purposes as well. Anderson (2014) makes the case to show how functional differentiation is very well suited to neuropsychological lesion evidence including double dissociation and we recommend his treatment for further interest in details.

Soft modularity as it is used in predictive processing means modularity as regulated by precision weighing. Functional differentiation has an important part to play, since, a given brain region will act autonomously if it has been doing so very reliably in the past and if current tasks are commonplace. Thus, a soft module can act as a Fodorian module if its success is trustworthy given the current problem. However, there are two clear additional changes. First, soft modules can also be formed and reformed depending on the task, these, called Transiently Assembled Local Neural Systems or TALoNS (Anderson, 2014) are composed of coalitions of different networks. "Distinctive, objectively identifiable, local processing organizations [...] emerge and operate within a larger, more integrative, framework in which functionally differentiated populations and sub-populations are engaged and nuanced in different ways so as to serve different tasks." (CLARK, 2016, p.150). Thus, if more fixed functional differentiated regions by themselves are not able to solve the task reliably (as determined by inverse variance), coalitions can be formed 'on the fly' to attempt another local resolution. An illustration could be a soft module that applies a multimodal effect on the interpretation of perception, such as when we become more immersed in a virtual reality (using current technology) because brain coalitions take haptic stimuli to corroborate visual 'hypothesis'. So one can see that even this second step (of an emerging modularity, in contrast to fixed systems) is still autonomous in the sense of DPT (as defined by Stanovich & Toplak, 2012). Because it is free from resource constraining, time-consuming, higher-level interference, such coalition still results in an unconscious, quick, working-memory free and effortless processing. Of course, one might argue that there might be such soft coalitions that do demand higher level input, if that is true, then, as we will respond later on, those are the kind that will be under scrutiny of Type 2 processing, precisely, they lose their autonomy.

Although there are soft modules that are assembled and disassembled, there should also be regions that are more constrained to their functional differentiation.[74] This is because they assume

74 This does not mean they do not participate in other coalitions. While *hyperpriors* are fixed because they have free pass over precision, as far as we know no *layer* is a priori impenetrable.

more hyperpriors than others. The visual cortex is a good candidate for such. Hyperpriors are general hypothesis that can be innate (in the sense of evolutionary psychology) or acquired by statistical learning (of the kind predictive processing describes) and gained a higher status because they have been working so well. Or as Friston *et al.* (2013, p.1) put it hyperpriors are "prior beliefs about [...] precision".[75] That is, they constrain precision weighing itself, thus they have a free pass over contingent judgments of reliability. There seems to be various of these hyperpriors related to vision, if we were to state them in language: 'that two objects cannot stay on the same space at the same time', 'that the world is composed of one coherent state or another', and 'that such a state must be constantly changing'.

Not surprisingly the visual system is the best example used by Fodorian modularity. In particular the classic version of the Muller-Lyer illusion is taken as an illustrative case. Clark (2016, p.200) explains how predictive processing can understand such illusion:

> "perceptual systems are well-calibrated as devices for mediating between sensory stimulation and action, and their deliverances (though subject to alteration by extensive re-training) are not simply overthrown by our endorsement of sentences such as 'yes, the two lines are indeed of equal length'. Endorsing such a sentence [...] does not adequately account for the full spectrum of lower level predictions and prediction error signals that construct that particular percept, so it is unable to overturn the long-term learning of the system." (CLARK, 2016, p.200).

The language input, even if conscious, is just one hypothesis which is not given as much weight as other perceptually convincing ones. In contrast, if stimuli are ambiguous, that is, visual predictions are unable to accommodate them, then even a sentence can help disambiguate. This occurs, for instance, when a priming word helps disambiguate a visual image. A famous example of this case is the dalmatian dog image (which appeared in Marr, 1982). The image looks to most people like a random splattering of ink but once they are *told* there is a dog to be found their recognition and experience change drastically. Also, once people find the dog, it is hard to see the image as random again. This suggests that once the visual system has stumbled upon a stable hypothesis (reliable) it stops accepting input from higher-level processing. Fodor (1983) would argue that this might be done by central contextual systems and not by encapsulated systems. What

75 Or even: beliefs that constrain even hyperparameters, which are the parameters that govern precision weighing.

happens is he makes his hypothesis less and less tangible and testable, and also less relevant. So safer penetration is cognitive penetration that takes into consideration: the existence of hyperpriors; the reliability of the hypothesis proposed by the soft module and the reliability of the impacting (penetrating) input. Based on how these determinate parameters are placed, soft modules may or may not be subject to higher-level interference.

Autonomy also has a second (but deeply related) property, that of being mandatory. The use DPT makes of the concept of a process being mandatory is by how it cannot be interrupted once it has started. Once a mandatory process faces its triggering stimuli it runs to completion. Now, this might seem at odds with the overriding function of Type 2 processes. If Type 1 processes are free from higher-level input and will not stop running then how could they possibly be overridden?

Answers to this question helps us understand what use DPT can make of the concept of autonomy. First, it is not that Type 1 processes will never receive input from higher-level systems, the point is that they are *not dependent* on (in contrast to Type 2, processes which are), and in most cases will act independently. This seems to fit really nicely with the reshaping of modularity proposed by predictive processing. That is, Type 1 processes will act independently, unless they are unreliable, which then makes them susceptible to influence from other regions. Second, Type 2 processing is usually said to be able to help train and re-train Type 1 processes, so some influence must be acceptable. Finally, and more directly, if these processes can be overridden but must run to completion, then they must either be stopped before starting or after they are done. Stopping them after they are done means not letting them influence behavior or processing any further. This seems to be most plausible interpretation of overriding. It most likely seems to be the case that we cannot stop Type 1 processes from starting either.

What all of this implies, we understand, is that Type 1 processes are not symbolically decomposable, one cannot override certain steps in its processing but not others. This, we argue, is because Type 1 processes are not expressions with neatly defined constituents but statistical predictions. Another way to put it is that priors that will govern how Type 1 processes will respond to a certain stimuli are already in place before the stimuli arrives (this is also what allows their predictive character).

Are these changes great trouble for DPT? Well, let us see how Stanovich reads Fodor. Fodorian criteria is listed in Stanovich (2004, p.38) as such: 1. Fast, 2. Mandatory, 3. domain specific, 4. informationally encapsulated, 5. cognitively impenetrable, 6. subserved by specific neural architecture, 7. subject to idiosyncratic pathological breakdown, and 8. ontogenetically deterministic (undergo a fixed developmental sequence). Where predictive processing argues for functional differentiation and not specialization and that soft modules can be formed and reshaped, Stanovich (2004, p.38) has similar worries (where TASS is a name for Type 1 processes):

> "Properties 6 to 8 follow from Fodor's (1983) emphasis on innately specified modules. However, they are not part of my conceptualization of TASS because, although innate modules are an important part of TASS, my conceptualization deems it equally important that processes can become part of TASS through experience and practice. In short, processes can acquire the property of autonomy."

Likewise, where predictive processing argues for safe penetration, Stanovich holds his judgment:

> "Whether a particular subsystem is informationally encapsulated or not—and thus whether or not it qualifies as a Fodorian module—is a frequent source of debate in cognitive science. In contrast, properties 1 and 2 are much less controversial, which is why I have emphasized them as central features of the TASS construct." (Stanovich 2004, p.38).

In conclusion, not only is predictive processing's accounting of modularity and autonomy compatible with DPT's[76] position, it seems to provide a new theoretically and philosophically rich framework which was absent after DPT's rupture with strict Fodorianism.

Finally, we arrive at less use of **working memory**. Suppose we lived in a simple world where predictive processing always worked. Every change in such world would come as no surprise to the agent. Priors would guarantee success in understanding even before problems were posed. What effort would this agent have to put into acting in this certain world? Not very much. What would he pay attention to? It is hard to say. What would he think about? He probably would have to daydream or confabulate about things. Daydreaming seems to be turning attention and effort to oneself and forgetting the world for a while. What seems to happen to attention and working memory in predictable situations is that it turns inwards, it starts to generate novelty or

76 Remembering that Stanovich is not only one of the main defenders of the theory but also the one who proposes autonomy as a defining feature.

monitor inner performance. This is actually observable in habituation, a phenomenon much known by psychologists (noticed by Sokolov, 1960) where exposure to repeated stimuli decreases attention paid to it. However, in our simple world thought experiment, we allow agents to turn inside and study their own inner uncertainty and generate thoughts. But if we want this world to be truly simple even inner processing could be so predictable causing the agent to habituate over not just outer stimuli but inner processing too.

This thought experiment is just a way of illustrating how working memory is an online and ever-ready mechanism for dealing with further uncertainties and unpredictable information. We will have more to say on working memory in the next chapter. What we want to note for now is that the more predictable a given state is, the less working memory resources systems will consume in processing it. Working memory is needed when predictive processing fails. Since this is a negative feature 'less working memory' there is little we can say positively about it. Thus, we will in the next sections come back to working memory as unrelated to predictive processing.

3.3 How predictive processing fails to account for Type 2 processing

We have seen now how predictive processing fits naturally with various features ascribed to Type 1 processes. However, it is now that we part ways with Friston, Hohwy, Clark and company. That is because they make the same mistake that other older cognitive scientists have, by supposing there is a simple general computing framework to account for all types of cognition. We think predictive processing is somewhat incomplete as the classical architectures were. We propose that, at least for DPT of reasoning and decision making, a dual framework is a better option, where predictive processing accounts for Type 1 processing and where the classical architectures as proposed by Newell (1980) accounts for Type 2 processing.

Instead of showing now how classical accounts deal with Type 2 processing, we leave that for a further chapter and focus on how predictive processing fails to account for Type 2 processing. We will start by showing how alleged attempts of predictive processing accounting for DPT fails (section, 3.3.1), then we will focus on some features of Type 2 processing that seem at odds with

predictive processing (section 3.3.2) and finally state a few theoretical reasons to explain why Type 2 processing is incompatible with predictive processing (3.3.3).

3.3.1 Predictive processing approaches to Dual Process Theory

Andy Clark (2016) directly touches on the subject of DPT by promising to account for it in subsequent parts of the book:

> "[…] some theorists (see e.g., Stanovich & West, 2000) have suggested a 'two systems' view that posits two different cognitive modes, one (' system 1') associated with fast, automatic, 'habitual' response, and the other (' system 2') with slow, effortful, deliberative reasoning. The PP perspective offers, as we shall see, a flexible means of accommodating such multiple modes and the context-dependent use of fast, heuristic strategies within a single overarching processing regime." (CLARK, 2016, p.245).

However, when truly getting into details of how PP can account for dual processes, Clark (2016) abandons Evans, Stanovich and Kahneman and instead works with a similar dual division of model-free and model-based processing proposed by Daw *et al*. (2005, 2011). As we have seen in chapter two, the details for defining dual process theories makes all the difference, and such difference is very much overlooked in Clark (2016). Predictive processing can account for lower and higher forms processing, but DPT is not this simplistic.

Daw *et al*. (2005, 2011) are computational neuroscientists working with reinforcement learning. They propose a dual division where model-free reinforcement learning follows Thorndike's (1911) classical law of effect principle which states that responses that produce a successful effect will more likely be repeated in similar situations. In contrast, model-based reinforcement follows Tolman's (1948) principle of latent learning, where learning does not occur with an immediate response and no obvious reinforcement can be deduced, rather, organisms seems to be using learning for inner models such as cognitive maps.

A problem here is that this dual division works best under Clark's (1997) older proposal where fast responses occurred because of a reactive framework (see Brooks, 1999), simple behavior components fine-tuned with the environment, whereas slower and careful thinking required higher level models. However, this has clearly changed in Clark (2016) where it is extensively argued that it is precisely rich generative models that enable fast and fluid action by

predicting outcomes. Clark (2016) would have to agree that expectations (in the form of rich generative models) are there precisely to help organisms act quickly in the world, they provide the grip necessary for frugal responses. In contrast, Type 2 processes actually need to be expectation-free (generative model-free), that is, they must compute their answers online based on symbol based reasoning. Recall that error is associated with attention and attention is associated with working memory and working memory with Type 2 processes. Thus it is when error is not resolved by expectations that working memory and Type 2 processes are called into play, when the generative model has failed.

Clark (2016) and Daw *et al*. (2005, 2011) should also agree that a 'model-free' principle has very little sense with the account they have been advancing, as there is no such thing as free sensory information but rather prediction error relative to some expectation. Interestingly, by treating Type 1 processes as model-free and Type 2 processes as model-based. Daw *et al*. (2005,2011) make the same Fodorian mistake of attributing knowledge-rich forms of processing to higher processing only, precisely what Clark (2016) is arguing against in the rest of his book. Further, for any of this to hold we would need to concede that their dual distinction could be transferred from learning to reasoning in the first place.

Clark (2016) then attempts to directly tackle phenomena which seems to be understandable in a dual division. He argues that a good model-based example is 'intuitive physics':

> "More complex (intuitively more 'model-rich', though this is now just another location along a continuum) strategies may also involve simplifications and approximations. A nice example is work [...] on 'intuitive physics'. Human agents are able to make rapid inferences about the physical behaviour of ordinary objects. Such inferences might include spotting that the pile of books or washing-up is unstable and at risk of toppling over, or that a lightly brushed object is going to fall and hit some other object." (CLARK, 2016, p.258).

Now this is the problem of having a simplistic reading of DPT, a problem dual process theorists have been to subject again and again. It becomes very hard identify and differentiate instances of Type 1 and 2 processing if you do not have strict defining features to follow. Why would intuitive physics be a Type 2 phenomenon if, by Clark's (2016) description, it does not seem to load heavily on working memory, is quick and effortless?

Clark (2016, p.258) goes on to explain that:

> "Underlying that capacity [...] may be a probabilistic scene simulator (a probabilistic generative model) able to deliver rapid verdicts on the basis of partial, noisy information. Such a simulator does not rely upon propositional rules but rather upon 'quantitative aspects and uncertainties of object's geometry, motions, and force dynamics'

Certainly, 'quantitative aspects and uncertainties of object's geometry, motion and force dynamics' are not the sort of content of working memory. Those types of properties are what not what we are used to grasping and holding in thought. Working memory is very much related to our awareness, where we seem to know what items we are rehearsing in memory, such as telephone numbers. Whereas the content of intuitive physics is not explicit in such manner. Further, this sort of processing seems to be quick and effortless. Do we stop to ponder about the quantitative aspects and uncertainties of objects geometry, motion and force dynamics in order to act? No.

We agree this seems to be a very interesting account of intuitive physics. The point here is that, by all agreed upon definitions summed up in chapter two by following the most respectful dual process theorists of reasoning, this is actually a Type 1 phenomenon. In fact, it is a 'model-based' Type 1 phenomenon, just to make it clear that Daw's *et al*. (2005, 2011) account does not transfer easily to Basic DPT. In our reading, it is actually extremely plausible that such a model-based phenomenon is Type 1, since it is precisely such knowledge rich generative models which allow organisms to think fluidly, fast, and effortlessly about the world.

3.3.2 Features not well explained in predictive processing

It is not only that predictive processing accounts of DPT do not work, but also that most Type 2 features are at odds with HBPP. Predictive processing cannot explain why there is **limited capacity** in general, which is one of the main discoveries in cognitive science. Limited capacity has been found in short-term memory (Miller, 1956), attention (Simon and Chabris, 1999) and competing tasks with divided attention (Schiffrin and Schneider, 1977). While it is possible for us to unconsciously realize a great amount of tasks, some of them compete for limited capacity, there is a limit to our multitasking abilities related to what sort of tasks these are. Thus, although we see, feel, hear, think and move, choose and act at the same time, we cannot solve puzzles, pay attention to the structure of music, do math problems and talk about philosophy all at the same time. That is because these latter tasks require much error debugging which need limited capacity resources. But

this part of needing limited capacity resources for multitasking is not included in the predictive processing architecture.

In the case of attention, it is true that if you have a well explained mechanism then limits follows natural from it. If precision can determine the trustworthiness of the sensory information, it seems natural that the highly weighted prediction error will account for attention (although see Ransom *et al.* (2016) for arguments that some sorts of voluntary attention are not well addressed in HBPP). There are thus two funnels in the bottom-up flow, the first is that not all sensory information but only prediction error continues to flow up the hierarchy and second that out of these errors only the most highly weighted will received the more specific care that we can call attention. However, there are a multitude of errors in relation to predictions, various of which should be highly weighted and it is not specified why these resources should be limited. More resources will be shared to highly weighted errors, but why are these resources so limited and why do they compete in concurrent tasks. If multimodal effects occur automatically and soft modularity is true than resources are being shared without attention frequently. If more resources are shared why would there be decrease in general capacity rather than an increase?

Also, it is hard to see how short term memory capacity fits with predictive processing. Why would there be a limit of nearly seven items? Where would this fit in a hierarchical predictive processing architecture? It is hard to say. We will propose in the next chapter that much of the content that comes to attention will be further processed by a classical architecture of Newell (1980) which is limited in order to follow realistic computability constraints as discovered in mathematics and artificial intelligence. Limited resources is a price paid for simulating an organized, coherent, sequence of thoughts that apply (or tries to apply) normative rules as proposed by the classical architectures.

Predictive processing does not necessarily shun **working memory**, but just to illustrate how important this concept is to such framework it is interesting to see how it is mentioned only once in Andy Clark's book (2016) and absent from Jakob Hohwy`s (2013) book and other work in predictive processing. Working memory is mentioned 119 times in Frankish and Evans' (2009) review of DPT, we say this just as a point of comparison. In other words, it is probably not a very central tenet of predictive processing. And there is every reason for working memory not to be a

relevant tenet of predictive processing. This is precisely because stronger load on working memory concerns cases where the information that needs processing is unpredictable, or is not well accommodated by any statistical judgment, in fact, if the general prediction by statistics schema fails deeply to account for some relevant data, then it seems plausible that another type of processing should be applied. When predictions are working, then, really, working memory is dispensable.

Moving on to **decoupled** processes, it is crucial to remember the comments we made on decoupling in chapter two. There we have that decoupled representations cannot be any type of considerations of possibilities but only those that are free from given environmental context. Now, in this chapter, we can start to disentangle counterfactual ways of thinking about the world, specifying details and showing precisely what predictive processing accounts for and what it does not. We understand Type 1 predictions can engage in counterfactuals and simulations. However, these counterfactuals and simulations can only be probable causes given a certain context. As Kahneman (2002) has argued, "Close counterfactual alternatives to what happened are perceived—one can see a horse that was catching up at the finish as almost winning the race."[77] So when Osman (2013) speaks of the theater where one engages in counterfactuals and simulations of possible occurrences these are based on the current state of the play and expectations of probable outcome. So stretching the example, were we watching a horror movie, based on some given hints, we could wonder: 'this indicates this character is the killer'. Because based on such hints and previous movies usually these leads indicated so and so. Such probable outcomes of present scenes are Type 1 predictions. However, were we to wonder deeper, to consider the plans of a tricky movie director, then we would need decoupled representations. For instance: 'maybe he is hinting that this character is the killer to surprise us by making another hidden one be the killer. What hidden character could be the one that turns to be the killer after all?". When engaging in such counterfactuals and simulations which are not directly inferred from the given hints (or are improbable) then we will need working memory and we probably will lose concentration from the film because we were wondering too much. Thus, decoupling for a Type 2 process must mean

77 Despite Kahneman's language, note that we do not need to get into the discussion of if counterfactuals are actually perceived or just effortlessly inferred.

simulating and using counterfactuals that do not follow direct consequences of the probable outcomes of the given states of the environment.

Seth (2014) proposes a predictive processing account of counterfactual processing which is very enlightening for our cause. Seth is worried about having ways of distinguishing phenomena with perceptual presence, such as an apple on the table, from those without it, such as an afterimage. Interestingly, his suggestion is that for an object to be understood as present, cognition must employ rich models of counterfactual states of such object. This is because a present object could be moved in various ways, we could touch it, it could fall, that is, it interacts causally with other objects in the world and humans predict such probable interactions, whereas an afterimage does not interact with anything.

> "I propose that normal (veridical) perception is underpinned by counterfactually-rich generative models, which means that these models encode not only the likely causes of current sensory inputs, but also the likely causes of those sensory inputs predicted to occur given a large repertoire of possible (but not necessarily executed) actions—hence the term counterfactual". (Seth, 2014, p.98)

Despite the stronger claim that any perceived object must be embedded in such rich counterfactual schema, what is important for us to notice is that by having a predictive processing schema it pretty much follows that what is present in the environment will be used to predict occurrences to come. So what will be inferred via Type 1 processes are not only actual states of the world but likely consequences as well, such as a horse winning a race or a likely plot outcome in a movie.

We can call the sort of processing we have been discussing 'situated counterfactuals', counterfactuals of likely states that will be given directly by our perceiving (or effortlessly inferring) of the world. Considering Seth's (2014) claim we have that our cognition over the given environment is composed not only of actual states put possible outcomes as well. Thus, to decouple from this, one needs a form of representation which can work ignoring both the actual state and also plausible possible occurrences. So in the horse race example, one could wonder 'what would happen if someone were to interfere with an obstacle in the race'. In such case, this would decouple from the actual states and the likely outcomes, that sort of process can reasonably be understood as needing working memory, effort and limited capacity.

So a decoupled representation in Stanovich and Toplak's (2012) sense must be one that is decoupled from the actual state and the counterfactual likely states of the world, because the likely counterfactual occurrences are at least somewhat tied to veridical perception. Without this clarifying distinction, not only would our approach in this chapter be problematic but decoupling as a feature of Basic DPT would most likely be false. Since, there could be countless examples of effortless, fast, working memory-free counterfactuals. In this sense, our approach is actually needed to save the decoupling feature. This distinction however is not some *ad hoc,* made up solution. It very much in terms with how phenomenologists differentiate anticipations from productive imaginings such as picturing in the mind (see Noë, 2004).

Now consider how predictive processing could deal with **decoupling** as we have described it. How could processing based on probabilistic expectations somehow work on improbable, unlikely, context-free information? Of course, there are very interesting predictive processing accounts of imagination and simulation (Grush, 2004; Clark, 2012, 2016). In these accounts it is understood that imagination is possible since generative models accumulate knowledge on perceptual content and when needed can simulate and work out counterfactuals. By raising the precision on the models and attenuation on error, simulations can be executed without interference from the world. At first, this could be seen a mechanism for producing decoupled representations. We could suppose that generative models can be manipulated so that non probable simulations are possible. But even granting this is possible, in what sense will reasoning based on these simulations follow the predictive processing schema? There are no determining causes from priors, no predicting based on occurrences to be made, no Bayesian inferences based on probabilities available. In what sense is this subsequent computation, to reason carefully, serially over counterfactuals and simulations, following predictive processing? Decoupled representations are needed precisely to free our reasoning from interferences with what is most probable, what is repeated, what we already know. Decoupling is used to form new takes on things, to view things differently, to try other strategies. In order to process such type of counterfactual content, logical rules can be used (or approximations) as well as other attempts of reaching normative rationality. This allows for deductions of outcomes that we could not have foreseen by keeping track of the world based on Bayesian inference.

Also, predictive processing cannot explain why there should be a difference between **implicit and explicit representations**. Why is it that some content cannot be grasped, put in words, worked in clarity and detail while others can? As we have argued, it is plausible that content represented in probability density distributions will not be perfectly grasped. However, why should there be content that we can easily 'manipulate', speak of, organize and use them vocally or in text? As we will argue in the next chapter, this is a property enabled by symbolic representations. Thus, there are no representational differences in predictive processing that explain the explicit and implicit distinctions in a format that repeatedly appears in psychology.

Further, every distributed processing framework which shuns **serial processes** as a whole will have difficulties to account for how people can end up with organized goals, thoughts and a coherent phenomenology in general. We all agree there are a multitude of systems in the brain which process information and have their own separate results. But even models which emphasize this multiple drafts view (see Dennett, 1991) need to postulate some mechanism by which these multiple systems can reach effective coherence from time to time. This is what inspired the now famous global workspace model of consciousness (Baars, 1988). We find it unlikely that we could simply do without these higher level global organizing systems. In the case of Dennett (1991), even when arguing for how cognition mostly produces drafts, he has to postulate a 'joycean machine' which serially organizes these drafts into 'stories'. We believe something of this sort is needed to account for Type 2 processes. This is also what accounts for **control** and why it is qualitatively different from **automaticity**. Where in predictive processing there is no explanation for why there should be controlled processes.

Finally, predictive processing does not explain why there is a division in reasoning tasks between **intuitive responses and normative responses**. Dual process theories were developed precisely to account for such evidence. So either predictive processing needs to account for why these two general profiles of responses obtain or it needs to explain how DPT (which does explain the evidence) fits in its framework. As we have seen, a successful predictive processing account of DPT has not been proposed. We believe such complete account could not be proposed for the reasons we have been speaking of and the ones we will sketch out in this next section.

3.3.3 Discussion on predictive processing's limits

We have seen in the previous sections that Clark (2016) does attempt to account for the DPT distinction but fails, and that some features used in Type 2 processing are not well accounted by predictive processing. Now we move on to show that there are theoretical reasons that explain why predictive processing does not account for Type 2 features. Let us start this section off by showing how Clark (2016) seems to admit from start that similar descriptions to Type 2 processing are not at the heart of the predictive processing story:

> "Prediction is, of course, a slippery beast. It appears, even within these pages, in many subtly (and not-so-subtly) different forms. Prediction, in its most familiar incarnation, is something that a person engages in, with a view to anticipating the shape of future events. Such predictions are informed, conscious guesses, usually made well in advance, generated by forward-looking agents in the service of their plans and projects. But that kind of prediction, that kind of conscious guessing, is not the kind that lies at the heart of the story I shall present (Clark, 2016, p.2)."

Although he admits there are these two senses of prediction, he later on does not explain why, in his model, there should be two types of predictions and how the second one comes about. Instead he argues for a grand unified theory of brain function (Clark, 2013a, 2016). Anderson and Chemero (2013) raise a similar worry. They argue Clark (2013a) is conflating two types of senses of prediction, prediction1 and prediction2.

> "The first sense of "prediction" (henceforth prediction1) is closely allied with the notion of correlation, as when we commonly say that the value of one variable "predicts" another. […] The second sense of "prediction" (prediction 2), in contrast, is allied instead with abductive inference and hypothesis testing. we don't think that evidence for predictive1 coding warrants a belief in predictive2 coding.And it is only from predictive 2 coding that many of Clark's larger implications follow." (Anderson and Chemero, 2013a, p. 24).

Anderson and Chemero (2013) ascribe properties to these two types of predictions of which we are not sure we agree with[78]. The point is that despite Clark's goal of a having a GUT, the dual process problem keeps showing up and he has to deal with it. As we understand it, prediction1 is

78 Such as abduction being entirely related to prediction 2.

well explained by predictive processing, but prediction2 goes further by requiring computations on propositional content which organize content in order to make higher predictions.

This dual process problem also shows up in the concept of surprisal, since surprisal occurs frequently by flagging errors in generative models. However surprise (in contrast to surpris-al) is relatively rare, and involves a sense in which the 'agent' is surprised and not generative models. Clark (2013a, p. 16) notices this problem and attempts to dismiss it as well, he writes:

> "[…] there seems to be a large disconnect between surprisal (the implausibility of some sensory state given a model of the world–see sect. 1.6) and agent-level surprise. This is evident from the simple fact that the percept that, overall, best minimizes surprisal (hence minimizes prediction errors) "for" the brain may well be, for me the agent, some highly surprising and unexpected state of affairs– imagine, for example, the sudden unveiling of a large and doleful elephant elegantly smuggled onto the stage by a professional magician."

Clark (2013a, p.16) then goes on to explain this and argues that "top-level theories of an initially agent-unexpected kind can still win out so as to explain away that highly-weighted tide of incoming sensory evidence." The problem, however, is what accounts for the reason why there should be an 'agent-unexpected kind' in the first place. Of course, as we see this, Type 2 reasoning processes seem to be clearly more available to introspection than Type 1 processes and hence a disconnect between surprise and surprisal would obtain.

This dual process problem shows up also nullifying predictive processing when Clark needs to speak of Type 2 characteristics. As we have seen, probability density distributions are responsible for much of what gives predictive processing its explanatory success. Representing information in such fashion allows for statistical processing of previous input and for generative guesses for future outcomes. There is a problem with this representation, however, which is keeping a probabilistic take on states of objects. Having this probabilistic state usually allows organisms to act more rapidly, but there are times when we need precise, definite, properly discrete information about an object. In such times only one answer is needed and related ones should not interfere. To account for this, Clark (2016) speaks of single peak probability distribution functions, representations where each distribution must have a single best explanation. Thus, instead of having various related peaks indicating possible outcomes, only one is enforced.

> "One fundamental reason that our brains appear only to entertain unimodal (single peak) posterior beliefs may thus be that— at the end of the day— these beliefs are in the

> game of informing action and behaviour, and we can only do one thing at one time." (Clark, 2016, p.188).

Now, what happens when you have a single peak probability density function is that it acts like a discrete symbolic representation. That is, all other possible states are denied in favor of a single active state. That means all of the predictive processing story about statistical encoding and generative models predicting probabilistically stops and some other form of computing needs to take place. When using single-peak probability density functions you lose the effects of having various related instances as possible outcomes to gain feasibility, you lose effective predictive processing.

Predictive processing stories present really detailed specifications on how it is that the brain is able to stay ahead of the world and predict consequences of current states. You have the Bayesian hypothesis, probabilistic representations, surprisal, prediction error, precision mechanism and so forth. However, surprisingly, not much is said about what goes on in cognition when predictions fails deeply and prior statistics and likelihoods are not enough to deal with the uncertainties that the world leads us to. How does a system stumble onto a better hypothesis? Of course, you could go on and say more about how such framework would account for the processing of significant error, however, it is not a part of the theory that they have dedicated as much time and effort as they have dedicated to the actual predicting schema. If statistics fail, are we to accept that the brain will go on and apply more Bayesian statistics to reach another hypothesis?

The ability to flexibly reformulate priors to interfere with brain normal responses is not very old in evolutionary terms. Humans and some other mammals seem to have developed this skill more than other beings. Such ability seems to go in hand with cultural development and developed manifestations of communication. Cultural development and advanced communication is allowed by the use of symbols. It is probable that working memory is a mechanism that uses symbols as means of holding concrete, discrete objects over and above what is a fleeting world of distributed probabilities.

Hirsh *et al*. (2013) argue for a similar position. Drawing from the field of narrative psychology (see Bruner, 1991) they argue that narrative representations could function as high-

level generative models for organizing content. This would allow for a function of integration, as they explain:

> "Although narratives can take many different forms, they are distinguished by their ability to compress and encode a great deal of information about the world, including the causal relations between events over time […] the planning and sequencing of goal-directed actions [...] the emotional significance of an event within a temporal context [...], the unfolding nature of personal identity [...], and the dynamic intentions of multiple social agents [...]" (Hirsh *et al.*, 2013, p.36).

Hirsh *et al*. (2013), like us, believe predictive processing can serve as an account for a great deal of phenomena but would work best if linked to a high-level serial, propositional organizing structure (see also the joycean machine in Dennett, 1991). Clark (2013a) faced with this position is dubious. If he needs to defend predictive processing as a unified theory then he will argue that these narratives can be scattered across different levels of the hierarchy in predictive processing. So Clark (2013b, p.61) defends predictive processing by responding that

> "Narrative structures, if they are correct, lie towards the very top of the predictive hierarchy, and they influence and can help coordinate processing at every level beneath. It is not obvious to me, however, that personal narrative needs to be the concern of a clearly demarcated higher level. Instead, a narrative may be defined across many levels of the processing hierarchy, and supported (in a graded rather than all-or-none fashion) by multiple interacting bodies of self-expectation". (Clark 2013b, p.61).

However, when he needs to account for Type 2-like features alone, he draws from Dennett's division of two systems. A while after arguing that an unmodified predictive processing schema accounts for Type 2 processes, Clark (2016) goes on to tackle the problems of human cultural advances. Now however he concedes to some sort of 'artificial second system':

> Self-produced (or mentally rehearsed) language would then emerge as a potent means of exploring and exploiting the full potential of our own acquired generative model, providing a kind of artificial second system for manipulating the precision-weighting of our own prediction errors— hence a 'neat trick' (Dennett, 1991) for artificially manipulating our estimations of our own uncertainty enabling us to make fully flexible use of what we know (Clark, 2016, p.284).

This 'second system' would be the application of language to tune precision weighting. But does this application of language, this new artificial system, compute following HBPP applying Bayeasin inferences to probability density distributions? There is no reason to believe so. It is a new mechanism that regulates precision with another sort of computation. Clark does not specify how this manipulation of precision by language could work. In our terms, this sort of work would be done by a classical architectures like Newell's (1980). Using something like the language of

thought (Fodor, 1975), this second mechanism formulates new options for generative models to use. Thus, in training a skill to automaticity, what occurs is that, in untrained people, generative models cannot figure out how to deal with the situation fluidly, thus processes using serial, limited, propositional, and thus Type 2 commands (such as 'use the breaks in curves' for driving) are necessary to reformulate hypothesis, enabling the application of new strategies. This training ends when the feedback from Type 2 processes are no longer necessary and generative models can account for the situation alone, automatically. What matters for the present purpose, is that something like this needs to be true, and it is not clear how this artificial second system could aid if not by some different sort of computing.

Likewise, Aaron Sloman (2013) in his criticism asks. "What else can brains do?" His point is that it seems plausible that brains compute in different fashions for different tasks. Although, we argue here for dual computing division, we do not want to say that the whole mind can be divided in these two forms of computation. We are working in a framework that we hope can serve at least for dual process theories of reasoning and decision making. The conclusion Sloman (2013, p.51) reaches is also that when you need to explain various different functions, the general schema of predicting the world is not related. He writes: "Many things brains and minds do, including constructing interpretations and extending their own meta-cognitive mechanisms, are not concerned merely with predicting and controlling sensory and motor signals." In sum, predictive processing is always timid in developing material over truly Type 2 processing occurrences.

3.4 Chapter conclusions

We can now have a general idea of HBPP, its applications and its limits. Hopefully, this will have convinced the reader that predictive processing seems to be more aligned with Type 1 features and that it fails when explaining Type 2 features. If this is the case, then we might have begun solving the unity problem. To end this solution, we now need to see how classical architectures explain Type 2 features but fail at Type 1 processing.

CHAPTER FOUR:
REFLECTING AS EXPRESSION SEARCH

4 Reflecting as expression search

In the last chapter we saw how predictive processing seems aligned with Type 1 features but not with Type 2 features. In this chapter, we want to show how classical architectures in cognitive science should not be abandoned completely, as they really are the best model for Type 2 processing. We will also hold a stronger explanatory hypothesis which states that while the brain does not have a classical computational structure, working memory is part of an emergent system that does seem to have such a classical architecture. Thus, Type 2 processing will obtain from strong use of a classical architecture and therefore will load heavily in its working memory.

We will start by reviewing some fundamental ideas of classical cognitive science (section 4.1), then we will show how such fundamental idea accounts for Type 2 features (4.2) by working out each one as we have done in previous chapters. Finally, we will show how classical architectures fail to account for Type 1 features (4.3), mostly because of the frame problem of cognitive science.

4.1 Computing and Classical Architectures

Concepts of classical architectures were developed in close relation to basic notions of computing. Now, while it is true that any model in cognitive science must bear some relation to information processing and computing, it is not true that such relation must be as described by classical architectures. Thus, there is an idea of how the mind works that is specific to classical cognitive science alone and not to other models, such as connectionism (McClelland, 2013), reactivism (Brooks, 1999) and Bayeasinism (Knill and Pouget, 2004).

One might argue that since computations can be understood as the workings of a Turing machine, and schemas like predictive processing (for instance) are computational, then such schemas are in essence like a Turing machine. If this were the whole story, then the sort of computational difference we search for could not be a useful solution to the unity problem.

It is true that different sorts of computation can be reduced to the workings of a Turing machine. However, it does not follow, that therefore they have the same implications for cognition.

Because, although they are reducible in theory, in practice, such reduction does not equate to the same performance, important differences obtain because of physical and time constraints. As Newell (1980, p.155) explains: "Every universal machine exhibits in some form all the properties of any universal machine. To be sure, differences exist among universal machines-in primitive structure, in processing times, in sensitivities, and in processing reliabilities."

We have something similar in mind. There are various architectures for computing, and while they can all be reduced to computations of universal machines, there are differences in processing times, sensitivities and reliabilities. And for models of cognitive science, what matters is precisely differences in processing times, sensitivities and reliabilities[79].

Classical architectures were more deeply related to Turing machines than a predictive processing architecture is. Computing structure of classical architectures were similar to Turing machines in that they were something like implementations with adaptations, or a specific use of the Turing machine idea.

A Turing machine (1936), in general, consists of a tape divided into squares where a symbol can be written, a reader that considers only one square at a time and a set of rules that work as a table of instructions for what the machine does. The reader examines the symbol on the tape and proceeds as the table of instructions determine. The machine, at any given moment, can replace a symbol on the tape, move to the left or right or stop. This machine is a mathematical abstraction, so although there are finite rules and finite symbols, its tape is infinite. However, a form of such abstract machine has long been instantiated in real, physical computational architectures. The common architectural form for computers is called the Von Neumann architecture.

The Von Neumann architecture (Von Neumann, 1945) is a design for implementing a digital computer. It consists of a processing unit, a control unit, primary storage and input/output mechanisms. The processing unit contains an arithmetic/logic unit which executes instructions by carrying out arithmetic or logical procedures on a binary code and a register that quickly stores steps of such operations. This is analogous to what reads and applies changes to the tape in the

[79] Newell (1980, p.178) called this worry "The Turing Tar fir" and rejects it on the grounds that "the question of interest here is precisely what structure provides flexibility".

.

Turing machine. A control unit contains an instruction register, which holds the instructions to be processed at a given time and mediates the flow of processing by providing timing and control signals. This is what makes the idea of a table of instructions physically possible. Such instructions as a whole are stored in the primary storage along with the data the processing unit previously operated on. So it is a memory that is analogous to the tape in the Turing machine but which also stores the instructions of the table as a whole. Input and output mechanisms are what enable the communication between this central core of computers with all its other parts, such as its external mass storage hard drive.

We can see that the implementation of a digital computer bears a close architectural resemblance to a Turing machine. This would not be true for the cases of a quantum computer, a parallel computer or a brain. But most importantly for our purposes, is that Newell's (1980) architecture of the mind also bears resemblance to the Von Neumann machine and to Turing machines. And not only in essence and structure but also in 'processing times, sensitivities and reliabilities'.

Newell's (1980)[80] architecture of the mind is called a physical symbol system and can be seen as a fundamental idea behind classical cognitive science. Like the Turing machine and the Von Neumann architecture, a symbol system is also a machine that reads, operates on and writes symbols that form expressions, and whose expressions can be coupled with distinct actions. A symbol system consists of a set of operators, a control unit, a memory and input/output functions. The memory stores symbols which together form expressions. The control unit is what performs 'interpretation' (more on this on the next section). Input-output functions are what guarantee communications with other parts of the system and with the environment. There are ten operators which take in symbols as input and produce symbols as outputs. For illustration, some examples are the 'read' operator, the 'write' operator, the 'copy' operator and the 'behave' operator. By manipulating expressions and by the coupling of such expressions with actions, this sort of symbol system is supposed to be responsible for any sort of cognition.

[80] Newell (1980) warns that such description of classical cognitive science might not be accepted as he gives it. The case is that no such universally agreed version will exist and that Newell's, Simon's and Fodor's version is usually a good reference to a generally agreed upon version of classical architectures.

The point for now is only to start showing how it is the case that, for classical architectures, it is not only the case that their computational commitments are reducible to Turing machines or digital computers, but that there are intrinsically related even in the way it functions, without a need for 'reduction'. Let us now go a bit further in the concepts

4.1.1 Basic concepts of symbols systems and heuristic search

We are very much interested in the notion of symbols insofar as it helps us differentiate classical architectures from predictive processing. Newell and Simon's (1976) review of their work on artificial intelligence sheds some light on the idea of symbols and related definitions. Newell and Simon (1976, p.116) state that "A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure)." Operations then manipulate such symbols, creating, copying, editing and destroying expressions.

Unlike the Turing machine, for a physical symbol system it is essential that it exists in a world of objects. That is because one of the fundamental roles of symbols and expressions is designating[81] objects. Newell and Simon (1976, p.116) explain that "an expression designates an object if, given the expression, the system can either affect the object itself or behave in ways dependent on the object" and thus gain "access" to the object via expression. As Newell (1980, p.156) puts it, "having X (the symbol) is tantamount to having Y (the thing designated) for the purposes of process P". Any object could be so designated via symbols and expressions. Thus, Newell (1980) claims that the numbers in a Turing machine are not symbols as in the case of physical symbol systems. That is because while digits in a Turing machine do work as tokens of expressions, they are not however coupled with some data structure.

Having symbols that discretely designate grants the system a few properties, such as transitivity. If X is a symbol that designates Y and Y designates Z, then X designates Z. Therefore, symbols are granted transitivity. Now, we only obtain this if X is granted a fixed and discrete value,

[81] Such notion is very much related to what philosophers call reference.

thus, there can be no uncertainty over the value of a symbol, as one would have over a variable. Another way to put this second commitment is that "at any time a symbol designates a single entity".

Designation also allows for universality. Universality is the features of physical symbol system by which they are free from content-wise limitations and can simulate any other universal machine, such as a Turing machine. This is because symbols can mimic the parts of another universal machine and then act as the other machine does. For instance, a programmer can simulate a simple Turing machine (of course with fixed digits and a fixed amount of tape) in a digital computer by creating tokens that refer to the tokens of what such simples Turing machine would have.

If an expression designates a process of the system and if the system can carry out such process, it is said that the system can interpret the expression. Another way to understand this is that expressions have some sort of meaning to the machine when actions (or processes) can be derived from them. Such a process is enabled by the control (a part of the symbol system). What the control does is to bring together, in a certain moment, data and operator. In Newell's (1980, p.159) own words "This organization implies a requirement for working memory in the control to hold the symbols for the operator and data as they are selected and brought together." Even more interesting for us is that "[...] working memory is an invariant feature of symbol systems". Interpreting expressions and thus applying operators is what makes for the production of new expressions. Designation and interpretation therefore are two crucial functions of a physical symbol system.

Being such a system alone by no means implies having granted problem solving abilities. So Newell and Simon (1976, p.120) need to explain how such systems exhibit problem solving abilities. They do so with the heuristic search hypothesis "The solutions to problems are represented as symbol structures. A physical symbol system exercises its intelligence in problem solving by search—that is, by generating and progressively modifying symbol structures until it produces a solution structure."

As Newell and Simon (1976) explain[82], for these systems problems come in a specific form. For there to have a problem, there must have a problem space, or a state-space. Such space includes all the possible steps the process could take when solving the problem, including the initial state, various solutions and steps in between. There must also have a test which states what state the machine must be in order to have reached a goal solution, or what kind of answer one would want of such problem. And finally there must be move generators which are processes for modifying one situation in the problem space into another.

Newell and Simon (1976, p.121) also state the conditions in which a machine would be solving problems intelligently:

> "The task that a symbol system is faced with, then, when it is presented with a problem and a problem space, is to use its limited processing resources to gen-erate possible solutions, one after another, until it finds one that satisfies the problem-defining test. If the system had some control over the order in which potential solutions were generated, then it would be desirable to arrange this order of generation so that actual solutions would have a high likelihood of appearing early. A symbol system would exhibit intelligence to the extent that it succeeded in doing this. Intelligence for a system with limited processing resources consists in making wise choices of what to do next."

So in the case of physical symbol systems, a problem space contains all the forms the expression could be in given the initial form, the move generators it has available and the test. Thus, each expression obtained in the process is a modification of the initial one. Further, the move generators incorporate rules, such as logical, arithmetical or whichever is desired. These rules contain both general problem solving rules and specific ones. For instance, a general one would be something like: 'do not produce contradictions'. A specific one would be related to the problem at hand, such as 'create the smallest expression possible'. At any given step, the modifications are not chaotic, but rather always depend on these rules incorporated by the move generators and on the current state of the process, a given place in the problem space.

The steps taken in a problem space form branches of a tree. A search tree, after a solution has been found, then exposes each step taken from the initial expression to the solution expression. If every state (or most states) of a problem space were visited (brute-force search), the process would incur in computational explosion, unless the problem space is too small (which does not

[82] See also Luger (2009).

occur for relevant problems) in the world. Thus, the search must be heuristic in the sense that it must take the fewest steps possible to achieve solutions. As Newell and Simon (1976, p.123-124) claim: "The task of intelligence, then, is to avert the ever-present threat of the exponential explosion of search" and "Our analysis of intelligence equated it with ability to extract and use information about the structure of the problem space, so as to enable a problem solution to be generated as quickly and directly as possible[83].

In tree search, it is thus desirable to decrease the rate of branching and when in each node, the subsequent direction should be carefully planned. Heuristic strategies, such as measuring the difference of the next node from the goal are one way of achieving this. As Luger (2009) explains, there are different mechanisms for achieving this, such as searching steps in the same level of a tree first, or going through the same line of descendance first (Depth-First or Breadth-First Search).

The details of this does not matter for our present purpose. Indeed, this sort of search in problem spaces done by physical symbol systems could even be applied to various domains, say searching books, a chain of genes, moves in chess and so on. For our interests, of course, we care if they can search solutions to problems in reasoning. It turns out Newell and Simon (1963) have done just such work, not only in building programs that solve logical tasks but in comparing their solutions with human subjects experimentally.

Newell and Simon (1963) developed The General Problem Solver (GPS[84]) as a program using the idea of a physical symbol system and heuristic search to mimic human thought. Newell and Simon (1963, p.114) explain the GPS can designate objects, initially of any sort, that then will be transformed by generators. The program detects differences in objects and organizes tasks into three types of goals: "Transform object A into object B", "Reduce difference D between object A and object B" and "Apply operator[85] Q to object A". Generators can vary, but for the task studied, they were rules of symbolic logic, such as A v B -> B v A and objects were the expressions of

[83] of course every time they speak of intelligence or of discovering the structures of the mind as a whole, for us they are speaking of Type 2 reasoning only.

[84] The GPS is an old program and has developed into various other programs, such as the Soar program. We can imagine it has flaws other newer ones do not. However, despite of what technical difficulties it may have faced, our point here, will be that in its original form it captured the essential idea of Type 2 processing.

[85] They call these 'operators', but I will call them generators because they do not have a similar sense to Newell's (1980) operators, and they do have a similar sense to Newell and Simon's move generators.

logic. Thus, the GPS transforms an object A into an object B by checking for differences and applying generators. In the case of solving a symbolic logic task, the GPS represents the initial expression and the goal expression, it then attempts to discover which generators must be applied in order to transform the initial expression into the goal expression.

They then test the potential of this program in a experimental setup in which a student of engineering attempts to solve a symbolic logic problem in front of a chalkboard as he speaks openly about the steps he is taking in doing so. They then compare the same problem with the activity of the GPS and find most steps are surprisingly similar. For instance, the student speaks of a step he is taking: "Well, looking at the left-hand side of the equation, first we want to eliminate one of the sides by using rule 8." (Newell and Simon, 1963, p.119) Then, they compare what is said with the processes the GPS takes: “We see here a desire to decrease LI or eliminate something from it, and the selection of rule 8 as the means to do this” (Newell and Simon, 1963, p.119).

A few things do not match. However, it is interesting to notice that if we were to compare two people with each other, the steps taken would not be exactly the same. They would only be alike in many ways. So really there is little difference in comparing the steps taken by the GPS, with the reasoning steps taken by students solving the same task[86]. This would not be the case, for instance, if a neural network were to solve the same logical problem. There would not even be steps to compare. Of course the neural network could mimic such type of performance, but that would be in an output level, simulating a similar serial machine, not internally. If we were to describe the inner details of a neural network, they would be about increasing or lowering activations of nodes.

One could ask, using some sort of modern Cartesian evil demon, how we could know that humans are using a system like GPS or if it is just a simulation, in the sense that underneath it would be just like a neural network mimicking a serial system. The answer is that it probably is a simulation, since the brain is a network after all. The point is a system like the GPS should need to be simulated in order for our reasoning skills to have such type of performance. That computations in a network are not of the same sort as computations of classical architectures was very well noticed by Fodor and Pylyshyn (1988). In fact, their point is that because of the properties of

[86] Of course, there is the difference of having to interpret the states of the GPS since they do not come in natural language replies.

computations in symbolic manipulation, the only way the human brain could achieve certain tasks is by having some sort of instantiation of a symbol manipulation system such as those described by Newell and Simon (1976).

4.1.2 Compositionality of symbols in expressions

Fodor and Pylyshyn (1988) noticed some essential differences between the processing of networks and that of classical architectures. Importantly, they noticed that some features are essentially available in classical architectures which seem to be lacking or at least trouble for networks. We must bear in mind that they were speaking of early connectionists models, which we do not address in this theses. However, we understand that many of their contrasts apply to predictive processing networks, or any computing in any network at all.

Fodor and Pylyshyn (1988) understand that semantics are differently mapped in networks and in classical architectures. Where in a classical architectures semantics are ascribed to expressions, in networks, semantics correspond to states of a network, particular configurations of nodes and how they activate each other. They argue such difference ends up influencing how computation preserves semantic relations in each architecture. To illustrate how predictive processing is like neural networks in this sense, consider this claim:

> "Prediction error driven supervision by the world can happen via representations of quickly changing states of the world and via representations of more slowly changing parameters of the world. The former is thought to occur in the brain's synaptic activity (the way neurons interact in the short term), and the latter in synaptic efficacy (the way connection strengths are set over the longer term) (Hohwy, 2013, p.49)."

Fodor and Pylyshyn (1988) explain that classical architectures have a combinatorial syntax and semantics. Particularly with a distinction between (1) atomic (singular symbols) or molecular (expressions) and (2) syntatic constituents which are also molecular or atomic (unitary or complex operations), so that semantic relations are a function the structure of 1 a 2. Furthermore, a molecular expression A&B literally contains the tokens A and B and the semantics of said expression depends on such atomic tokens, whereas when a network represents A&B there is no causal structural connection between the constituents within such expression. A&B in a network corresponds to a

certain pattern of activation or state the network is in. That is, there is not a pattern for A and a pattern for B that together cause a pattern for A&B, When such expressions are produced, the patterns of activation do not relate to the structure of the content it represents. Representations are distributed in networks and parts of representations are not causally connected with the parts of the network.

In a classical architecture, the combinatorial syntax and semantics of a logical language, a language of thought (LOT, Fodor, 1975) or the inner language of the computer are the medium of computation. This means that if you were to change the symbols of an expression it would completely alter the state of transition of a classical machine. In the case of a network, on the other hand, the representations are not themselves the medium of computation. Representations are mapped to the states of a network but they are not computed on, and they do not cause the state transitions of the machine directly. The medium of computation are the transactions between neurons and they alone are not representations. To change the course of a network you need to change activation rules or the state of some elements. Thus, were you to change what the network designates, it would go on to do the same physical transactions as before.[87] Because of this difference, in classical architectures, content is directly constrained by the causal roles of computing whereas in networks this is not so.

Not only is content directly constrained but also lexical items follow the principle of compositionality. Fodor and Pylyshyn (1988) explain the principle of compositionality as the fixed semantic role a lexical item plays in an expression[88]. This implies that in another expression the same lexical item plays the exact or a very similar semantic role.

[87] Fodor and Pylyshyn (1988) hold the view that the content of the representation did not interfere with how the state of transitions of the machine performs at all. Changing the representations, what the network designates, would not change how the machine behaves. This is only true in the sense that the machine would go on to the same activities, would go through the same state transitions. However, since its outputs would be meaningless in this new form, this changed version would be like an untrained network, which would then require more training in order to behave properly. Thus, the content matters in the sense that it is what determines if a network is trained or not. If we were talking about a brain, such a change would lead to the reformulation of the network, new training would ensue because error would accumulate. Anyway, Fodor and Pylyshyn (1988) are correct in stating that there is a difference in how the content relates with the computation in classical architectures and networks.

[88] Fodor and Pylyshyn (1988) argue intensively that systematicity is a main component of classical architecture and that compositionality is one of the factors which grant systematicity. However, systematicity is not of very much interest to us as compositionality is. Also compositionality is used not only in systematicity, but to show how two expressions are semantically connected.

We know that large part of our mental functioning must be context-dependent as humans are very sensitive to these changes. However, we can also think about things without taking context into play. So it is no shame to think that some mechanism must preserve context-independent meaning of lexical items. As Fodor and Pylyshyn (1988, p.30) claim "[...] to the extent that the semantic value of these parts is context-independent, that would explain why these systematically related thoughts are also semantically related." To be clear, we are not speaking of natural language, we have no serious hypothesis to offer on language processing. We are speaking about the role of symbolic expressions in computational mechanisms which we suppose are responsible for thinking.

Fodor and Pylyshyn (1988) take compositionality to be necessary for formal deduction. The example they offer is (1) Turtles are slower than rabbits, (2) Rabbits are slower than Ferraris. Therefore, turtles are slower than Ferraris. They argue that if the relation that holds between turtles and rabbits are not the same as that which hold between rabbits and Ferraris then it is hard to see why the inference should be valid. This example is especially interesting because the speed of an object in the physical world is usually something with a wide variety of variable positions. However, to work this out in logic, we must objectify this variable quantity into a symbol with a fixed value in order to proceed.

Therefore, in a sense, working logically already means being context-insensitive and compositionality dependent. Because if one starts to question the meanings of the lexical items deeply, they could be taken to mean something else ("how slow?" "slow in what kind of terrain?" "how much time is involved?"),

Fodor and Pylyshyn (1988) state that constituents must preserve lexical meaning across expressions. Another good way to put this is that lexical items must play the same role in various expressions in order for the full exploration of consequences of axioms to obtain. Suppose you have the equation X + (3 - 2) = 4*7. You would need a few steps of computation in order to reach the conclusion that X is 27. But you can only reach such conclusion if, in each step, all lexical items, numbers in this case, remain the same and are kept in memory. If any of the numbers were to change or were you to forget details of how the expression stands after a certain step, you would lose all your progress. It is the same for computational expressions. Some lexical items must have

core meanings in order for the exploration of problem spaces to obtain. If this is changed, in another representation, the system might gain some abilities, but it will certainly lose the exploration of consequences of axioms. Of course we could switch the lexical items of the equation with a letter that stands for a probability density functions instead of numbers and still solve an equation likewise. But notice that, in such case, we have a fixed representation of the function, say P. This is a constituent lexical item which can figure in expressions, it is represented symbolically. It would remain the same in further steps of the equation. In contrast, a predictive processing system does not represent probabilities symbolically, it represents previous states of layers of neurons (which are mapped to events in the world) probabilistically. P would be a symbolic stand in for a density function (a collection of positions of variables along a continuum). Without this symbolic stand-in (P) the function could not act as a constituent of an expression.

The differences between exploring the world and exploring mathematics have been noticed long ago in philosophy. It has also been known that you cannot easily and adequately map well defined concepts with events in the world. Here we have something similar. You cannot track the world with fixed symbols but you cannot track mathematics and logic using raw probability density functions either, you need compositionality in this case. In fact, a hypothesis we can extract from this is that predictive processing systems would be subject to bias from lack of compositionality, such as mistakes in transitivity, failures in noticing necessary character of formal rules and so on, precisely the type of mistake Type 1 processing incurs in (see Tversky, 1969).

Fodor and Pylyshyn (1988) have influenced the debate between connectionists and classicists. But they certainly do not go without criticism. The arguments against them are in the form of attempts to show that neural networks could manage to work out systematicity and compositionality. Chalmers (1996), for instance, argues that if they are right in all of their points then minds cannot be based on networks, which is absurd because our brain is a network. We agree that they have criticized networks more than they should have wanted to and would respond that it might be that networks can work such features out, after all, computations should, in principle, exhibit universality.

It is plausible that a computational network should be able to achieve any computational feature. However, a network would need to be trained to exhibit compositionality whereas these

features come naturally for classical architectures, and what matters for us is that operational steps and semantic relations are tracked because of such easy access to compositionality. Also, it is clarifying to notice that when we need to train an architecture to do something really different from what it can do naturally, from what follows from its basic 'inner structure' (in Newell's terms), by installing too much modifications and ad hoc procedures, what you might end up doing is training this architecture to work in another way and thus ultimately to simulate another type of architecture. If the semantic content is not nearly directly mapped to the process of computation, then what is being computed is another architecture which in turns computes the content. Thus, for cognitive science, it is less important that any architectures could simulate another one, it is rather important what the differences of architectures are when they are in one shape or another simulated or not. The point of Fodor and Pylyshyn (1988) is that even though our brain is a network, because of some properties the mind has, this network must, at some level, to be like (or similar to) a simulation of a physical symbol system.

4.2 How classical architectures account for Type 2 features

The general idea we hold is that Type 2 processing works like a classical machine for reasoning, such as a GPS. However, it only exists in the wider setup of a predictive processing network. Thus, when facing a reasoning problem such as the ones exposed in chapter one, Type 2 processing opens a problem space containing an expression that designates the initial problem (as it is written on the paper, or better yet, how it was digitalized or interpreted) and an expression that designates a solution, which was produced by a probabilistic prediction (Type 1 processing). Having the initial expression and the predicted expression in the problem space, Type 2 processing then uses its move generators to attempt to reduce differences between them and sometimes finds different solutions in such path, or illuminates something that previously had not come about.

Move generators (or operators in the GPS) are mechanisms that apply rules, which might be fed from different sources, such as logic, mathematics or philosophy (say Occam's razor). These generators are likely to be flexible, in that they can change depending on the problem. Thus, although the basic structure is that of a logical machine that works on symbolic expressions it could

be set up to apply paraconsistent rules, for instance. This is possible because although it does not work with contradictory expressions it could work with expressions that designate contradictory expressions. Therefore, it is free to work out any sort of principle to solve tasks. It is not hard to imagine, for instance, tasks in which there are more than one solution which depend on Type 2 processing, depending on which principles are used. It can even apply generators that stem from knowledge of probability, and even still be completely different from a process which is essentially probabilistic such as predictive processing. In fact, if we take the Linda problem this is what seems to happen. The Type 2 solution uses a rule of probability "the probability of two or more events occurring together is less than or equal to either of them occurring individually", while Type 1 solution, realized by a predictive processing (essentially probabilistic) system simply goes for the amount of times 'social justice' and 'feminist movement' appeared together previously. When social justice obtains feminist movement will most likely be follow, independently of what other content is present 'bank teller'. The failure to observe the difference in the internal structure of a system (and how this constrains its behavior) and the rules it can apply has plagued the history of psychology of reasoning with confusion[89].

We want to make it clear that we are taking 'classical architecture' and 'predictive processing' both as whole packages. As we have been stressing, computations have universal features, classical architectures could work with representations of probabilities and predictive processing could be realized by a serial machine. But this is out of their standards. To claim that we are taking the whole package means that we are taking features of classical architecture and predictive processing that usually come together in all levels. Therefore, we are speaking of a classical architecture in the form of a serial physical symbol system performing heuristic search such as a GPS (Newell and Simon, 1963, 1976, Newell, 1980) which are responsible for Type 2 processes and predictive processing as a hypothesis about how networks in the brain form a system that encodes probabilistic representations of stimuli which are used to infer properties of objects in the world, being responsible for Type 1 processing.

[89] This is also related to 'the rationality wars'.

This is our general hypothesis, to show that it obtains we will proceed like we have in the previous chapter, by showing how each Type 2 feature is best explained by classical computing.

4.2.1 Working memory

In the last chapter we spoke of how working memory was not used in predictive processing literature. But why would it if it is not a feature of how these networks work. In contrast, a working memory is a necessary component of a classical architecture, both structurally and functionally. So in a Von Neumann architecture there is a primary storage for holding what to do and what is done, which is basic for the functioning of the machine. More importantly, in a physical symbol system, the model proposed by Newell (1980) for classical cognitive science, you need a similar component that stores operators and expressions which are being used in a given moment. Remembering Newell's (1980, p.159) words "This organization implies a requirement for working memory in the control to hold the symbols for the operator and data as they are selected and brought together." and "[...] working memory is an invariant feature of symbol systems".

A working memory in cognitive psychology is usually taken to be a system with executive functions and not only a storage. Perhaps there is some vocabulary mistake in this, since a memory per se does not have executive functions. However, the term is already in massive use, so we should continue with it but remembering that it also has executive functions. As Baddeley (1992, p.557) explains "Although concurrent storage and processing may be one aspect of working memory, it is almost certainly not the only feature". In fact, it is such executive functions which pushed the need for the concept of a working memory instead of just a short-term storage. Baddeley (1992, p.556) explains that "This definition has evolved from the concept of unitary short-term memory system. Working memory has been found to require the simultaneous storage and processing of information".

Instead of being just a short storage the model also includes "an attentional controller and the central executive, supplemented by two subsidiary slave systems" (BADDELEY, 1992, p556). These slave systems are storages for different type of content, such as phonological or visual. More important for present purposes are the 'attentional controller' and 'the central executive'. It seems

these claims on the processing abilities of working memory are not as clear as what has been said of its storage function. For instance, Baddeley (1992) claimed, as we showed above, that the attentional controller was an additional component, but he also claims "the central executive [...] is assumed to be an attentional-controlling system"[90].

We understand executive functions are equivalent to the application of operators in Newell's (1980) architecture or to the functioning of a processing unit of a Von Neumann architecture which carries out logical or arithmetic procedures. As for the attentional controller, it is not directly related to attention as in the psychological concept, but to 'attention' as in a Turing machine which can only focus on certain elements each moment. This function would also be something like the control unit of the Von Neumann architecture which mediates the flow of processing by providing timing and control signals.

So when we say that Type 2 processing depends on working memory we are saying that a temporary storage is needed but also other mechanisms which mediate symbol processing. We are actually saying that something like the physical symbol architecture of Newell (1980) is necessary. Certain operators must be applied to elements of this storage and there must be a control of which expressions are being used in a given moment. We have two choices here, one is to say that the concept of the working memory actually refers to Newell's (1980) physical symbol architecture as a whole, or that it is the storage component of such architecture. Since the literature (Baddeley and Hitch, 1974, Baddeley, 1992) sustains the importance of executive functions which differentiates working memory from the concept of short-term memory, we should stay with the first choice.

[90] we only added the "[...]" because of grammar, the original citation is "Working memory has been found to require the simultaneous storage and processing of information. It can be divided into the following three subcomponents: (i) the central executive, which is assumed to be an attentional-controlling system, is important in skills such as chess playing and is particularly susceptible to the effects of Alzheimer's disease; and two slave systems, namely (ii) the visuospatial sketch pad, which manipulates visual images and (iii) the phonological loop, which stores and rehearses Speech-based information and is necessary for the acquisition of both native and second-language vocabulary. Working memory has been found to require the simultaneous storage and processing of information. It can be divided into the following three subcomponents: (i) the central executive, which is assumed to be an attentional-controlling system, is important in skills such as chess playing and is particularly susceptible to the effects of Alzheimer's disease; and two slave systems, namely (ii) the visuospatial sketch pad, which manipulates visual images and (iii) the phonological loop, which stores and rehearses Speech-based information and is necessary for the acquisition of both native and second-language vocabulary."

That working memory is not only a memory but a system which has very similar (if not the same) properties to that of Newell's (1980).

As we saw Newell's (1980) architecture maintains properties of a Von Neumann architecture which maintains (or instantiates) properties of Turing Machines. So if this logic (and our hypothesis) is correct there should also be some similarity of between working memory and Turing machines.

First, it is enlightening to notice that Turing started to think about his machine by trying to mimic what he was doing in his own abstract thought[91]. The processes he was executing when doing mathematics for instance. Thus, since we must process in working memory what we are thinking consciously and with effort, which clearly was the type of thought he had to engage in for his work, what he probably was doing then was an inspection of the functioning of his own working memory. If this is the case, it would also be no surprise to find similarities of working memory and a Turing machine.

Consider this part of Turing's (1936, p. 250) intuitive argument:

> "The behaviour of the computer at any moment is determined by the symbols which he is observing, and his "state of mind " at that moment. We may suppose that there is a bound B to the number of symbols or squares which the computer can observe at one moment. If he wishes to observe more, he must use successive observations. We will also suppose that the number of states of mind which need be taken into account is finite. The reasons for this are of the same character as those which restrict the number of symbols. If we admitted an infinity of states of mind, some of them will be "arbitrarily close" and will be confused. Again, the restriction is not one which seriously affects computation, since the use of more complicated states of mind can be avoided by writing more symbols on the tape".

This description is like that of working memory in various ways. We can see that if we switch the term 'computer' by 'working memory' in this citation. By doing so every claim continues to be true. If fact, it could equally be that he is describing working memory:

[91] This is usually repeated by people but we are not sure of the exact source. One citation from his paper which shows this comparison is: "We may compare a man in the process of computing a real number to machine which is only capable of a finite number of conditions" (TURING, 1936, p.231). Also there is an intuitive argument which starts with "Computing is normally done by writing certain symbols on paper." (1936, p. 249), and goes on to explain a machine to do what we do when writing symbols on papers.

1. The behaviour of working memory at any moment is determined by the symbols which he is observing, and his "state of mind " at that moment.
2. We may suppose that there is a bound B to the number of symbols or squares which working memory can observe at one moment.
3. If working memory wishes to observe more, it must use successive observations.
4. We will also suppose that the number of states of mind which need be taken into account is finite.
5. More complicated states of mind can be avoided by writing more symbols on the storage components of working memory.

Obviously this switching of terms in Turing's words would not work were we to use `predictive processing` or `Type 1 processes`. The statements would then be false. It seems like Newell's (1980) architecture is adequate in many ways to serve as a model of working memory whereas predictive processing is not. How does this relate to Type 2 processes however?

We have been stressing that Type 2 processes are those that load heavily on working memory. So we might want to think that Type 2 processes are those that are executed by a system like Newell's architecture. On the other hand, of course working memory processes could only be restating what Type 1 processes had already arrived at. This is shown, for instance, by the computerized wason task (Evans, 1996). Also, we allow by definition that Type 1 processes might load weakly in working memory. A better claim we are ready to hold is that for us to consider a token process as Type 2, conclusions to such problem must be reached only after the use of such distinct computational methods of Newell's architecture. That is, something must be found in heuristic search which was not found in predictive processing in order for a process to be considered Type 2. Of course, methodologically it seems unlikely at the moment that we would be able to verify this in psychology or neuroscience. So methodologically we should keep the criteria of identification of a Type of process presented in chapter 2 which is just not as specific as this one.

A strong metaphysical hypothesis we can hold is that human working memory is literally a classical architecture simulated by the brain, or a component of such, and also that its executive

functions are literally the application of operators as in Newell's symbol systems. This would be a problem if the whole mind was said to work in this fashion. But in our case it is only Type 2 processes that are realized by such architecture, which are a very limited class of mental functions. A weaker hypothesis would be that Type 2 processes have similar features to that of classical architectures but there is no metaphysical commitment implied. Both do the job of solving the unity problem.

4.2.2 Decoupling

We saw in chapter two and in the last chapter that the meaning of decoupling needs detailed attention. In chapter two we saw that Osman (2013) argues for examples of effortless mental simulations. In the last chapter we argued that such effortless mental simulations could be dealt with by predictive processing. However, as we pointed out in both chapters, decoupling cannot be taken to mean any sort of imaginative production. We argued that effortless mental simulations, or situated counterfactuals, have the characteristic of being mediated by context. More precisely, that such simulations only follow from probable occurrences in the present environment. We saw that predictive processing did not account for the simulation of improbable (but reasonable) occurrences.

Now we want to claim that a symbol system applying heuristic search could account for decoupling as in simulating and using counterfactuals that do not follow direct consequences of the probable outcomes of the given states of the environment. If Type 2 processing is somewhat like the GPS system of Newell and Simon (1963), then given its inputs and its goals it should open a problem space for investigating possibilities based on its move generators. Following our hypothesis, given a problem and Type 1 predictions about it, a physical symbol system would represent these in a problem space and use heuristic search to discover (or 'make sure') if one follows from the other. In doing so, it investigates a space of possibilities (instead of probabilities) based on the generators that it will use. Decoupling occurs because this investigation is not an investigation of the world or of its probable future states but an investigation of consequences of rules and expressions.

So in the example of a person watching a horror movie. If she was an expert, it might be that just by Type 1 predictions she would figure out who the killer was. If she was not, however, or if the movie was really tricky, then she would have to explore alternatives. The problem would be 'which out of N is the killer'. The prediction would be 'character X'. But she could further ask herself, "what are the consequences of having character Y or Z as the killer?". In that case, she would not have followed a likely option but instead start to investigate other options based on rules, knowledge stated in propositions (of social relationships, for instance), logic, and whatever move generator her heuristic search could apply in this case. Of course, this does not mean that Type 1 predictions would not come about while her heuristic search was carrying on. Indeed there probably would be a strong communication of Type 1 and Type 2 processes. However, only by opening the space of possibilities for heuristic search instead of probabilities based on previous events that she could start investigating further characters which seemed unlikely, at first, to be the killer.

One of the biggest problem classical artificial intelligence had was precisely to determine the likely causes of events. That is because various improbable outcomes would be taken as important to be analyzed. As Fodor (1987) has claimed, of course in exaggeration, if a classical AI was to monitor change in the world and if it were to turn on the refrigerator, in theory, it would have to consider all the particles that change state once it is on. This is somewhat true of pure Type 2 processes. Of course their relevance are constrained by Type 1 processing, since it needs Type 1 predictions to start functioning. However, we can notice if we wonder deep in thought that we are soon to lose relevance from our task because there are so many possibilities to be investigated which the present context does not prime for. The frame problem of classical artificial intelligence was precisely that of being decoupled from the world. It is therefore, interesting to think that classical architectures would be good models of decoupled processes.

One way to put it is, if all information processing is an investigation of state spaces, that heuristic search and predictive processing investigate differently. Predictive processing would investigate a reduced space and limiting the search based on previous occurrences. Primarily investigating what is probable and what is possible leads us to very different paths. Taking some examples from science for an illustration, not surprisingly, studying what is probable leads us to statistics and probability. Whereas studying what is possible leads us to math, modal logic and

philosophy. Studying the world *empirically*, through science, ultimately requires statistics and probability. Studying *theories* of science leads us to math (theoretical physics), and analysis of how propositions stand. You need probability to obtain evidence and predict consequences of such evidence and you need possibilities to investigate consequences of axioms, rules and propositions.

4.2.3 Slowness

It is well known that you need to have time in order to be able to work out Type 2 processes. Figuring out the improbable consequences of things is not what we do naturally. In contrast, we know now that personal computers are very fast and they perform heuristic search in speeds we cannot even dream of doing. So at first it seems this is not a feature of symbol systems that is related to Type 2 processing. We suspect this is the case because of hardware conditions. As we know the first computers ever invented were much slower than the ones we have today. So it is true that having the best hardware for processing in a given way is tantamount to fast processing. In contrast, the brain is a network of cells, so simulating a classical architecture is not what is natural of it. That we organize our goals explicitly and that we investigate possibilities better than other animals seems to be true. It also seems to be true generally that we are better at Type 2 processing than other animals are. For instance, no other animal even knows what mathematics is, and are not able to explore consequences of axioms (although, of course, they can know about quantities). So it seems to be true that Type 2 processes is an unnatural function of the mammal brain. If we follow the hypothesis that Type 2 processing is the result of operations of simulated classical architecture in the brain, then it would make sense to assume that such simulated architecture does not have the appropriate hardware conditions to perform with the speed of computers built just for such functions.

Further, the feature in Basic DPT reads ‘slower in comparison’. Which in our expanded hypothesis would mean that classical architectures are slower than hierarchical bidirectional predictive processing. We do not have computers with hardwares in the forms of networks, much less ones that compute probabilistically in hardware. We only know this by simulations. Anyhow,

we do have reason to believe that networks are faster. As Fodor and Pylyshyn (1988, p.35) comment:

> "in the time it takes people to carry out many of the tasks at which they are fluent (like recognizing a word or a picture, either of which may require considerably less than a second) a serial neurally-instantiated program would only be able to carry out about 100 instructions many thousands — or even millions — of instructions in present-day computers (if they can be done at all)."

Of course, by defending classical architectures Fodor and Pylyshyn (1988, p.39) go on to argue that these are issues of the implementation level. In fact that any speed issue should be so. "The moral is that the absolute speed of a process is a property *par excellence* of its implementation." If this is the case, then, apparently we have two reasons to think that that type 2 processes in the current developing framework would be slower. First because network processing will tend to be faster in comparison and second because, as physiology teaches us, the brain does not have the appropriate hardware for the implementation of a fast classical architecture.

4.2.4 Effort

The notion that DPT brings of effort is based on folk psychology, introspection, subject report and physiological symptoms of being tired and stressed. Thus, people have a certain natural notion of what it means to be in an effortful mental state. They also report on this while having to execute tasks with higher need of Type 2 processing and eventually become more tired (see Kahneman, 2011). However, for our level of explanation, we are looking for a related information processing or computational reason for Type 2 processing to be more effortful, which of course must bear some relation to the data gathered by aforementioned means.

Like slowness, it could also be the case that Type 2 processes are more effortful because they are based in an architecture which is unnatural for the brain. However, there seems to be an even deeper reason for thinking that they would be more effortful. As we have seen, if the job of the brain is to minimize prediction error, in the free energy formulation, then the brain attempts to minimize free energy by getting predictions right. As we have been arguing, the higher the prediction error, the more Type 2 processing will be needed. Thus, higher need of Type 2 processes

are related to higher free energy, entropy and information (in the sense of Shannon and Weaver, 1949).

Piccinini (2015) explains this notion. Considering two variables a1, which has a high probability of occurrence and a2 a very low probability, the occurrences of either of them generates information since it resolves the uncertainty of the state before their occurrence. Piccinini (2015, p.229) goes on to explain that "The occurrence of a2 generates more information than the occurrence of a1, because it is less expectable, or more surprising, in light of the prior probability distribution." In predictive processing, the unexpected requires reformulation of the models. More information in computing means that more steps will need to be taken in order to solve a given problem. Of course, if something were predictable it would need less steps to arrive at. This way, we can ascribe effort even to a computer, where more effort occurs when it has to process more information, an algorithm with more steps, for instance.

More information is related to more free energy. To minimize a greater amount of free energy will take more time and work. This is more effortful than having predictions ready that minimize free energy as quickly as possible. Therefore, when probabilities fail the system needs to start investigating possibilities. It searches for other possible solutions by means of heuristic search. Heuristic search will be related to more information and time because it does not have probable solutions ready. Instead, it needs to investigate its state-space almost from scratch. We say 'almost' because it is heuristic and hence it will also have tricks to get the correct solution faster, unlike brute-force search which would investigate the state-space from beginning to end. We believe reports of effort would be related to executing more heuristic searches. Thus, reports of effort by subjects would seem to be based on cognitive-informational and physical constraints of reality.

4.2.4 Control

In chapter two we saw that the notion that some processes are controlled is not so easy to define as its counterpart of being automatic. That is because control usually is defined in relation to a subject having more decision over his thoughts and behavior. However, the subject, usually is taken to be the whole person and not a system. Since we are describing the internal systems of a

subject, the subject himself should play no role in the explanation. In contrast, automaticity is defined by characteristics of the system, as processes that are overlearned and nearly always become active in response to a particular input configuration. Also, this traditional way of looking at control is very much related to the concept of consciousness, and since consciousness is yet another feature in DPT we will discuss this notion in its topic.

Because of these conceptual troubles, we stated the opposite of automaticity to find an adequate counterpart to automatic processes. We defined controlled processes as those that must work out in real-time (do not invoke previous solutions) to produce output with some novelty for the system at that moment. Obviously, predictive processing needs to invoke previous solutions, but for heuristic search such is not necessary. Newell's and Simon's (1963) GPS had no previous knowledge of solving formal logic problems, all it had was efficient move generators that by means of comparison (or means-ends analysis) searched the problem space effectively. It is not, however, that knowledge cannot be used for a solution to be considered controlled. Rather that the solutions are produced in the moment the problem comes, if it applies knowledge to find a new solution to an old problem, that would still be a controlled process. As we have stressed, Classical architecture could have move generators based on knowledge. But using knowledge to find a new solution to a problem is different from applying old solutions to problems, even if these old solutions end up eliciting new events in the world.

Finally, classical architectures also provide control in the sense of organization. Physiology and neuropsychology teaches us the brain is a conjunction of multiple organs wired in a network. Out of all of the processes that occur contradictions are should be commonplace. As Baars (1988) and Dennett (1991) have argued, there probably is some sort of serial organization schema for ordering our thoughts and organizing behavior. Dennett (1991) and Hirsh *et al*. (2013) have argued that expressions and narratives have such power of providing control to an otherwise messy agglomeration of organs. Classical architectures are well suited for such a job such as scheduling, sorting, and limiting content to a few commands and a few items.

4.2.5 Low Capacity

As we have stressed, claiming that Type 2 processes have limited capacity means that they work with a limited amount of inputs from selective attention, they are usually disturbed by competing tasks and are limited by the capacity storage of short-term memory. This makes the idea of low capacity very related to working memory, and it could be included as part of it. However, it is important to notice that low capacity is related to working memory if we are using working memory as an explanation. A system could be limited because of other issues. Therefore, it is not granted that low capacity is related to working memory, only that it is likely. So it is interesting to see separately why classical architectures also have low capacity and why their structure explain this limitation.

We saw that predictive processing is actually a mechanism for overcoming limited capacity. In contrast, you get limited capacity as a necessary consequence of using classical architectures. The architecture exposed by Newell (1980) uses heuristic search to deal with its own low capacity. As Newell and Simon (1976, p.120) explain: "Physical symbol systems must use heuristic search to solve problems because such systems have limited processing resources; in a finite number of steps, and over a finite interval of time, they can execute only a finite number of processes." This limitation is not only a general computational problem but a problem of their architecture. In Newell and Simon's (1976, p.120) words: "[...] all universal Turing machines suffer from it. We intend the limitation, however, in a stronger sense: we mean practically limited."

The contrast they make between limited in principle and limited in practice is essential to us. All computational processes are limited because of their nature but some can overcome this by implementational strategies, such as networks. Thus, Newell and Simon (1976, p.120) go on to say that "We can conceive of systems that are not limited in a practical way, but are capable, for example, of searching in parallel the nodes of an exponentially expanding tree at a constant rate for each unit advance in depth."

They go on to explain that their physical symbol systems are unlike systems that are not practically limited. Much to the contrary, their low capacity implies that these machines need to

solve problems as if they were treated one at a time. With this we also explain the serial feature, which we will therefore not come back to.

> "The fact of limited resources allows us, for most purposes, to view a symbol system as though it were a serial, one-process-at-a-time device. If it can accomplish only a small amount of processing in any short time interval, then we might as well regard it as doing things one at a time. Thus "limited resource symbol system" and "serial symbol system" are practically synonymous. The problem of allocating a scarce resource from moment to moment can usually be treated, if the moment is short enough, as a problem of scheduling a serial machine. (Newell and Simon, 1976, p.120).

The issues of selective attention, difficulty in multitasking and short-term memory also follow from Newell's (1980) architecture. Scheduling a serial machine comes with the limitations that to have goals completed effectively focus must be given to limited tasks and goals. Attempting to deal with many goals at once would open a number of problem spaces that would make it unable to search in a realistic amount of time, resulting in computational explosion. Also, limitations in its working memory make it so that either elements are discarded and forgotten or they are reduce to a small quantity. So focusing on effective means of dealing with less content is a necessary condition for architectures that are any similar to Newell's.

4.2.6 Explicitness

We saw in chapter two how the concept of explicitness is very much related to others such as consciousness and controlled processes. However, we defined it specifically as a matter of representational format. More precisely a representation is explicit when it has an accessible representational format. By this we mean that subjects seem to grasp such content with ease and they verbally report having done so. This contrasts with content which is fuzzy and one does not know how to speak or even think clearly about it.

We can see how classical architectures can have a more fixed access to the content it deals with than networks, especially predictive processing networks. There is the difference between symbolic representations and probability density functions. As we mentioned in the last chapter, Clark (2016) admits that sometimes values in a density function need to be reduced to only one, in an all or nothing fashion. We argued this would be just like turning it into a symbolic representation. This eliminates uncertainty and we think it is related to subjects being able to grasp the content.

You can grasp something that is clearly defined but you cannot easily grasp the meaning of something like values in a probability density function in one 'shot'. They are fuzzy because they cannot be simply well defined. It is precisely their fuzziness that allows for context-sensitivity and fast processing. When you digitalize this whole density function into a symbol you lose its variability but you gain in certainty. It is as if you pretend that such symbol accounts for the function. A simple way of thinking about this is by alluding to an analogy in math when mathematicians condense a whole function into the variable X. When they say "let us call this huge function F(x)", F(x) refers to a huge amount of content, but the symbolic stand-in does not carry with it all of the properties the referred object had. However, the symbolic stand-in is graspable.

The reason it is graspable seems to be because working memory can store it and use it in symbolic manipulation. Working memory cannot store all of the values in a probability density function. But it can store F(x) as stand-in for such values. When it stores F(x), this symbol can then be treated as a constituent in an expression. When that occurs, the classical architecture can easily work with compositionality. The point for compositionality in making content accessible is that manipulations to these expressions can then be easily tracked. Rules and semantic content become related to the inner structure of the computation. Then, when taking a content as a symbolic object we can easily identify it in various expressions and we will not lose hold of its identity. Whereas values in a density function might lose their identity, in fact, we should want that to happen if context is to shape their identity.

Even the steps in processing become objects because they too can become stored expressions. Therefore, when we are reasoning in a syllogism we can keep premises in working memory and also the steps used to extract one from the other. Of course these are fleeting, but also, the way to make them less fleeting is by reducing uncertainty and naming a step or a premise by a letter, say P. So it seems plausible that representations in classical architectures should make both content and steps of processing more graspable because of ease in determining their identity, reducing uncertainty. Therefore, if our hypothesis holds until now, we should want to speak of explicit representations as symbolic and implicit ones as probabilistic or multi-valued.

4.2.7 Consciousness

Type 2 processes are said to involve consciousness. Consciousness does involve many theories and definitions. In chapter two we used consciousness as in the access the subject has to his own mental contents. However, as we have stressed, because of the risk of regress, the concept of a subject cannot play a role in the explanation of the subject's mind. So when we speak of access we must mean that some processes have limited and special access to mental contents and that such processes are causally related to subject's reporting awareness.

We know that much goes on in our brains that we could never know about. For example what processes are behind the computations of an image. However, it also is the case that we can know the processes behind the solution to a reasoning problem. So when we solve a reasoning problem we might remember the steps taken and what kinds of rules we had to apply in order to reach the solution. Now, if we know how we did it and we can report this various times then it is probably the case that in our cognitive mental architecture there is a storage for this sort of content where there are none for other types of content.

To say that Type 2 processes are related to consciousness is thus to claim that much of such processes are at least temporarily stored[92]. Thus, we come back again to short-term memory. If there is no short-term memory tracking which steps were taken in the recent past, then it is reasonable to assume that such content would be lost. Also, if a lot is going on in parallel such as in networks it is harder to track, store, or talk about these processes. If the content is lost, how could the system know about it? So to have consciousness of some content or some reasoning steps, storage is needed.

It should come as no surprise that classical architectures have exactly such sorts of storage. They have storages for representing the current state of the machine, the machine in a few states before, as well as the operators that were used. One could then ask, "why aren't classical architectures in artificial intelligence conscious then?" The answer is straightforward, it is because the concept of consciousness we used is simplified. The concept of consciousness in the question above refers to something complete, with phenomenal aspects, behavioral aspects, causal aspects

[92] Remembering that explicitness refers to the content which has an accessible format which *can* be stored, where consciousness here refers to the access of content that are (currently) being stored in working memory.

and also storage aspects. When we speak of consciousness as in access we are limiting the concept. This is necessary in order to work out the details of a cognitive process. Thus, in effect, all we would be saying is that classical architectures can have access to its 'states of mind' (to use Turing's expression). Which only means it can recall expressions used and return to previous states.

4.2.8 Cognitive ability and normative responding

Cognitive ability seems to be related to flexibility in thought. It should seem pretty clear now how classical architectures enable flexibility in thought by analyzing possibilities. Although they are intrinsically deterministic systems, which might suggest rigidity, in truth, the way heuristic search functions causes effects much to the contrary. As we have seen, computing based on probability can be an excellent option for getting through the world we live in. However, it might also be that focusing on previous events constrains thinking in such a way that it does not see possible (according to relevant rules) but improbable alternatives. Heuristic search investigates the problem space slowly in a way that allows for other possible alternatives, enabling flexibility. Thus, it seems likely that individuals that use more Type 2 processing would do better in cognitive ability tasks.

Accordingly they also would tend to focus on normative responding. Because heuristic search in an investigation of problem spaces based on rules, or consequences of axioms, it is likely that answers from such system will be aimed at normativity. This does not mean it will always answer according to normative standards, only that it will be aimed in such direction. This means that, when a classical architecture for reasoning such as the GPS fails, its mistakes will be attempts to act normatively. These sorts of mistakes occur in the GPS and in people when using Type 2 processes. These are fails such as in applying a rule, or in getting the wrong order of an operation, which are different from fails of Type 1 processing which are usually unrelated to the normative response completely but are just related to how previous occurrences have happened together.

4.2.9 Other correlational features line

We have spoken about the seriality of Newell's (1980) architecture and so what is left is that Type 2 processes are abstract and recent in evolution.

The property of being abstract is related to the symbolic function. When you abstract away from content, what you seem to do is grasp a few more important items that need further working on. It is unlikely that classical architectures can handle the process of abstraction itself. However it should be the case that they work mostly abstracted symbols rather than with objects rich in content themselves or even how these objects are usually represented in the brain. A symbol is a pretend item that does not account for the completeness of density functions related to the object it represents much less to the object itself. So in a sense it is an automatic abstraction. But there is a further sense in which symbols can be grouped into classes and thus be abstracted in such a way. Here compositionality seems to do the trick, since including objects as properties of classes is enabled by transitivity allowed by compositionality. Of course classes here are abstract and more rigid then when we classify items in the world by perception. The latter probably does not work by means of classical architecture for reasons we will see in the section to come.

Finally, there is a sense in which, if there is a classical architecture being simulated in the brain, that it must necessarily be more recent in evolutionary terms. That is because if it depends on the brain's networks for existence then it means these networks were already doing something else before. That is, if there is such a thing as a classical architecture in the mind, they must necessarily come as an adaptations to existing organs, and not as the main functions of the organ itself, since as we know in physiology there is nothing like a classical hardware in the human body.

4.3 How classical architectures fail to account for Type 1 Processing

We have been over the features of DPT various times in the thesis now and if our arguments that Type 2 features are explained by classical architecture and that Type 1 features but not 2 are explained by predictive processing, then the reason why each Type 1 features are not classical should also already be deductible. Thus, to avoid redundancy we will in this section by other means and show that limitations of classical architecture related to these features have been evidenced by malfunctioning of artificial intelligence, under the name of the frame problem. With this point we

close our work and we hope this ending will function like a bootstrap knot which further explains how Type 1 process are based on predictive processing and Type 2 based on classical architectures. This bootstrap knot is also related to reaching success where Fodor had failed in the relation of his dual process theory with cognitive architectures.

An objection to the core of our thesis could be raised: “well since a classical machine is universal, since it can apply even probabilistic rules and norms, why would there need to be different types of processes such as in dual process theory?” Again, the answer is because of differences in processing times, sensitivities and reliabilities. For instance, in principle, in an idealized environment, a symbolic expression manipulator, such as Newell’s (1980), could be used to sense, navigate and act in a world relevantly. However, in practice, in our world, that is not so, because it would take too long to process the needed content, it would be hard to sense the most important facts, and thus it would be unreliable for the task. This is well known now by the failures of artificial intelligence, which we will explore in the next section. Our point is that these failures demonstrate that architectures based on Newell’s (1980) or similar ideas will not work for perceptual, context-dependent, quick reasoning, and ‘perception-like’ functions. This is recognized by Newell and Simon (1976, p.124): “Human levels have not yet been nearly reached by programs that have a complex perceptual "front end": visual scene recognizers, speech understanders, robots that have to maneuver in real space and time.” In fact, they never did reach any relevant level in these ends. In contrast, what is working now for speech comprehension, visual scene recognizer and maneuvering in space and time are networks and rich statistical models (see, for instance, Dahl, *et al*., 2012 and Lotter *et al*., 2016).

As we have stressed, Kahneman & Frederick (2002, p.50) claimed that intuitive thinking is “perception-like” and that “intuitive prediction is an operation of System 1”. Further, that “The boundary between perception and judgment is fuzzy and permeable: the perception of a stranger as menacing is inseparable from a prediction of future harm”. Our argument then, aside from what we have already stressed about each independent feature, is that if this is even somewhat correct about how Type 1 processes are like, then we would expect classical architectures to fail in working with them because of the frame problem.

4.3.1 The frame problem of classical architectures

The frame problem emerged as a result of a methodology in AI which is now retrospectively called Good Old-Fashioned Artificial Intelligence (GOFAI, see Boden, 2014 for a recent review). This sort of AI is an instantiation of more abstract work such as Newell's (1980). The project of GOFAI is to make the computer agent represent the world by means of formal symbols which stand for aspects of the world in an internal model. By means of an external programmer, instructions are given for how to manipulate these formal representations.

McCarthy and Hayes (1969), following the GOFAI agenda, started to develop their AI by naming the complete state of the universe at a given time a 'situation'. However, given that the universe is too large, only facts about the situation were to be represented, instead of the whole situation. These facts were used to infer further facts about the current situation or about future situations. It is understood that an action changes the current situation, however it does not change all elements of the situation, there are things that change and things that do not change in a situation given the action. A problem arose in the task of discovering how an agent can determine, in a situation, what should change and what should not, given a specific action. Originally they had to type in an incredibly huge list of statements, called frame axioms, determining what would not change in result of an action. McCarthy and Hayes (1969) saw this issue as one amongst many other technical difficulties they had discovered and coined the term frame problem to refer to this specific difficulty. However, this technical difficulty turned out to be serious enough to require extensive debates between philosophers and artificial intelligence scientists (see Pylyshyn, 1987).

To illustrate the frame problem, Dennett (1987) asks us to imagine a robot whose task is to go into a room and get an extra battery from a wagon that has a time bomb attached to it. In its first attempt, it pulls the wagon out of the room but explodes, since it brings the bomb along. So its designers had to teach it how to plan considering the side effects of its actions (considering changes and non-changes of the situation after its action). However, since it had to consider so many non-changes, such as that the color of the room will not change, that the wagon will not change its

shape, that the battery will not change any of its physical properties either, and that the wagon will not break by being pulled, the bomb exploded before it could act. Hayes (1987) argues this is the central point of the frame problem. That since these things could in principle change, these frame axioms were needed in order to assert that in that specific time, in that situation, there would be no changes.

> In this ontology, whenever something MIGHT change from one moment to another, we have to find some way of stating that it DOESN'T change whenever ANYTHING changes […] That is the frame problem. If there are 100 actions, and 500 labile properties and relations, then we might need 50,000 of these silly "frame axioms" [...] (Hayes, 1987, p. 125).

As Janlert (1987), Haugeland (1987) and Dreyfus (1972) notice, this seems to be a problem that arises from representing the world in such quasi-linguistic form, considering situations as the state of the universe and dealing with knowledge in encyclopedic statements. Various sub-instances of the frame problem for classical architectures can be identified (see Pylyshyn 1987 and Bellini-Leite, 2017), for our current purpose let us take the problem generally.

Janlert (1987) argues the original frame problem is just one manifestation of a general problem of dealing with the phenomenon of change. His general definition is the following: "The general frame problem is the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented" (Janlert, 1987 p.7). Dennett (1987) defines it a little differently. He suggests the problem is one of making good use of the knowledge one has given the requirements of the time-pressured world. Dennett (1987, p.52) suggests "the frame problem […] concerns how to represent (so it can be used) all that hard-won empirical information […] a problem that arises independently of the truth value, probability, warranted assertability, or subjective certainty of any of it. Even if you have excellent knowledge (and not mere belief) about the changing world, how can this knowledge be represented so that it can be efficaciously brought to bear?" He argues the agent that solves the frame problem ignores most of what it knows and works with the portion of his knowledge that is relevant to the situation at hand. Also, it must know how to effortlessly choose the portion of knowledge he is to work on.

Fodor (1987, p.148) emphasized the relation of the frame problem with induction: “The frame problem and the problem of formalizing our intuitions about inductive relevance are, in every important respect, THE SAME THING”. This relation to induction led him to characterize the frame problem as Hamlet’s problem. “If, for example, you undertake to consider a nonarbitrary sample of the available and relevant evidence before you opt for a belief, you have the problem of when the evidence you have looked at is enough. You have, that is to say, Hamlet's problem: when to stop thinking”.

And in 2001, his focus was mainly on abduction. He suggests that a reliable abduction uses the whole background of epistemic commitments to plan and to fix beliefs. However, in a feasible mechanical agent the computability constraint does not permit the use of such a giant web of beliefs for a single planning or belief fixing check. So Fodor (2001, p.38), concludes “How to make abductive inferences that are both reliable and feasible is what they call in AI the frame problem”. Fodor (2001) came to believe that cognitive science can only deal with modular, informational encapsulated processes, all else being impossible for the science as it is, especially AI which has to put the theory in practice: “So if a lot of quotidian cognition is abductive, and if there are intrinsic tensions between abduction and computation, why would you even expect that our robots would work? The failure of our AI is, in effect, the failure of the Classical Computational Theory of the Mind to perform well in practice” (FODOR, 2001, p.38).

Of course these various claims cannot be taken so literally together, if the frame problem were the same as the problem of formalizing induction than it would not be the same as the problem of abduction. However, Fodor does seem to have a unified position. His main point is that computation will never deal with a contextually changing environment because it cannot deal with information that is not encapsulated. And such follows from his philosophical commitments “Duhem and Quine were right that considerations relevant to rational epistemic assessments can come from anywhere in a belief system” (Fodor, 2001 p.32).

Finally, Dreyfus and Dreyfus (1987) also argue there is a general frame problem. In fact, Dreyfus and Dreyfus (1987) was the first one to hint at it. Their contribution to the analysis of the problem consists primarily in applying a distinction between knowing-that and knowing-how. They criticize Dennett (1987) for thinking that what enables us to do quotidian tasks, such as

household chores, is our knowledge of the facts about those things. Dreyfus and Dreyfus (1987) argue that using "knowing-that" alone to solve mundane tasks is what generates the frame problem. In contrast, they suggest representing the world cannot help agents in dealing with situations in the fashion AI developers desired. "If we stood outside the world and represented states of affairs as meaningful objects, situational similarity recognition would be mysterious, indeed." (Dreyfus & Dreyfus, 1987, p.101). They argue, in contrast, that we do not process sentences of facts about the world, rather we use our know-how.

It seems Dennett, Dreyfus, Janlert, and Fodor all agree that the frame problem is one of dealing with relevance and context. As Pylyshyn (1987, preface X) summarizes: "This, the problem of relevance, is what many believe lies at the heart of the frame problem, and which will continue to be a serious problem long after all the minor technical problems […] have been dealt with."

We believe that the type of processes that enables an agent to act relevantly in a context-rich and changing world are Type 1 processes. When our internal processes are tuned with the environment around us they work effortlessly, fast, automatically and with little use of working memory. This is not such a simple conclusion however, Fodor had imagined just the opposite. That is why it will be important to speak of his diagnosis so we can properly invert it.

4.3.2 Fodor's dual process theory and the frame problem

Fodor is best known for being one of the preeminent defenders of classical architectures. However, it is also true that, like Dreyfus, he worries that the classical architectures cannot be a decent model for the mind. This has puzzled some (like Pinker, 2005) over what Fodor's point is. Well, this apparent puzzle is shattered by understanding Fodor's division of mind into two. As we have seen in the introduction of this thesis, Fodor has a dual process theory of his own and understands cognition is divided into modular input systems and global central systems. The main difference that separates these two types of cognition (thus answering the unity problem in his view) is the amount of content they are sensitive to. Therefore, Fodor (1983) argues modules are sensitive only to their inner content, they are encapsulated, and are thus 'unaware' of what happens in other modules of the mind. In contrast, central processes are sensitive to any content present in

a person's web of belief, they are thus isotropic and Quinean. He understands that while classical architectures have been successful for understanding modular systems, they will never be a framework for isotropic and Quinean processes.

Fodor (2001) also argues that for the classical architectures to work as a complete model of the mind, our minds would need to execute only local and not global processes because the causal powers of classical architectures depends on syntax. Classical architectures work with computations on mental representations and such computations are sensitive only to the syntactical properties of representations and these cannot vary according to context. Fodor (2001) claims the identity of a mental representations depends on its syntactical form and such form would need to vary in global abductive thoughts. Fodor (1983, 1987, 2001) understands the frame problem proves his point, since serious success has been reached in understanding non-contextual systems, and the attempt to use classical architectures to model central processes has not worked. Fodor's diagnosis of classical architectures is, therefore, not very much puzzling if you understand his duality of mind. However, his diagnosis forces him to commit to a strategy in which the problems found in artificial intelligence are not used to reshape his theory, rather the theory stands still and the problems are abandoned as impossible or as needing profoundly (some radical change in foundations of cognitive science) new insights.

4.3.3 Inverting Fodor's diagnosis

Fodor's diagnosis seems quite inadequate. Instead of admitting that something is wrong with his duality of mind he rather wants to claim that the problem is in cognitive science itself. We think that if we tweak his diagnosis, we can see that it is his duality of mind and not cognitive science that should be abandoned. This can be done if we realize what exactly is flawed in this diagnosis. There are two points to be made. First, it is true that the classical architectures as he formulated (Fodor, 1975, Fodor and Pylyshyn 1988) are limited in the ways he mentions. However, it is false that cognitive science is limited to his formulations of the classical architectures. We understand connectionism, dynamical system approaches, reactive approaches and the predictive processing approaches are also information processing theories that are somewhat limited but not

in the exact same ways as classical architectures. The second point is that if it were true that perceptual systems were as encapsulated as he claims, then in fact it would be impossible for central processes to deal with content. However, this is true only because he delivers too much load for central process alone to deal. In contrast, it has been one of the main successes in cognitive psychology to show that central capacities are actually limited, as evidenced in short term and working memory (Miller, 1956), attention (Simon and Chabris, 1999), and competing tasks with divided attention (Schiffrin and Schneider, 1977). If Type 2 processes are inherently limited, as are classical architectures, why should they have to deal with global contextual processing? The mistake here must be in theory, not in these famous studies presented by cognitive psychology.

Our proposal is rather that Type 2 processes do not deal with such global contextual processing, they are limited, in just the way frame problem constrained processes need to be. There must be, surely, some other explanation for how we are able to entertain context sensitive processing. In our view it is Type 1 processes which provide such contextualized content when we stop treating them as encapsulated limited machines and understand them as embedded in a sea of Hierarchical Bidirectional Predictive Processing (HBPP). But supposing Type 1 processing provides contextualized and relevant information for Type 2 processing then even though Type 2 processes are limited, they work only with relevant information to begin with, they are, if you will, somewhat blind to useless content.

We do not want to claim that abduction is realized independently by either type of processing. Abduction has aspects of intuitive and thoughtful processing and is most likely an emergent capability spanning from the interaction of these two types of processes and most likely others not directly involved in reasoning. As we have seen in the last chapter when we spoke of autonomy, switching encapsulated systems with HBPP has implications for relevant and contextual cognition without requiring central processes. If Type 1 problems follow predictive processing and Type 2 processes follow classical architectures, this will provide an answer to the unity problem which renders DPT feasible and cognitive science existent[93]. Thus this inversion of Fodor's diagnosis can be seen as one of the main insight which drove us to write this work.

[93] (alluding to Fodor's 'first law of the nonexistence of cognitive science')

The mistake Fodor made was in supposing that Type 2 processes had to be contextual and tied to relevance, when in fact it is not. It is rather that Type 1 processes brings only the relevant content for Type 2 processes to work with. Stanovich (2004, p.48) had also identified that Type 1 processes are inconsistent with the traditional effort of artificial intelligence. He says, where 'TASS' stands for Type 1 processes and 'analytic system' for Type 2 processes:

> "Such a view of the differences between TASS and the analytic system is consistent with a long-standing irony in the artificial intelligence literature: those things that are easy for humans to do (recognize faces, perceive three dimensional objects, understand language) are hard for computers, and the things it is hard for humans to do (use logic, reason with probabilities) computers can do easily. The current view of the differences between TASS and analytic processing removes all of the air of paradox about these artificial intelligence findings. Computers have built up no finely-honed TASS subsystems through hundreds of thousands of years of evolution, so the things that the massively parallel and efficient human TASS systems do well because of this evolutionary heritage computers find difficult. In contrast, the analytic system of humans—the serial processor necessary for logic—is a recent software addition to the brain, running as somewhat of a kludge (in computer science, an inelegant solution to a problem) on massively parallel hardware that was designed for something else. In contrast, computers were originally intentionally designed to be serial processors working according to the rules of logic." (Stanovich, 2004, p.48).

We believe this long-standing irony existed because artificial intelligence was treating the whole mind as if it were composed of Type 2 processes. With this changing, the progress of artificial intelligence seems also to be developing in contextual relevant areas such as visual recognition and speech comprehension (you can easily notice this by using face recognition and voice recognition apps on smartphones, but see also, for instance, Dahl, *et al*., 2012 and Lotter *et al.*, 2016). If our analysis is correct then predictive processing should at least be able to shed some light on some aspects of the frame problem or function as a framework able to treat the problem differently. Therefore, to finish the argument we will explore in which ways predictive processing could propose changes for the frame problem of classical architectures.

4.3.4 Possible changes to the frame problem of classical architectures

This section has no intention of claiming that predictive processing will have an a priori solution to the frame problem. It is rather a sketch to show how it might propose different strategies

to be tested, strategies we cannot have any a priori guarantees will work, that do not follow from necessity, but that seem to differ quite a lot from the ones of classical cognitive science had proposed and could turn out to be feasible in real-life information processing agents.

The problem of modeling change and non-change resulted from an attempt of representing a changing world in static representations. Janlert (1987) argues we would need a representational form permitting a changing, complex world to be efficiently and adequately represented. In predictive processing the world is not represented, as in entirely mirrored, but there is a function of tracking this changing world using correlations stored in probabilistic representations. It is by having correlations between inner states related to stimuli (instead of represented facts) stored in probability density functions (instead of in discrete symbols) that predictive processing can have an alternative for modeling change.

Further, taking stored information of past stimuli correlations and contrasting them to current patterns of sensory input by applying Bayesian inference leads to predictions about what sort of stimuli could perturb the perception system next. This allows for simple, frugal, moment by moment, predictions about future states of the world. Also, objects are not identified discretely but rather always probabilistically. When identifying a chair, it is not that our cognition has a representational format that is able to put together the exact characteristics that form a chair in every situation. Rather predictive processing provides means to identify objects by likelihoods and similarities. Thus, a three legged metal object might invite sitting as other four legged wooden ones have because they share at least some similar properties. The point is not that this sort of qualification helps define a chair, but that it enables the agent to decide to sit or not. Definitions are not necessary in order to allow agents to act fluidly in the world.

Strategies for dealing with time pressure are incorporated in every aspect of predictive processing architecture. The whole advantage of having a system that is always predicting is precisely staying ahead of time. So, every solution found in a predictive processing architecture will be approximate and will sacrifice precision for speed. As Clark (2016, p.250) puts it:

> "Cheap, fast, world-exploiting action, rather than the pursuit of truth, optimality, or deductive inference, is now the key organizing principle. Embodied, situated agents, all this suggests, are masters of 'soft assembly', building, dissolving, and rebuilding

> temporary ensembles that exploit whatever is available, creating shifting problem solving wholes that effortlessly span brain, body, and world."

Instead of having to organize a multitude of facts about a changing world, updating 'beliefs' in predictive processing occurs every time error units flag mistakes in models proposed for acting in a given situation. Hence, if this is true, perception is, by its own explanatory mechanism, a process of updating generative models.

Also by using prediction one already is implicitly searching knowledge that was related to the environmental item in the past. So, at least in a first moment, probabilistic information will be posed and not any other type of knowledge will be brought along

. As Clark (2016, p.296) explains: "Perception-action loops are fundamental; low-cost, representationally efficient options are preferred; and the continuous stream of error-minimizing action allows for the recruitment and use of arbitrarily complex suites of external resources—resources that are now simply swept up in the ongoing circular causal flow." The predictive processing architecture is argued to have biologically realistic applications (see Friston 2005, 2010), since it follows computational neuroscience, that is, it is worried about how neurons and real (in contrast to artificial) neuronal networks could apply such computations. For instance, Friston (2005, p. 823) while explaining the value of having a hierarchical organization with error and representation units argues for a realistic implementation:

> "Responses of (representational) units […] depend only on the error at the current level and the immediately preceding level. Similarly, the error units are only connected to representational units in the current level and the level above. This hierarchical organization follows from conditional independence and is important because it permits a biologically plausible implementation, where the connections driving inference run only between neighbouring levels."(Friston 2005, p.823, our added parenthesis)

Predictive processing is also related to feasible machine learning (see Dayan et al., 1995) and realistic new attempts at robotics (see Park et al., 2012; Yamashita & Tani, 2008). It could be argued that Bayesian optimality is intractable, and so the underlying mathematical structure proposed by predictive processing would fail to work well. However, as Clark (2016, p.259) argues, because of the goal of reaching feasible applications, "Implausible implications of pervasive brute optimality are thus abandoned in favour of strategies that deliver some combination of efficacy, reliability, and energetic efficiency".

Wheeler (2008) mentions that there is a difference between sensitivity to the current context and crossing knowledge from one context to the other. As Clark (2016, p.141) notes "within the predictive processing paradigm, such context-sensitivity becomes ´[…] pervasive and 'maximal'. This is due to the combination of hierarchical form with flexible 'precision-weighting'". So the difference in these two types of problems seem to stem from where precision is allocated. Intra-contextual (in one context) problems will most likely be related to focusing on error unit corrections while inter-contexual (cross-contextual) will necessarily invoke strengthening precision in representational units. There are surprising details of an event's current contingencies that only error units will be able to flag since they will not be encoded statistically by generative models and these will be necessary to act successfully intra-context. On the other hand, what best helps an organism in another situation might be to apply knowledge from other related domains, and these relations will be brought about because the current state of the world will perturb the perceptual system in some related way to what it has done so in other contexts, and this relation would be captured by predictive processing by means of the correlational data perception works on, bringing about the relevant cross-contextual knowledge. There have been recent proposals on how this (transfer of knowledge) could be accomplished (see Salakhutdinov, et al., 2013).

Finally, predictive processing is a great example of how to overcome problems related to treating the whole mind as symbolic or propositional expressions. Nowhere in all the work that has been done in predictive processing has it been stated that propositional content needs to be stored and worked with in order to achieve effective processing. Processing in this framework occurs by Bayesian inferences in probability density distributions. So predictive processing does explain how cognition could work without storing facts in propositions. Much to the contrary, the problem for predictive processing is explaining the contrary, how propositions could be well formed, organized and used in a HBPP structure.

So although we cannot say that we have a solution the frame problem, it seems we can say the problem is more directly related to classical architectures. Thus, we understand that these failures of the classical architecture (where predictive processing not necessarily fails) shows how poor it executes in tasks which are seamlessly effortless for humans. Such tasks, we have argued, are related to most features of Type 1 processes such as: working with less load on working

memory, autonomously, faster, nearly effortlessly, automatically, with high capacity, intuitively, in parallel and contextualized. Therefore, it seems plausible to conclude that problems in artificial intelligence (GOFAI) have shown that classical architectures cannot handle Type 1 processes very well.

CONCLUSION

Now at the end of these four chapters we feel safe to conclude that the case for a dual framework for DPT is, at least, plausible. We have argued that predictive processing can account for Type 1 processing while classical architectures are still best for Type 2 processes. This endeavor is meant to solve the unity problem as posed by Samuels (2009). From chapter three we could see that the predictive processing account explains why some reasoning processes would have Type 1 features. In the last chapter we explain likewise, how Type 2 features stem clearly from the classical architectures.

We are aware that by having an argument for each feature of DPT in relation to both frameworks we have probably said too much. What we mean by this is that some of these arguments might turn out to be wrong. Although at the same time, if some of them are wrong, this does not disqualify the thesis, as most of them, we believe will be correct. This is similar to the defining feature vs. correlational feature issue. We think there are some features which really make the case for our proposal, and it is these features that we can properly argue to be explained, in a stronger sense of the word, by the framework we are proposing. We can name a few: Working memory as a component of classical architectures, explicit representations as symbolic and implicit ones as probability density functions; Effort as free energy accumulation and automaticity occurring when predictions work with low-error.

Because we focused mainly on showing that features could be related to these frameworks we did not have space to work out various other topics of interest, such as elaborating an answer to the reference problem, dealing with how the two processes interact in more detail or showing how the framework as a whole explains evidence in the reasoning literature. To cover for this fault, we will leave hypothesis on these topics with instructions for further research.

We left the reference problem open for Basic DPT and we had good reasons for doing so. However, it seems our conclusion, which goes beyond Basic DPT, seems to imply a few things about the reference problem. We believe much of the brain could work like predictive processing. But it is not everything that processes in that way which could be said to be Type 1. Therefore, we would need to specify what it is that works in such way that can be said to be Type 1. We understand that any predictive processing that influences a response to a personal-level problem solving task directly can be said to be Type 1. What we mean by directly is that it should figure in the answer to this 'important question': "what cognitive reasons caused X to respond to problem Z with the answer Y?. What we mean by personal-level is that such a problem cannot be something like what neurons have to deal with in image processing, but rather, something that is recognized as a problem by a person. That is, which can be verbally stated as a problem the person recognizes to be solving. However, there is no sense in which there is a system 1 for problem solving. For instance, a perceptual system might be involved in a judgment of face recognition that answers the problem: "decide which of these men seem to be the most angry to you and why". Therefore, we cannot frame what systems will always be involved in problem solving and which will not, they will vary according to task. Also, the statement that multiple systems will be involved is probably true. But there is also no sense in referring to these multiple systems as Type 1 systems, since the same systems could be involved in other tasks or even in a Type 2 response.

Type 1 refers to processes, processes executed to help in problem solving which can stem from multiple systems. However, all of these processes will have been processed in a HBPP fashion. Type 2 also refers to processes, processes executed to help in problem solving which can stem from multiple systems but that necessarily involve intense participation of a classical architecture system with a working memory. Now, likewise, we cannot say that this classical architecture is a system 2, because it might work for other things, such as rehearsing, focusing attention to perceptual events, helping in speech, and so on.

One could be tempted to say that there is a system 2 which is this classical architecture simulated by the brain and a system 1 that is the rest of the brain working. However, we are unconvinced that pairings of features in DPT would hold for all sorts of tasks. These would probably vary for recognition tasks, singing, dancing, talking, and so on. And as we stressed

extensively in chapter two it is the clusters of features which define DPT, if we change these clusters freely, there would be no DPT to defend. So we do not think we can speak of system 1 or 2 or multiple system ones and twos. What we can say is that there are probably various systems in the brain, using diverse computational principles and when an answer to a problem-solving task was generated by processes following predictive processing then we can say we have a Type 1 process. Whereas if the answer to the problem-solving task was directly influenced by a classical architecture simulation in the brain then we can say we have a Type 2 process. Of course the identification of types of processes is to be done via features, as Basic DPT methods established in chapter two. The simple answer to the reference problem would then be 'types', instead of systems, minds or modes. What we did was just explain a bit more of what we mean by 'types' then previously had been done in the literature.

Stating that the reference problem is solved in such fashion has a few implications. Brain systems are left for what they are: control systems, emotional response systems, language systems, spatial systems and so forth. They are not to be further characterized by DPT. One could wonder what systems apply these two distinct kinds of computational principles, and thus argue that our proposal also avoids the explanation of the reference problem. The answer is that various brain systems might apply these computational principles, however, there is no use in or safe way to refer to systems as being in the class 1 or 2 on the basis of which type of processing they use. For instance, it could be the case that visual or face recognition systems apply predictive processing for quick judgment or to start a chain of inferences. There is no reason to suppose a visual system should be termed a System 1, and the fact that they might apply predictive processing does not make the case. It is reasonable, in contrast, that a language system will work both types of computational principles, which would leave us with a doubt on how to classify it in CKT.

We do not want to claim that the brain as a whole applies both sorts of computations 'the brain as a whole' is a strange object, since it applies countless functions; to say that the brain applied a computational principle in a general sense is perhaps another way to avoid the reference problem, since the answer is not precise. It is more appropriate to think that each instance of a Type 1 or Type 2 application could, in principle, be localized. However, this would be ambiguous for our current technology to determine, since we would expect these to be applied by the systems

which are designed to the task at hand. So what we would see in neuroscience if this were correct is only what we already see. That is, when the task requests control, language and visual functions, we see control, language and visual brain systems activate, since the computations are being applied in such areas in those cases. Therefore, this answer is barely predictive for science; it is just a philosophical solution to the reference problem. Again, this is why DPT's scientific value is set in its features discrimination and not in an answer to the reference problem.

On the other hand, however, this proposal helps guide research to the best track. For example, if accepted, it will stop researchers from searching for two different systems for reasoning in the brain. It is plausible to think that an approach like Schneider and Chein's (2003) is possibly more fruitful, since it could be read not as attempting to map systems 1s or 2s but as attempting to discover how specific regions react when important distinctions in DPT are tested, in this case, the difference between the brain in automatic versus the same brain in controlled processing. Unlike Schneider and Chein's (2003) proposal however, we favor the idea that such shifts will not be task-independent. Rather, each task should require a different arrangement of brain functioning for Type 1 versus Type 2. It is unlikely that all Type 1 responses to different tasks should be correlated with the same patterns of activations. For instance, Type 1 answers to math problems should have little in common with Type 1 judgments of human behavior. Although, following the results of Schneider and Chein (2003) and DPT predictions we should tend to see a more homogeneous pattern even cross-task for Type 2 processes, since it will recruit fixed limited resources such for conflict resolution and working memory. But we would not identify a System 2 then, rather just working memory and other 'central systems'.

If we say there are two computational principles related to each type of processing, and if these principles are applied in areas which are not directly related to thinking and reasoning, then we should expect to find similarities there, since it is these computational differences which explain the cluster pairing. If one is satisfied with the cognitive kinds thesis (CKT) then he could go on to say that what characterizes a system as 1 or 2 is whether it processed via predictive processing or via symbolic processing. However as we explained in chapter two we found a methodological issue in CKT where multiple features could be mixed with multiple theoretical suggestions and we would not be able to tell them apart. One solution for a DPT of the mind and not only of reasoning is

perhaps to find the contrary, instead of multiple features, just a few agreed upon ones. If this is found we believe it might be related to the core distinction of two computational principles (HBPP and Classical) governing all of cognition. If we had to name a few that might apply to the whole mind it would be similar to those we mentioned were better explained by our core thesis: Working memory, explicitness/implicitness, effort and automaticity.

We must add a caveat for our general proposal. We probably should not want to claim that all that has been described as a Type 1 process will be accounted by predictive processing. That is because too much has been described as Type 1. However, if our explanation works well and more often than not, then perhaps it can even clarify what should be put together as Type 1 (predictive processing) and what should not.

As for the default-interventionist model in our proposal we have that predictions will always come as default answers. Interventions will be executed if, when predictions are given low precision weigh, answers in heuristic search on problem spaces using relevant generators come up with new possibilities. This is also probably how training of Type 1 processes by Type 2 processes occur. If the heuristic search continually shows a new path, then eventually predictive processing schema will migrate to this new solution. After the new successful solution is found a couple of times, similar problems will stop being treated by Type 2 processes and come by default as a prediction (like in automaticity training). We also hypothesize that the highest the free energy the more Type 2 processing will be necessary, and of course, when free energy is at minimum is when predictive processing is working in its best. Thus, we have here a few hypotheses for how the default-interventionist model could work in cognition that could be explored in further work. It is likely that a solution to the frame problem will come from using the best of what different computational principles can generate for acting fluidly in the world. In this sense, we hope our framework could at least displace the myth that Type 2 processes are context-sensitive abduction systems.

We have mentioned in chapter four how HBPP accounts for intuitive (pragmatic, evolutionary and heuristics) explanations of the evidence. We also want to add that there seems to be various cases where Type 1 processing responds to what is usually previously found and Type 2 responses are derived from logical or step-by-step reasoning. For instance, in Piattelli-Palmarini's

(1994) example of the lead and feather heuristic we can suppose that people will think that the weight of an object determines the impact it can cause when colliding because *usually* it does so and predictive mechanisms generalized this as a prior for future instances. That is, HBPP mechanisms track these causes as highly probable. Understanding that this is not so, requires Type 2 processes to examine the reasons analytically, separating weight, impact and relations with the wind.

Likewise, in the Linda problem, *usually* the women that are engaged in social justice and philosophy are also a feminist, so these instances speak higher than the analytic processes that would need to determine the conjunction effect.

In the Wason selection task the intuitive answer follows the idea that the consequent of an implication might be related to the antecedent. Well in many cases effects are very much related to the causes, Type 2 processing is needed to show that in this case, because of logical principles the antecedent will not help in determining the answer to the question. We do not mean the antecedent fallacy is not a fallacy, only that causes and effects usually have an explanation for why they are together, and that Type 1 processes care more that they always come together than they care about the fact that logically we cannot affirm the consequent in a formal implication.

There seems to be a pattern of responses based on the criteria of focusing on things which usually show up in the world together and there might be probabilistic rules as to how they affect how people will respond, even biasing formal normativity in their responses. This is a hypothesis over the evidence of chapter 1 related to our framework that could be further investigated.

With that covered, we hope to have clarified at least some details about how HBPP relates to the DPT evidence base. We also hope to have offered some discussion on the reference problem and on the interaction of the two processes. But most importantly, at the end of the day, we hope to have proposed a profound and detailed answer to the unity problem of dual process theory.

Appendix I – Internal issues of Dual Process Theories

(Q1) Can dual process theories that come from other evidence bases such as social cognition, learning, language and neuroscience be united with dual process theories of reasoning and decision making?

(Q2) Do the clusters of features really divide perfectly into two groups?

(Q3) If most features are only correlational, which are the defining ones?

(Q4) Are defining features needed for a coherent theory?

(Q5) Do these noticed distinctions point to two minds, systems or types of processing?

(Q6) Is a group of features older in evolution than the other?

(Q7) Do these two processes share the same knowledge base and goal structure?

(Q8) Do they operate in parallel and compete for control of behavior, or do they cooperate, with production of Type 1 default responses that are then assessed and sometimes overridden by Type 2 processes?

(Q9) What processes determine if Type 2 processes are called in or not?

Appendix II – Basic Dual Process Theory features

Type 1	Type 2
Defining Features	
Less or no use of WM	Loads strongly on WM
Autonomous	Can use decoupled representations
Faster in comparison	Slower in comparison
Less effort in comparison	More effort in comparison
Core Correlational Features	
Automatic	Controlled
High Capacity	Low Capacity
Implicit	Explicit
Unconscious	Conscious
Uncorrelated with cognitive ability	Correlated with cognitive ability
Tends towards intuitive responses	Tends towards normative responses
Other Correlational Features	
Parallel	Serial
Contextualized	Abstract
Older in evolution	Recent in evolution

Appendix III – Definitions of Dual Process Features

Working memory: Follows Baddeley and Hitch's (1974) model. Working memory is a short-term storage that holds and processes about seven items at a time. The effort increases as the number of processes it needs to realize increases; too much load will make it malfunction. It is said to have a verbal storage to deal with linguistic representations and a visual-spatial storage to deal with imagery. Later increments to the theory also add an episodic memory buffer.

Decoupled: We can define decoupled representations as those that do not directly follow from (or ignore) context or what is given perceptually.

Autonomy: we can say that Type 1 processing is autonomous because the execution of Type 1 processes is mandatory when their triggering stimuli are encountered, and they are not dependent on input from high-level control systems.

Speed: When solving the same task, a Type 1 process will always work faster than a Type 2 process. This 'faster' should be understood in the sense of coming first to mind.

Effort: Effort is a simple idea that relates to how tiresome using a process is to an individual.

Automatic: Automatic processes are overlearned and nearly always become active in response to a particular input configuration.

Controlled: processes that must work out in real-time (do not invoke previous solutions) to produce output with some novelty for the system at that moment.

Capacity: claiming that Type 2 processes have limited capacity means that they work with a limited amount of inputs from selective attention, they are usually disturbed by competing tasks and are limited by the capacity storage of short-term memory. The contrast is that Type 1 processes have high capacity, but that is no surprise since by Type 1 we are referring to processes executed by a massive collection of systems

Implicit and Explicit: the implicit/explicit distinction refers to the content's format, whether it has an accessible representational format (explicit) or not (implicit).

Consciousness/unconsciousness: refers to the access one has to the content of the environment, of the body and mental processes in a given moment.

Cognitive Ability: Cognitive ability has its origins in Spearman's G (1904) (General Intelligence). It is related to success in learning, problem solving and professional life. Although it is hard to define what it refers to, it is measured with tests and is a generally accepted feature in the psychology community for its predictive success.

Intuitive responses: Responses explained either by Darwinian algorithms, relevance (in the sense of Sperber and Wilson, 1986) or heuristics or responses that follow Type 1 features.

Normative response: responses that respect normative criteria such as logic, math or an optimal probability calculus.

Parallel processes: Processes that occur concomitantly., for instance in a neural network.

Serial processes: Processes that occur step-by-step, one at a time.

Contextualized and abstract: the first refers to processes that are fixed by content and context and the latter are content and context-free.

Old or Recent in evolution: Concerns if processes had their origins further back in the evolutionary timescale or appeared more recent

BIBLIOGRAPHY

ACKERMAN, P.; HEGGESTAD, E. Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, v.121, p.219–245.

ADLER, J. Abstraction is uncooperative. *Journal for the Theory of Social Behavior*, v.14, p.165-181, 1984.

ANDERSON, M. *After phrenology*: Neural reuse and the interactive brain. Cambridge, MA: MIT Press, 2014.

ANDERSON, M.; CHEMERO, A.The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster, *Behavioral and Brain Science*s, v.36, n.3), 2013.

BAARS, B. J. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press, 1988.

BADDELEY, A. Working Memory. *Science*, v.255, 1992.

BADDELEY, A.; HITCH, G. Working Memory. *Psychology of Learning and Motivation*, v.8, p.47–89, 1974.

BALDWIN, M. A New Factor in Evolution. *The American Naturalist*, v.30, n.355, 1896.

BELLINI-LEITE, S. The embodied embedded character of system 1 processing. *Mens Sana Monographs*, v.11, 2013.

BELLINI-LEITE, S. The revisionist strategy in cognitive science. In: Frederick Adams, João Eduardo Kogler Jr., Osvaldo Pessoa Jr. (eds.) Cognitive Science: Recent Advances and Recurring Problems. Proceedings of the 10th Brazilian International Meeting of Cognitive Science. Wilmington: Vernon Press, 2017 (forthcoming).

BETSCH, T. The nature of intuition and its neglect in research on judgment and decision making. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making*. New York: Erlbaum, 2008, p. 3-22.

BODEN, M. GOFAI. In: K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence*: The Frame Problem in Artificial Intelligence. Cambridge: Cambridge University Press, 2014.

BOYD, R. Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, v.61, p.127–48, 1991.

BRIGHTON, H.; GIGERENZER, G. Probabilistic minds, Bayesian brains, and cognitive mechanisms: harmony or dissonance. In: Nick Chater & Mike Oaksford (eds.), *The Probabilistic Mind*: Prospects for Bayesian Cognitive Science. Oxford: Oxford University Press, 2008.

BROOKS, R. *Cambrian intelligence*: the early history of the new AI. NY: Bradford books, 1999.

BROWN, H.; FRISTON, K Dynamic causal modelling of precision and synaptic gain in visual perception - an EEG study. *Neuroimage*, v.63, 2012.

BRUNER, J. The narrative construction of reality. *Critical Inquiry*, v.18, n.1, 1991.

CACIOPPO, J. T., PETTY, R. E., FEINSTEIN, J., & JARVIS, W. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, v.119, p.197–253, 1996.

CARRUTHERS, P. An architecture for dual reasoning. In: Evans, J. & Frankish, K. (Eds.) *In Two Minds:* Dual Process and Beyond, Oxford: Oxford University Press, 2009, p.109-127.

CASSCELLS, W; SCHOENBERGER, A. GRABOYS, T. Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine*, v.299, n.18, 1978.

CHAIKEN, S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, v.39, 1980.

CHALMERS, D. Facing up to the problem of consciousness. *Journal of Consciousness Studies*. v.2, n.3, 1995.

CHALMERS, D. Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 1996.

CHEN, S.; CHAIKEN, S. The heuristic-systematic model in its broader context. In: Chaiken & Trope (Eds.), *Dual Process Theories in Social Psychology*. New York: The Guilford Press, 1999. p. 73–96

CHENG, P.W., & HOLYOAK, K.J. Pragmatic reasoning schemas. *Cognitive Psychology*, v.17, 1985.

CHOMSKY, N. *Reflections on language*. New York: Random House, 1975.

CLARK, A. *Being there*: Putting brain, body and world together again. Cambridge, MA: MIT Press, 1997.

CLARK, A. Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, v.121, n.483), 2012.

CLARK, A.Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science*s, v.36, n.3, 2013a.

CLARK, A. Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Science*s, v.36, n.3, 2013b.

CLARK, A. The many faces of precision. *Frontiers in Psychology*, v.4, n.270, 2013c.

CLARK, A. *Surfing Uncertainty*: Prediction, Action, and the Embodied Mind. Oxford University Press, 2016.

COHEN, L. Can human irrationality be experimentally demonstrated?. *Behavioral and Brain Sciences*, v.4, p.317-370, 1981.

COSMIDES, L. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, v.31, p.187-276, 1989.

COSMIDES, L. *Deduction or Darwinian Algorithms*? An explanation of the "elusive" content effect on the Wason selection task. Doctoral dissertation, Harvard University. University Microfilms #86-02206., 1985.

COSMIDES, L.; TOOBY, J. Origins of domain specificity: The evolution of functional organization. In L. Hirschfeld and S. Gelman (Eds.), *Mapping the Mind*: Domain specificity in cognition and culture. New York: Cambridge University Press, 1994, p. 84-115.

COSMIDES, L.; TOOBY, J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, v.58, n.1, p.1-73, 1996.

COSMIDES, L.; TOOBY, J. No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition*, v.75, p.1-79, 2000.

COSMIDES, L.; TOOBY, J. Unraveling the enigma of human intelligence: Evolutionary Psychology and the multimodular mind. In: R. Sternberg & J. Kaufman (Eds.) *The Evolution of Intelligence*. Hillsdale, NJ: Erlbaum, 2001, p. 145-198.

COSMIDES, L.; TOOBY, J. Evolutionary Psychology: New Perspectives on Cognition and Motivation. *Annual Review of Psychology*, v.64, p.201-229, 2013.

DAHL, G.; Yu, D.; DENG, L.; ACERO, A. "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition". *IEEE Transactions on Audio, Speech, and Signal Processing*, v.20, n.1 , 2012.

DAW, N.; NIV, Y.; & DAYAN, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, v.8, n.12, 2005.

DAW N.; GERSHMAN, S.; SEYMOUR, B.; DAYAN, P.; DOLAN, R. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, v.69, 2011.

DAWKINS, R. *The Selfish Gene*. Oxford: Oxford University Press, 1976.

DAYAN, P.; HINTON, G.; NEAL, R. The Helmholtz machine. *Neural Computation*, v.7, 1995.

DE NEYS, W. Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, v.17, 2006.

DENNETT, D. *Brainstorms*. New York: Bradford Books, 1978.

DENNETT, D. Cognitive Wheels: The Frame Problem of AI. In: Z. Pylyshyn (ed.), *The Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987. p.41-64.

DENNETT, D. *Consciousness explained*. New York: Black Bay Books, 1991.

DREYFUS, H. *What Computers Can't Do*: A Critique of Artificial Reason. New York: Harper & Row Publishers, 1972.

DREYFUS, H & DREYFUS, S. How to Stop Worrying about the Frame Problem Even though It's Computationally Insoluable. In: Z. Pylyshyn (ed.), *The Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987. p.95-111.

EGNER, T.; MONTI, J.; SUMMERFIELD, C. Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, v.30, n.49, 2010.

ELQAYAM, S. & EVANS, J. Subtracting 'ought' from 'is': Descriptivism versus normativism in the study of the human thinking. *Behavioral and Brain Sciences*, v.34, n.5, 2011.

EVANS, J. *Bias in Human Reasoning*: Causes and Consequences. Brighton, UK: Erlbaum, 1989.

EVANS, J. Deciding before you think: Relevance and reasoning in the selection task. British *Journal of Psychology*, v.87, n.2, 1996.

EVANS, J. On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, v.13, n.4, 2007.

EVANS, J. Dual-processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, v.59, 2008.

EVANS, J. How many dual-process theories do we need? One, two, or many? In: Evans, J. & Frankish, K. (Eds.) *In Two Minds*: Dual Process and Beyond, Oxford: Oxford University Press, 2009, p.33-54.

EVANS, J. Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental review*, v.31, 2011.

EVANS, J. Dual-Process Theories of Deductive Reasoning: Facts and Fallacies. In: Holyoak, K. & MORRISON, R. (Eds.) *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press, 2012. p.115-133.

EVANS, J.; CURTIS-HOLMES, J. Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, v.11, p.382–389, 2005.

EVANS, J.; LYNCH, J. Matching bias in the selection task. *British Journal of Psychology*, v.64, p.391-397, 1973.

EVANS, J.; OVER, D. *Rationality and Reasoning*. East Sussex: Psychology Press, 1996.

EVANS, J.; STANOVICH, K. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, v.8, n.3, 2013a.

EVANS, J.; STANOVICH, K. Theory and Metatheory in the Study of Dual Processing: Reply to Comments. *Perspectives on Psychological Science*, v.8, n.3, 2013b.

ERAÑA, A. Dual process theories versus massive modularity hypotheses. *Philosophical Psychology*, v.25, n.6, 2012.

ERNST, M.; BANKS, M. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, v.415, 2002.

FIEDLER, K. The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, v.50, p. 123-129, 1988.

FITZGERALD, T.; MORAN, R.; FRISTON, K.; DOLAN, R. Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation. *Neuroimage*, v.107, 2015.

FODOR, J. *The Langue of Thought*. New York: Crowell, 1975.

FODOR, J. *The Modularity of Mind*. Cambridge: MIT Press, 1983.

FODOR, J. Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres. In: Z. Pylyshyn (ed.), *The Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987. p.139-149.

FODOR, J. *The Mind Doesn't Work That Way*. Cambridge: MIT Press, 2001.

FODOR, J; PYLYSHYN, Z. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition,* v.28, 1988.

FODOR, J.; PIATTELLI-PALMARINI, M. *What Darwin got wrong*. New York: Ferrar, Straus, Giroux, 2010.

FRANKISH, K. *Supermind and Supramind*. Cambridge: Cambridge University Press, 2004.

FRANKISH, K. Systems and levels: Dual-system theories and the personal–subpersonal distinction. In: Evans, J. & Frankish, K. (Eds.) *In Two Minds*: Dual Process and Beyond, Oxford: Oxford University Press, 2009, p.89-107.

FREDERICK, S. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, v.19, n.4, 2005.

FRISTON, K. Learning and inference in the brain. *Neural Networks*, v.16, n.9, 2003.

FRISTON, K. A theory of cortical responses. *Philosophical Transaction of the Royal Society of London B: Biological Sciences*, v.360, n.1456, 2005.

FRISTON, K. Hierarchical models in the brain. *PLoS Computational Biology*, v.4, n.11, 2008.

FRISTON, K. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, v.11 n.2, 2010.

FRISTON, K.; LAWSON, R.; FRITH, C. On hyperpriors and hypopriors: Comment on Pellicano and Burr. *Trends in Cognitive. Sciences*, v.17, n.1, 2013.

FRISTON, K.; STEPHAN, K. Free energy and the brain. *Synthese*, v.159, n.3, 2007.

GIBSON, J. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

GIGERENZER, G. How to make cognitive illusions disappear: beyond heuristics and biases. *European Review of Social Psychology*, v.2, p.83-115, 1991.

GRIGGS R.; COX, J. The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, v. 73, p.407–420, 1982.

GRUSH, R. The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, v.27, 2004.

GOEL, V. Cognitive neuroscience of deductive reasoning. In: Holyoak, K. and Morrison, R., (eds.) *Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 2005. p. 475–492

GOEL, V. Anatomy of deductive reasoning. *Trends in Cognitive Sciences* v.11, p.435–441, 2007.

GOULD, S. *Bully for the Brontosaurus*. New York: Norton, 1991.

GOWATY, P.; HUBBELL, S. Bayesian animals sense ecological constraints to predict fitness and organize individually flexible reproductive decisions. *Behavioral and Brain Science*s, v.36, n.3, 2013.

HAUGELAND, J. An Overview of the Frame Problem. In: Z. Pylyshyn (ed.), *The Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987. p.77-93.

HARMAN, G. *Change in View*: Principles of Reasoning. Cambridge: MIT Press, 1986.

HAYES, P. What the Frame Problem Is and Isn't, In: Z. Pylyshyn (ed.), *The Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987. p.123-137.

HIRSH, J.; MAR, R.; PETERSON, J. Personal narratives as the highest level of cognitive integration. *Behavioral and Brain Sciences*, v.36, 2013.

HOHWY, J. *The predictive mind*. Oxford: Oxford University Press, 2013.

HORN, L. *A natural history of negation*. Chicago: Chicago University Press, 1989.

HOSOYA, T., BACCUS, S. A. & MEISTER, M. Dynamic predictive coding by the retina. *Nature* v.436, n7, 2005.

JANLERT, L. Modeling Change–The Frame Problem of AI. In: Z. Pylyshyn (ed.), T*he Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987. p.1-40.

JOHNSON, M.; MORTON, J. *Biology and cognitive development*: The case of face recognition. Oxford: Blackwell, 1991.

KAHNEMAN, D. *Maps of bounded rationality*: a perspective on intuitive judgment and choice. Nobel prize lecture, 2002. Online at: http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahnemann-lecture.pdf

KAHNEMAN, D. *Thinking, Fast and Slow*. New York: Farrar, Strauss, Giroux; 2011.

KAHNEMAN, D.; FREDERICK, S. Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases*: The psychology of intuitive judgment. Cambridge, UK: Cambridge University Press, 2002. p. 49–81.

KAHNEMAN, D., FREDERICK, S. Frames and brains: Elicitation and control of response tendencies. *Trends in Cognitive Science*, n.11, 2007.

KAHNEMAN, D.; SLOVIC, P.; TVERSKY A. *Judgment under Uncertainty*: Heuristics and Biases. New York: Cambridge University Press, 1982.

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

KEREN, G., & SCHUL, Y. Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, 4, 533–550, 2009.

KNILL, D.; POUGET, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, v.27, n.12, 2004.

KRUGLANSKI, A. W.; GIGERENZER, G. Intuitive and deliberative judgments are based on common principles. *Psychological Review,* v.118, p.97–109, 2011.

KRUGLANSKI, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, v.8, p.242–247, 2013.

LAKATOS, I. *The Methodology of Scientific Research Programmes*: Philosophical Papers Volume I. Cambridge: Cambridge University Press, 1978.

LESLIE, A. Pretense and representation: the origins of ‘‘theory of mind’’. *Psychological Review*, v. 94, p.412–426, 1987.

LIEBERMAN, M. Reflective and reflexive judgment processes: a social cognitive neuroscience approach. In: Forgas J., Williams K., von Hippel W., (Eds). *Social Judgments*: Implicit and Explicit Processes. New York: Cambridge University Press, 2003, p. 44–67.

LOTTER, W.; KREIMAN, G.; COX, D. Deep predictive coding networks for video prediction and unsupervised learning. arXiv:1605.08104, 2016.

LUGER, G. Artificial intelligence: structures and strategies for complex problem solving. New York: Pearson, 2009.

LUPYAN, G. Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, 2015.

MARKOV, N.; VEZOLI, J.; CHAMEAU, P.; FALCHIER, A.; QUILODRAN, R.; HUISSOUD, C.; LAMY, C.; MISERY, P.; GIROUD, P.; ULLMAN, S.; BARONE, P.; DEHAY, C.; KNOBLAUCH, K.; KENNEDY, H. Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *The Journal of Comparative Neurology*, v.522, n.1, 2014.

MARR, D. *Vision*: A Computational Investigation into the Human Representation and Processing of Visual Information. Cambridge: MIT Press, 1982.

MCCARTHY, J.; HAYES, P. Some philosophical problems from the stand point of artificial intelligence. In: B. Meltzer & D. Michie (Eds.), *Machine Intelligence 4*. Edinburgh, Scotland: Edinburg University Press, 1969.

MCCLELLAND, J. Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, v.4, 2013.

MERCIER, H. & SPERBER, D. Intuitive and reflective inferences. In Evans, J. and Frankish, K. (Eds.) *In two minds*: Dual processes and beyond. Oxford University Press, 2009, p.149-170.

MILLER, G. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, v.63, p,81- 97, 1956.

MILLER, G. *Psychology*: The science of mental life. New York: Harper & Row, 1962.

NEISSER, U. *Cognitive Psychology*. New York: Appleton Century-Crofts, 1967.

NEWELL, A. Physical symbol systems. *Cognitive Science* 4, 1980.

NEWELL, A.; SIMON, H. GPS: A Program that Simulates Human Thought. In: E. Feigenbaum & J. Feldman (Eds.) *Computers & Thought.* New York: McGraw-Hill Book Company, 1963. p.279-293.

NEWELL, A.; SIMON, H. Computer Science as Empirical Inquiry: Symbols and Search. *The ACM Communications*, v.19, n.3, 1976.

NICHOLS, S.; STICH, S. *Mindreading*: an integrated account of pretence, self-awareness, and understanding other minds. Oxford University Press, Oxford, 2003.

NISBETT, R.; PENG, K.; CHOI, I., NORENZAYAN, A. Culture and systems of thought: holistic vs. analytic cognition. *Psychology Review*, v.108, 2001.

NISBETT, R.; WILSON, T. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, v.84, n. 3, 1977.

NOË, A. *Action in perception*. Cambridge: MIT press, 2004.

OSMAN, M. An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, v.11, p.988–1010, 2004.

OSMAN, M. A Case Study: Dual-Process Theories of Higher Cognition--Commentary on Evans & Stanovich (2013), *Perspectives on Psychological Science*, v.8, n.248, 2013.

QUINE, W. V. *Ontological Relativity and Other Essays*. Ch. 5, Columbia University Press, 1969.

PARK, C.; LIM, J.; CHOI, H.; KIM, D. Predictive coding strategies for developmental neurorobotics. *Frontiers in Psychology*, v.7, n.3, 2012

PIATTELLI-PALMARINI, M. *Inevitable Illusions*: How Mistakes of Reason Rule our Minds. New York: John Wiley & Sons, 1994.

PICCININI, G. *Physical computation*: a mechanistic account. Oxford: Oxford University Press, 2015.

PINKER, S. *How the mind works*. New York: W. W. Norton & Company, 1997.

PINKER, S. So how does the mind work?. *Mind and Language*, v.20, n.1, 2005.

PHILLIPS, W.; CLARK, A.; SILVERSTEIN, S. On the functions, mechanisms, and malfunctions of intracortical contextual modulation. *Neuroscience and Biobehavioral Reviews*, v.52, 2015.

PYLYSHYN, Z. (ed.) *The Robot's Dilemma*: The Frame Problem in Artificial Intelligence. New Jersey: Ablex Publishing Corporation, 1987.

REBER, A. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, v.6, p.317-327, 1967.

REBER, A. *Implicit Learning and Tacit Knowledge: an essay on the cognitive unconsciousness*. Oxford, UK: Oxford University Press, 1993.

RANSOM, M; FAZELPOUR, S.; MOLE, C. Attention in the predictive mind. *Consciousness and Cognition*, 2016.

RAO, R.; BALLARD, D. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, v.2, n.1, 1999.

ROBERTS, M.; NEWTON, E. Inspection times, the change task, and the rapid response selection task. *Quarterly Journal of Experimental Psychology Section A*: Human Experimental Psychology, v.54, p.1031–1048, 2001.

SALAKHUTDINOV, R.; TENENBAUM, J.; TORRALBA, A. Learning with hierarchial-deep models. *IEEE transactions on pattern analysis and machine intelligence*, v.35, n8, 2013.

SAMUELS, R. The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In: Evans, J. & Frankish, K. (Eds.) *In Two Minds*: Dual Process and Beyond, Oxford: Oxford University Press, 2009, p.129-146.

SAMUELS, R.; STICH, S.; BISHOP, M. Ending the rationality wars: How to make disputes about human rationality disappear. In: E.Renee (ed.), *Common sense, reasoning and rationality*. New York: Oxford University Press, 2002. p.236-268.

SAMUELS, R.; STICH, FAUCHER. L. Reason and Rationality. In: I. Niiniluoto, M. Sintonen, & J. Wolenski (eds.). *Handbook of Epistemology*. Dordrecht: Kluwer, 2004. p.1–50.

SCHACTER, D.; TULVING, E. *Memory Systems*. NY: Bradford books, 1994.

SCHNEIDER, W.; CHEIN, J. Controlled and automatic processing: Behavior, theory, and biological processing. *Cognitive Science*, v.27, p.525–559, 2003.

SETH, A. A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience*, v.5, n.2, 2014.

SHANNON, C.; WEAVER, W. The Mathematical Theory of Communication. Urbana: University of Illinois Press, 1949.

SHIFFRIN, R; SCHNEIDER, W. Controlled and automatic human information processing I: detection, search and attention. *Psychological Review*, v.84, p.1–66, 1977.

SIMON, D.; CHABRIS, C. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, v.28, 1999.

SLOMAN, A. What else can brains do? *Behavioral and Brain Science*s, v.36, n.3, 2013.

SLOMAN, S. The Empirical Case For Two Systems of Reasoning. *Psychological Bulletin*, v.119, 1996.

SOKOLOV, E. Neuronal models and the orienting reflex. In: M. A. B. Brazier (Ed.), *The central nervous system and behavior* (pp. 187– 276). New York: Josiah Macy, Jr. Foundation, 1960.

SPERBER, D., CARA, F.; GIROTTO, V. Relevance theory explains the selection task. *Cognition*, v.57, p.31-95, 1995.

SPERBER, D.; WILSON, D. *Relevance:* Communication & Cognition. Oxford: Blackwell, 1986.

STANOVICH, K. Concepts in developmental theories of reading skill: cognitive resources, automaticity, and modularity. *Developmental Review*, v.10, p.72–100, 1990.

STANOVICH, K. *Who is Rational*? Studies of Individual Differences in Reasoning. Mahwah, N.J.: Lawrence Erlbaum Associates, Inc., 1999.

STANOVICH, K. *The Robot's Rebellion*: Finding meaning in the age of Darwin. Chicago: University of Chicago Press, 2004.

STANOVICH, K. Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In: Evans, J. & Frankish, K. (Eds.) *In Two Minds*: Dual Process and Beyond, Oxford: Oxford University Press, p.55-88, 2009.

STANOVICH, K. *Rationality and The Reflective Mind*. Oxford: Oxford University Press, 2010.

STANOVICH, K; WEST, R. Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, v.4, p.193–230, 1998a.

STANOVICH, K; WEST, R. Individual differences in framing and conjunction effects. *Thinking and Reasoning*, v.4, p.289–317, 1998b.

STANOVICH, K; WEST, R. Individual differences in rational thought. J*ournal of Experimental Psychology*: General, v.127, p.161–88, 1998c.

STANOVICH, K; WEST, R. Who uses base rates and P(D/~H)? An analysis of individual differences. *Memory & Cognition*, v.28, p.161–79, 1998d.

STANOVICH, K.; WEST. R. Individual Differences in Reasoning: Implications for the Rationality Debate?. *Behavioural and Brain Sciences*, v.23, n.5, p.645–665, 2000.

STANOVICH, K.; WEST, R. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, v. 94, p.672–95, 2008.

STANOVICH, K. E.; TOPLAK, M. E. Defining features versus incidental correlates of type 1 and type 2 processing. *Mind & Society*, v.11, p.3–13, 2012.

STEIN, E. *Without good reason*: The rationality debate in philosophy and cognitive science. Oxford: Oxford University Press, 1996.

THORNDIKE, E. *Animal Intelligence; Experimental Studies*. New York: The Macmillan Company, 1911.

TOLMAN, E. Cognitive maps in rats and men. *Psychological Review*, v.55, n.4, 1948.

TOMASELLO, M. Culture and Cognitive Development. *Current directions in psychological Science*, 2000.

TOPLAK, M.; WEST, R.; STANOVICH, K. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition*, v.39, 2011.

TOPLAK, M.; WEST, R. STANOVICH, K. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, v.20, n.2, 2014.

TOOBY, J; COSMIDES, L. The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environments. *Ethology and Sociobiology*, v.11, p.375-424, 1990.

TURING, A. On computable numbers with an application to the entscheidungs problem. *Proceedings of the London Mathematical Society* 2, v. 42, n.1, 1936.

TVERSKY, A. Intransitivity of preferences. *Psychological Review*, v. 76, n.1, 1969.

VON NEUMANN, J. The First Draft Report on the EDVAC. *IEEE Annals of the History of Computing*, v.15, n.4, [1993] 1945.

WASON, P. The processing of negative and positive information. *Quarterly Journal of Experimental Psychology*, v.11, p.92-107, 1959.

WASON, P. Reasoning. In Foss, B. (Ed.), *New Horizons in Psychology*. Harmondsworth: Penguin Books, 1966. p.135–151.

WEISS, Y.; SIMONCELLI, E.; ADELSON, E. Motion illusions as optimal percepts. *Nature neuroscience,* v.5 n.6, 2002.

WHEELER, M. Cognition in Context: Phenomenology, Situated Robotics and the Frame Problem. *International Journal of Philosophical Studies* 16; v.3, p.323-49, 2008.

YAMASHITA Y.; TANI, J. Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Computational Biology*, v.4, n.11, 2008.