THE RELATIONSHIP BETWEEN RATIONALITY AND REASONING IN RATIONAL CHOICE AND BEHAVIOURAL ECONOMICS

Thesis submitted at the University of East Anglia
for the degree of Doctor of Philosophy

by

Antonios Staras
School of Economics
University of East Anglia

June 2019

Abstract

In normative economics and behavioural social sciences, rationality is described as a set of rationality axioms on preferences. A common explanatory strategy is to attribute deviations from standard decision theory axioms to 'reasoning errors' using the dual-system model. The main idea is that reasoning errors are often sourced in the fast and automatic System 1 and that the logical System 2 has the capacity to correct these errors. Very little effort has been made to explain what this logical reasoning is, what are its 'limitations', and how it leads to preferences that satisfy these axioms.

My thesis explores the relevance of John Broome's (2013) philosophical argument for normative and behavioural economics that the notion of 'reasoning' is separate from that of 'rationality'. I propose a simple 'Broomean' model of reasoning as a conscious, explicit, and rule-guided mental process – what cognitive scientists and behavioural economists call System 2 – and investigate the extent to which rational preferences can or cannot be reached by this type of reasoning.

Chapters 1 and 2 develop the formal framework that allows us to capture and disentangle the notions of reasoning and rationality. Chapter 2 concludes that reasoning is successful in achieving some but not all requirements of the theory. One implication of this is that automatic processes jump in where reasoning fails to lead to rational preferences. Chapter 3 uses this framework to discuss general problems and famous paradoxes to expected utility theory in decision theory and behavioural economics.

Contents

List of Tables

Acknowledgements

First and foremost, I would like to thank Robert Sugden and Franz Dietrich. I have been extremely lucky to have two outstanding supervisors who cared so much about my research project, who continuously inspired me and listened to my ideas so patiently, and who I look up to as researchers and as persons. Without their support, I would have never been able to develop and finish this thesis. Thank you Bob for all you have done and are still doing for me, and for sharing with me your passion for research. Thank you Franz for always having tried so patiently to understand my incomprehensible and messy notes, and for never giving up on me.

I have greatly enjoyed my time at the University of East Anglia. One of the best things about it is the very good environment of our PhD office. Special thanks to Anwesha, Balazs, Emike, Mengjie, Paul, Ritchie, and Yannis. I am also immensely grateful to the PhD community at the University of Paris 1 that made my research visit there so enjoyable. I would also like to thank my housemates, Marco, Julia, and Sara for being there when I needed them the most. I wish to acknowledge that this thesis would not be possible without the full studentship support I received from the Faculty of Social Sciences.

For obvious reasons, I would like to thank my family. I could simply not pursue this PhD without the constant support of my lonely sweetheart and the confidence and encouragement of my amazingly understanding mother. Πατέρα σε ευχαριστώ για όλα!

Introduction

Can rational choice be reached by reasoning? This thesis is about answering this question. It is a question I wanted to ask after reading John Broome's answer (Broome, 2007) to Kolodny's article (Kolodny, 2005) on the nature of 'rationality requirements'. Then I read Broome's book, *"Rationality through reasoning"* (2013), in which he argues that, theories of rationality such as rational choice theory can be described by requirements of rationality (e.g. transitivity of preferences), and often neglect to explain the 'reasoning' by which one comes to satisfy these requirements. In his words,

> "[they] seem to think they have finished their job when they have described the requirements of rationality. [...] I think these authors must believe that, once you know what requirements there are, that knowledge directly supplies you with premises you can use in active reasoning. They must believe that, starting from knowledge of a particular requirement, you can reason your way actively to satisfying that requirement" (2013, pp. 208-9)

So according to Broome, knowledge of the requirements does not provide the self-help tools necessary to become rational if one is not already. This argument rests on a careful distinction between 'requirements of rationality' and 'rules of reasoning'. Broome's informal analysis of rationality and reasoning is very well received among philosophers but is almost unknown to economists. So the first two chapters set up the modelling framework which can represent (i) an agent's 'mental states', (ii) rationality requirements on an agent's set of mental states, and (iii) rules of reasoning by which an agent can create new mental states, given existing ones.

The first chapter describes conventional rational choice as a theory of rationality requirements on mental states while the second chapter formalises reasoning and addresses the main question of this thesis: whether and in what ways one can come to satisfy rational choice requirements by reasoning. Then it relates the formal analysis to existing literature, particularly 'System 1/ System 2' models in psychology and behavioural economics. The third chapter uses this framework to reconstruct

how Savage (1954, pp. 101-3) resolved his personal problem of discovering that his preferences over Allais's gambles violated the sure-thing principle which is an implication of his own axioms of expected utility theory.

Following are detailed descriptions of each chapter.

0.1 Chapter 1

The framework starts with the fundamental notion of mental states as attitudes such as beliefs or intentions towards particular propositions, and from it builds a notion of rationality as requirements on an agent's set of mental states. Different theories of rationality recognise some specifications of attitudes and propositions. I represent Broome's theory of rationality and conventional rational choice as two theories of requirements, the first from philosophy and the second from economics.

Following Broome, with mental states I represent the agent's internal language relevant to practical reasoning. A person will at least have the attitude of belief and of intention towards propositions when engaging with practical reasoning, such as the belief that it is sunny or the intention that I go sailing.

For choice theory, I consider a non-empty set X of mutually exclusive choice options such as goods or political candidates, and the attitudes of preference and indifference towards them. In each choice context, certain options from X are feasible. Conventional choice theory implicitly assumes the feasible set to be known. I enrich the theory in our framework with beliefs about what the feasibility set is and intentions prior to any choice. Following Broome, choices are not mental states, but are rather caused by intentions.

Eight requirements make together my 'Broomean' theory of rational choice. Of them all, Economic Enkrasia is the most 'Broomean' one; the choice-theoretic counterpart of Broome's ordinary Enkrasia. It differs from ordinary Enkrasia in that intentions respond to preferences and feasibility beliefs rather than ought-beliefs.

I classify rationality requirements in a taxonomy of requirements typically found in choice theoretic axiomatisations. The taxonomy consists of four types of requirement: i) Completeness requirements (e.g. completeness of preferences), ii) Consistency requirements (e.g. non-contradiction of preferences), iii) Closedness requirements (e.g. transitivity), and iv) Negative Closedness requirements (e.g. negative transitivity).

I show that, except for a particular case that, if a requirement is of one type then it is not of any other type, and that every possible single requirement with two or three mental states is a conjunction of requirements from the taxonomy. Then

I define a single type of requirement (all types of requirement considered above are special cases of this type) and prove that every possible single requirement is a conjunction of generalised requirements of this type with a finite number of mental states. This classification will help us in the next chapter to answer Broome's main question which is at the core of this Thesis; whether reasoning helps to achieve any type of requirement.

0.2 Chapter 2

Chapter 2 unpacks Broome's account of reasoning in a simple model of reasoning for rational choice.[1] This account explores the relations of consequence that hold between attitudes (e.g. beliefs, intentions), and not between propositions or sentences as usual (2013, p. 254); an idea that captures the intuitive notion of everyday human reasoning as a mental activity conducted in a language through which you give rise to new attitudes from existing ones following 'reasoning rules'. These reasoning rules are restrictive in two ways. First, they create rather than remove attitudes; for instance, no rule removes the preference for x to y given the preference for y to x. Second, new attitudes follow from the presences of attitudes rather than the absences of attitudes; for instance, no rule forms a preference based on the absence of other preferences.

Drawing on the classification of the formal structure of requirements developed in Chapter 1, I show that, rule-following reasoning (i) is always capable of achieving Closedness requirements, (ii) cannot achieve Consistency requirements, (iii) can achieve Completeness or Negative Closedness requirements, but usually only at the cost of creating inconsistencies; and so give a partially negative answer to the main question of the thesis. I study some basic requirements of rational choice; completeness and transitivity of weak preferences and Economic Enkrasia, and show how far one can go when reasoning to satisfy them. And with it, how far our negative answer reveals gaps in Broome's theory, or deficiencies in choice theory and behavioural economics.

I relate the implications of the results derived from this model to existing literature, particularly 'System 1/ System 2' models in psychology and behavioural economics. The central idea is that judgment and choice is an interplay between the automatic, non-verbal, and associative "System 1" and the conscious, explicit, and rule-guided "System 2". System 1 generates impressions, intuitions, feelings,

[1] The thesis reports my own work. In the course of my PhD I had the opportunity to discuss extensively my research with my supervisors. Our paper Dietrich et al. (2019) has been the product of these discussions.

and impulses, and System 2, operating on the outputs of System 1, forms explicit beliefs, intentions and preferences (Wason and Evans, 1974 and Kahneman, 2011).

A growing number of behavioural economic models have incorporated the dual-system hypothesis to model choice. I identify three prominent approaches. In the first approach, choice is the result of two types of conflicting selves, the System 1 'irrational' self and the System 2 'rational' self that solve a maximisation problem against each other. In the second approach, the two selves are allies that work together to solve the maximisation problem. The third approach is to understand them as complementary systems contributing towards the mental process. I argue that this model of reasoning can be understood as an explicit description of System 2 reasoning and that the negative results make a case against the first two approaches. The two systems complement each other in the sense that they are able to process different types of information. System 2 processes mental states which are attitudes towards propositions while System 1 processes mental states which are non-propositional. Using the model, some examples of System 1 and 2 mental states and their interaction are presented.

0.3 Chapter 3

A negative conclusion of Chapter 2 is that explicit reasoning cannot achieve consistency requirements. In plain English, this result says that, in reasoning explicitly you cannot conclude in the *absence* of an attitude, a thesis that Broome agrees with (2013, p. 278). Broome's own interpretation is that when explicit reasoning fails 'automatic processes will normally prevent you from having contradictory beliefs' to achieve consistency (2013, p. 278). But there is not much discussion what to make of this. Chapter 3 gives a possible explanation. Using the formal analysis in the first two chapters, I reconstruct how Savage resolved his personal problem of discovering that his preferences over Allais's gambles violated the Sure-thing principle which is a basic requirement in Savage's framework.

Chapter 1

Rational choice as a theory of rationality requirements

1.1 Introduction

According to John Broome (2013), theories of rationality such as rational choice can be described by 'requirements of rationality', e.g. transitivity of strict preferences. Requirements of this sort require something on a person's set of mental states: her beliefs, her preferences, and so on to be properly related to each other. Theories of rationality that are solely described by requirements of rationality neglect to explain the 'reasoning part' by which one can come to satisfy these requirements (2013, pp. 208-9). A person acquires or drops a particular mental state through active reasoning, which is a rule-following conscious mental act that applies directly on the contents of a person's mental states (2013, p. 153). This argument rests on a careful distinction between 'requirements of rationality' and 'reasoning'. Others, particularly (Kolodny, 2005, 2007), have questioned whether this is the right way to think about requirements and reasoning.

Many rationality requirements are or can be expressed as an *'If ... then ...'* statement whose consequent is a proposition about a single mental state and the antecedent is a proposition about a set of mental states, while rules of reasoning describe, in a quite similar form, a way in which a conclusion follows from a set of premises. For example, transitivity of strict preferences states that, for all options x, y, z, *if* x is strictly preferred to y and y is strictly preferred to z, *then* x is strictly preferred to z. Analogously, the reasoning towards my conclusion, 'I prefer x to z', follows from my premises, 'I prefer x to y and I prefer y to z'. Because of the structural analogy between requirements expressed as conditional statements and reasoning rules, philosophers and economists often think that requirements apply

on processes. For example, Kolodny has defended the view that requirements require you to *do* something, that requirements apply on processes (Kolodny, 2005, 2007)[1]. And Savage has famously paralleled his axioms of expected utility theory with the principles of logic, suggesting that they can be used as reasoning templates: "Pursuing the analogy with logic, [he says] the main use I would make of P1 [completeness and transitivity] and its successors is normative, to police my own decisions for consistency and, where possible, to make complicated decisions depend on simpler ones" (Savage, 1954, p. 20). There is room for debate about whether rationality requirements 'really' apply on processes or not, and hence about whether Broome's rigid separation of requirements and reasoning reflects the best way of thinking about how a person can come to satisfy rationality requirements. But there is no doubt that many theorists describe rational choice as a set of axioms on preferences. To this end, this chapter analyses the formal structure of requirements – and indeed of any rationality requirement – setting aside reasoning.

I identify four requirement types that have an *'If ... then ...'* structure typically found in choice theoretic axiomatisations: i) completeness requirements (e.g. completeness of preferences), ii) consistency requirements (e.g. non-contradiction of preferences), iii) closedness requirements (e.g. transitivity), and iv) negative closedness requirements (e.g. negative transitivity). I show that, except for a particular case, no requirement is of two types. I then define a single type of requirement (all types of requirement considered above are special cases of this type) and prove that every possible single requirement is a conjunction of generalised requirements of this type. The proposed taxonomy of requirements is essential for answering in Chapter 2 whether reasoning can bring an agent to satisfy each of the theory's requirements; the main question of this thesis.

The second part of the chapter is less formal. It provides examples of the use of the taxonomy in theories of rationality. Particularly, I use rational choice as a main example of a theory of rationality requirements. This version of rational choice is inspired by Broome's philosophical analysis of mental states as types of attitudes such as beliefs or intentions towards particular propositions. I use mental states to represent the agent's internal language relevant to practical reasoning. A person will at least have the attitude of belief and of intention towards propositions

[1]According to Broome, rationality requirements have wide scope: for example, an enkrasia requirement takes the form 'Rationality requires of you that if you believe you ought to do F then you intend to do F'. According to Kolodny, rationality requirements have narrow scope: for example, an enkrasia requirement takes the form 'If you believe you ought to do F, rationality requires of you that you intend to do F'. Naturally, Broome is primarily concerned with synchronic requirements: requirements on attitudes held simultaneously, while Kolodny is primarily concerned with process requirements: requirements that require you to do something over time.

when engaging with practical reasoning, such as the belief that it is sunny or the intention that I go sailing. For rational choice theory, I consider a non-empty set X of mutually exclusive choice options, goods or political candidates and so on, and the attitudes of preference and indifference towards them. In each choice context, certain options from X are feasible. I enrich the theory with beliefs about what the feasible set is and intentions prior to any choice as choices are not mental states but are rather caused by intentions. Conventional choice theory implicitly assumes the feasible set to be known and describes choices as choice functions that assign the objects chosen from the feasible set. So this new framework partly recasts a simple version of rational choice in terms of an agent's set of mental states; her preferences, feasibility beliefs, and intentions. The reason for doing so is to explicitly include more of the agent's mental states (her beliefs, intentions, preferences, ...) that are implicitly involved when considering requirements of rational choice.

The chapter proceeds in the usual order: Section 2 gives the abstract definitions of the framework. Section 3 presents the taxonomy of requirements. Section 4 provides examples of the use of the taxonomy in theories of rationality. Section 5 concludes. Proofs omitted from the main text are in an appendix.

1.2 Formal concepts and definitions

In this framework an agent operates *in* a *theory of rationality* in which:

1. the agent has mental states which are attitudes towards particular propositions.

2. the agent can create new mental states by following rules of reasoning.

3. there are combinations of mental states which are allowed by requirements of rationality.

Formally, the quadruple $(\mathcal{L}, \mathcal{A}, \mathcal{T}, \mathcal{S})$ denotes the agent's environment where $\mathcal{L}$ contains the possible objects of attitudes, $\mathcal{A}$ the possible attitude types such as a belief, an intention, a preference, $\mathcal{S}$ captures how the agent can create mental states from existing ones, and $\mathcal{T}$ captures which combinations of mental states are rationally allowed, i.e. the theory's requirement. $\mathcal{T}$ captures an external notion of rationality, i.e. that of being rational which depends on the requirements of a theory of rationality, whereas $\mathcal{S}$ captures an internal notion of rationality, i.e. that of becoming rational that depends on the agent's perception. So tuple $(\mathcal{L}, \mathcal{A}, \mathcal{T}, \mathcal{S})$ captures a

view of rationality according which rationality is i) non universal: different theories of rationality recognise some combinations of attitudes and propositions, and ii) independent of the agent's perception.

1.2.1 Mental states

A "mental state" is the fundamental notion in this model. I build this notion considering a non-empty set $\mathcal{L}$ of **propositions** and a non-empty finite set $\mathcal{A}$ of **attitude types**. Any attitude type $a \in \mathcal{A}$ comes with a number of places n_a in $\mathbb{N}^+ = \{1, 2, ...\}$, and a domain of propositions $D_a \subseteq \mathcal{L}$. I call an attitude type towards particular proposition or propositions a propositional attitude, more simply, an attitude or a mental state.

Definition 1.1. A **mental state** is any tuple $(p_1, p_2, ..., p_k, a)$ that satisfies the following properties:

1. a is an attitude type in $\mathcal{A}$ and

2. $p_1, p_2, ..., p_k$ are propositions in D_a, where $k = n_a$.

Any such tuple $(p_1, p_2, ..., p_k, a)$ is an attitude type a towards $p_1, p_2, ..., p_k$. The number of places and the domain tell us that the attitude type applies to combinations $p_1, p_2, ..., p_k$ of k propositions that belong to D_a. For example, (I bike, I walk, preference) is a mental state that consists of a preference attitude that applies to pairs of propositions that belong to the relevant domain D_a of preference.

A remark about notation: Typically, I will use labels such as $m, m_1, m', ...$ to denote mental states; and labels such as $M, M_1, M', ...$ to denote any set of mental states that are *logically possible*. I will write $\mathcal{M} = \{(p_1, p_2, ..., p_k, a) : a \in \mathcal{A}, k = n_a, \text{ and } p_1, p_2, ..., p_k \in D_a\}$ throughout to denote the set of all *logically possible* mental states.

Moreover, I will use the term "mental constitution" or "constitution" to refer to a particular subset of $\mathcal{M}$ that describes the agent's psychology as a combination of her mental states held at a single time. Formally:

Definition 1.2. A **constitution** is a set of mental states $C \subseteq \mathcal{M}$.

An example is the constitution $C = \{(p, p', a), (p', p'', a)\}$. Throughout, I will write $\mathcal{C} = 2^{\mathcal{M}}$ to denote the set of all constitutions. Constitutions in $\mathcal{C}$ will typically be denoted by labels such as $C_0, C_1, C', ...$ I will say that the pair $(\mathcal{L}, \mathcal{A})$ denotes your "mental structure".

1.2.2 Requirements

Given a mental structure $(\mathcal{L}, \mathcal{A})$, I can define the notion of a "requirement". Requirements *allow* some constitutions from $(\mathcal{L}, \mathcal{A})$. Formally:

Definition 1.3. A **requirement** is a set of constitutions $R \subseteq \mathcal{C}$. A constitution C is said to:

1. EITHER *satisfy* R if $C \in R$.
2. OR *violate* R if $C \notin R$.

It is useful to say that, an implication of the definition of a requirement as a set R of allowed constitutions is that a requirement rules out some set of constitutions, i.e. the complement of the allowed constitutions $R' = \mathcal{M} \setminus C$. I now give an example of a requirement, more precisely of a requirement schema, or more simply of a schema.

Transitivity of strict preferences: For any $x, y, z \in X$,

if $(x, y, \succ) \in C$ and $(y, z, \succ) \in C$, then $(x, z, \succ) \in C$.

Requirements are requirement schemas since they involve parameters. In the example, $\succ$ is a two-place attitude in $\mathcal{A}$ interpreted as a strict preference and propositions x, y ,$z \in \mathcal{L}$ are the parameters of the requirement. A constitution C satisfies transitivity if C satisfies all instances of the requirement schema above.

Remark 1.1. By inspection of Definition 1.3, one observes that:

1. if $R = \mathcal{C}$, then the requirement is satisfied by all constitutions and is the **tautological requirement**.
2. if $R = \emptyset$, then the requirement is satisfied by no constitution and is the **contradictory requirement**.

1.3 A taxonomy of requirements

Let a **type of requirement** be a set $\mathcal{R} = \{R_1, R_2, ...\}$ of all the requirements that belong to the same type and a **taxonomy of requirements** be a set T of types of requirement $\mathcal{R}$. Since I am interested in analysing the formal structure of requirements, and in particular of requirements written as *'If ... then ...'* statements, it is convenient to introduce the following notation. I use m^+ to describe a proposition about the presence of a mental state, e.g. $m \in C$; and m^- to describe a proposition about the absence of a mental state, e.g. $m \notin C$. When I consider sets of mental states $M = \{m, n, \ldots\}$ I use:

(i) c^+ for a conjunction of presences m^+, n^+, ... for M,

(ii) c^- for a conjunction of absences m^-, n^-, ... for M,

(iii) d^+ for a disjunction of presences m^+, n^+, ... for M,

(iv) and d^- for a disjunction of absences m^-, n^-, ... for M.

I can now ask what are the different types of *'If ... then ...'* statements whose consequent is a proposition about a single mental state and the antecedent is a conjunction of mental states. I identify four requirement types.

(i) A completeness requirement is of the form 'if c^- then m^+'. For example, if $(x, y, \succ) \notin C$ and $(y, x, \succ) \notin C$ then $(x, y, \sim) \in C$ is a completeness requirement for preferences.

(ii) A consistency requirement is of the form 'if c^+ then m^-'. For example, if $(x_1, x_2, \succ) \in C$ and $(x_3, x_4, \succ) \in C$ and '...' and $(x_{n-1}, x_n, \succ) \in C$ then $(x_n, x_1, \succ) \notin C$ is a consistency requirement for preferences, i.e. acyclicity of preferences.

(iii) A closedness requirement is of the form 'if c^+ then m^+'. For example, if $(x, y, \succ) \in C$ and $(y, z, \succ) \in C$ then $(x, z, \succ) \in C$ is a closedness requirement for preferences, i.e. transitivity of preferences.

(iv) A negative closedness requirement is of the form 'if c^- then m^-' (equivalently, 'if m^+ then d^+'). For example, if $(y, x, \succ) \notin C$ and $(z, y, \succ) \notin C$ then $(z, x, \succ) \notin C$ is a negative closedness requirement for preferences, i.e. negative transitivity.

What is particular about the four-type taxonomy is that it represents all possible ways of linking M to a single mental state m and that analogously rules specify all 'correct' ways of linking premises to a single conclusion. I represent these types as a table of *'If ... then ...'* statements linking M to a single mental state m shown as Table 1. Rows are different antecedents and columns are different consequents. The contrapositives of these requirements can also be written as a table of *'If ... then ...'* statements shown as Table 2.

Apart from one exception, one cannot express a requirement belonging to one type in the four-type taxonomy T as a requirement belonging to another type in T. The exception is the case of closedness and negative closedness requirement in which the antecedent is a singleton because if we switch this antecedent with the consequent and negate them, closedness becomes a negative closedness. More precisely:

		then	
		m^+	m^-
if	c^+	closedness	consistency
	c^-	completeness	negative closedness

Table 1.1: A four-type taxonomy of requirements.

		then	
		d^+	d^-
if	m^+	negative closedness	consistency
	m^-	completeness	closedness

Table 1.2: A four-type taxonomy of requirements again.

Definition 1.4. A taxonomy of requirements T is **weakly exclusive** if, if a requirement R is of one type in T then it is not of any other type in T, i.e. $\mathcal{R} \cap \mathcal{R}' = \emptyset$ for any distinct requirement types $\mathcal{R}, \mathcal{R}' \in T$.

Proposition 1.1. *Let T be a taxonomy consisting of completeness, consistency, closedness, and negative closedness requirements. T is weakly exclusive, except in the case of a closedness or negative closedness requirement in which the antecedent is a singleton.*

Can any possible single requirement be expressed as a conjunction of requirements in the four-type taxonomy? To show this, it is useful to consider the following more explicit way of conceptualising the four-type taxonomy:

(i) A requirement R is a completeness requirement if it is of the form 'if absence of all elements of $M \setminus \{m\}$ then presence of m' with respect to a set of mental states $M \subseteq \mathcal{M}$ with $M \neq \emptyset$ and $m \in M$; formally: $M \cap C \neq \emptyset$.

(ii) A requirement R is a consistency requirement if it is of the form 'if presence of all elements of $M \setminus \{m\}$ then absence of m' with respect to a set of mental states $M \subseteq \mathcal{M}$ with $M \neq \emptyset$ and $m \in M$; formally: $\neg M \subseteq C$.

(iii) A requirement R is a closedness requirement if it is of the form 'if presence of all elements of M then presence of m' with respect to a set of mental states $M \subset \mathcal{M}$ with $M \neq \emptyset$ and a mental state $m \in \mathcal{M}$ with $m \notin M$; formally: $M \subseteq C \Rightarrow m \in C$.

(iv) A requirement R is a negative closedness requirement if it is of the form 'if absence of all elements of M then absence of m' (equivalently, 'if presence of m then presence of at least one element of M') with respect to a set of mental

states $M \subset \mathcal{M}$ with $M \neq \emptyset$ and a mental state $m \in \mathcal{M}$ with $m \notin M$; formally: $M \cap C = \emptyset \Rightarrow m \notin C$.

Note two things. The first is that there is a tight relationship between completeness and consistency requirements and between closedeness and negative closedness requirements. For any given M, there is exactly one completeness and one consistency requirement with respect to M. Similarly, for any given M and m, there is exactly one closedness and one negative closedness requirement with respect to M and m. A completeness requirement says that if all but one of the elements of M are not in C, then the remaining element is in C. A consistency requirement says that if all but one of the elements of M are in C, then the remaining element is not in C. What one of the two requirements says about C, the other says about the complement of C. Similarly, a closedness requirement says that if all the elements of M are in C, then m is in C whereas negative closedness says that if all the elements of M are not in C, then m is not in C. What one of the two requirements says about C, the other says about the complement of C. A special case is that if $M \setminus \{m\}$ is a singleton, the contrapositive of a closednesss requirement is a negative closedness requirement, hence Prop. 1. More formally:

Definition 1.5. For any two requirements R and R', R is the **dual** of R' if, for all $C \in \mathcal{C}$, $C \in R$ if and only if $\mathcal{M} \setminus C \in R'$.

The second is that, by definition, a requirement allows some set of constitutions R. Therefore, it rules out some set of constitutions, i.e. the complement of the allowed constitutions (i.e. $R' = \mathcal{M} \setminus C$). Some requirements rule out exactly one constitution: (i) $\emptyset$ can be ruled out by a completeness requirement, (ii) $\mathcal{M}$ can be ruled out by a consistency requirement, (iii) $\mathcal{M} \setminus \{m\}$ can be ruled out by a closedness requirement, and (iv) $\{m\}$ can be ruled out by a negative closedness requirement. And since any requirement can be understood by the constitutions it rules out, every possible requirement can be written as a conjunction of requirements that rule out exactly one constitution.

Definition 1.6. A taxonomy of requirements is **weakly exhaustive** if every possible single requirement is a conjunction of requirements from the taxonomy.[2]

[2]Weak exhaustiveness implies a notion of exhaustiveness from which exhaustiveness is weak and weak exclusiveness implies a notion of exclusiveness from which exclusiveness is weak. I state them for completeness. A taxonomy is (i) *strongly exhaustive* if every possible single requirement is of one of the types, i.e. $\bigcup_{\mathcal{R} \in T} \mathcal{R}$ contains all requirements R, and (ii) *strongly exclusive* if no requirement of one type is a conjunction of requirements of other types, i.e. no requirement in some $\mathcal{R} \in T$ is the conjunction of a set of requirements, i.e. $S \subseteq \bigcup_{\mathcal{R}' \in T \setminus \mathcal{R}} \mathcal{R}'$.

		then			
		c^+	d^+	d^-	c^-
if	c^+	(1)	(2)	(3)	(4)
	d^+	(5)	(6)	(7)	(8)
	d^-	(9)	(10)	(11)	(12)
	c^-	(13)	(14)	(15)	(16)

Table 1.3: An expanded table of *'If ... then ...'* statements.

		then			
		c^+	d^+	d^-	c^-
if	c^+	$\mathcal{R}1$	$\mathcal{R}2$	$\mathcal{R}3$	$\mathcal{R}4$
	d^+	$\mathcal{R}5$	$\mathcal{R}6$	$\mathcal{R}4$	$\mathcal{R}7$
	d^-	$\mathcal{R}8$	$\mathcal{R}9$	$\mathcal{R}1$	$\mathcal{R}5$
	c^-	$\mathcal{R}9$	$\mathcal{R}10$	$\mathcal{R}2$	$\mathcal{R}6$

Table 1.4: A ten-type taxonomy of requirements.

Proposition 1.2. *Let T be a taxonomy consisting of completeness, consistency, closedness, and negative closedness requirements. T is weakly exhaustive if and only if $\mathcal{M}$ has no more than three mental states.*

What are some of the requirement types that are not in the four-type taxonomy? To answer this consider an expanded table of *'If ... then ...'* statements linking one set M to another set N shown as Table 3.

What are the requirement types in this table? Since we are concerned only with formal structure, we can transpose M and N. This gives us 10 requirement types (Table 4). (1) and (11), (2) and (15), (4) and (7), (5) and (12), (6) and (16), and (10) and (13) are equivalent. This gives us 6 requirement types. Each of (3), (8), (9), or (14) is its own equivalent. This gives us 4 additional requirement types.

Can any of these be expressed as a conjunction of requirements in the taxonomy? In particular, $\mathcal{R}2$, $\mathcal{R}3$, and $\mathcal{R}10$ cannot be expressed as a conjunction of requirements in T. These are the requirements with the form 'if c then d'. This result motivates the following single type of requirement called conditional completeness requirement.

Definition 1.7. A requirement R is a **conditional completeness requirement** if it is of the form 'if presence of all elements of M then presence of at least one element of N' with respect to a pair of sets of mental states $M, N \subseteq \mathcal{M}$ with $M \cap N = \emptyset$ and $M, N \neq \emptyset$; formally: $M \subseteq C \Rightarrow N \cap C \neq \emptyset$.

One obtains this type of requirement if she relaxes the antecedent of the negative closedness requirement in the original form to be a nonempty set N of mental states. Using the simplified notation introduced at the beginning of this section,

this requirement is of the form 'if c^+ then d^+'. The following type of requirement is a generalised version of conditional-completeness requirements which no longer impose that $M, N \neq \emptyset$ but that $M \cup N \neq \emptyset$ called a unified requirement.

Definition 1.8. A requirement R is a **unified requirement** if it is of the form 'if presence of all elements of M then presence of at least one element of N' with respect to a pair of sets of mental states $M, N \subseteq \mathcal{M}$ with $M \cap N = \emptyset$ and $M \cup N \neq \emptyset$; formally: $M \subseteq C \Rightarrow N \cap C \neq \emptyset$.

Remark 1.2. This requirement is its own dual. The unified requirement says that if all the elements of M are in C, then some m from N is in C. The contrapositive of this is: If no element of N is in C, then some element of M is not in C, i.e. if all elements of N are not in C, then some element of M is not in C, i.e. the 'dual'.

All types of requirement considered above are special cases of this type.

Proposition 1.3. *Unified requirements generalise completeness, consistency, closedness, and negative closedness requirements.*

A special case of the requirement considered above is the requirement that imposes $N = \mathcal{M} \setminus M$. This requirement uniquely rules out M. Since any requirement can be understood by the constitutions it rules out, every possible requirement can be written as a conjunction of requirements that rule out exactly one constitution. I will now show that every possible single requirement is a conjunction of generalised requirements of this type with a finite number of mental states.

Proposition 1.4. *Let T be a taxonomy consisting of unified requirements. T is weakly exhaustive if $\mathcal{M}$ has a finite number of mental states.*

So every requirement of a theory can be written as conjunctions of requirements from T as defined above.

1.4 Theories of rationality

Note that the definition of a requirement treats requirements in a generic way; a requirement is not necessarily a requirement of rationality, or bound to a specific theory of rationality. Rationality is one of many possible sources of requirements: other sources of requirements might be morality, prudence, fashion, or Catholicism.[3] More precisely, given a mental structure $(\mathcal{L}, \mathcal{A})$, a "theory of rationality requirements" is

[3]See also (2013, p. 116)

a given set $\mathcal{T}$ of constitutions which are deemed rational.[4] I will define a theory of rationality by its requirements.[5]

Definition 1.9. A **theory of rationality requirements** is a requirement, denoted $\mathcal{T} \subseteq \mathcal{C}$. A constitution C is:

1. EITHER *($\mathcal{T}$)-rational* if $C \in \mathcal{T}$.
2. OR *($\mathcal{T}$)-irrational* if $C \notin \mathcal{T}$.

Remark 1.3. By inspection of Definition 1.9, one observes that:

1. if $\mathcal{T} = \mathcal{C}$, then the theory is **tautological**.
2. if $\mathcal{T} = \emptyset$, then the theory is **contradictory**.

Definition 1.10. A requirement R is called a **requirement of theory** $\mathcal{T}$ or simply a $\mathcal{T}$-requirement if R follows from $\mathcal{T}$, i.e. $\mathcal{T} \subseteq R$.

Remark 1.4. Given a theory $\mathcal{T}$, a constitution is $\mathcal{T}$-rational iff it satisfies all $\mathcal{T}$-requirements.

I now use the analysis of the formal structure of requirements above and discuss its use in two examples of theories of rationality: (i) Broome's philosophical analysis of rationality and (ii) rational choice theory under certainty. Taking inspiration from Broome's analysis of rationality I develop a 'Broomean' version of rational choice.

1.4.1 Broome's analysis

Broome is primarily interested in representing the agent's internal language relevant to practical reasoning. Broome's intuition is that you reason with the *marked contents* of your mental states expressed in a natural language (2013, pp. 253-4). The marked content is a pair that consists of a marker that specifies a type of attitude and the proposition. Since Broome's analysis is concerned with the kind of reasoning that operates on mental states that correspond to marked contents, I use a simple model in which mental states represent the marked contents relevant to practical reasoning. These mental states are formed by the sets:

1. $\mathcal{L} = \{p, q, ...\}$ which is sufficiently rich as to contain all relevant propositions, and

[4] Informally, a theory of rationality is given by its requirements. Since the conjunction of requirements of a theory is again required, one can form the conjunction of all requirements of the theory which yields the theory's strongest requirement, a particular set of constitutions which are allowed according to that theory.

[5] In these cases, requirements are internal to a theory or a conception of rationality

2. $\mathcal{A} = \{int, bel\}$ which consists of the intention and belief attitudes.

His analysis focuses on beliefs and intentions, aiming to keep his analysis within the domain of internal reasoning prior to any action. There are other attitudes as well such as desires, hopes, and others. One could even speak of graded beliefs and desires as such. But beliefs and intentions are essential for practical reasoning: a person will at least have the attitude of belief and of intention towards propositions when engaging with practical reasoning, such as the belief that it is sunny or the intention that I go sailing.[6]

Here are some examples:

- the mental state (p, int) describes an intention *to do something*.

For example saying to yourself *"I shall go sailing"* is normally the way in which you express in language your intention to go sailing. The convention I adopt is to use the variables $p, q, ...$ to refer generically to propositions in $\mathcal{L}$, and text in quotes such as *"it rains"*, *"I shall go biking"*, *"if it rains tomorrow, then the game will not take place"* to express, in language, the marked contents of your attitudes. I use *italics* such as *it rains*, *I go sailing*, *if it rains tomorrow then the game will not take place*,... to denote the unmarked content of your attitudes. So *I go sailing* is the unmarked content of the intention in the example and (*I go sailing*, *int*) is the marked content of this intention.

- the mental state (p, bel) describes a belief *that something is the case*.

For example saying to yourself *"if it rains tomorrow, then the game will not take place"* is normally the way in which you express this belief of yours. In the sentence the marker is silent.[7] The marked content of this belief is (*if it rains tomorrow then the game will not take place*, *bel*).

Broome discusses various basic theoretical and practical rationality requirements, in particular, the *modus ponens*, *non-contradiction of beliefs*, *means-end*, *non-contradiction of intentions*, and *enkrasia* requirements. I now give examples of these schemas of requirements:

- The *modus ponens* requirement schema:

  if $\{(p,\ bel),\ (if\ p\ then\ q,\ bel)\} \subseteq C$, then $(q,\ bel) \in C$

[6] In discussing practical reasoning in terms of beliefs and intentions, Broome is following a common practice among philosophers. An analysis of attitudes that is focused on intentions appears in (Bratman, 1987). Bratman thinks of intentions as some sort of plans towards the action.

[7] It is a intrinsic property of English that you can express a belief without a marker (2013, pp. 254-5).

- The *non-contradiction of beliefs* requirement schema:

  if $(p,\ bel) \in C$, then $(not\ p,\ bel) \notin C$

Broome extends his analysis of requirements to apply to intentions.

- The *means-end* requirement schema:

  if $\{(p,\ int), (q\ is\ a\ means\ implied\ by\ p,\ bel), (q\ is\ up\ to\ me,\ bel)\} \subseteq C$, then $(q,\ int) \in C$

- The *non-contradiction of intentions* requirement schema:

  if $(p,\ int) \in C$, then $(not\ p\ ,\ int) \notin C$

Enkrasia is another requirement of practical rationality. It involves a belief that you ought to do something, a 'normative' belief, and an intention to do it.[8]

- The *enkrasia* requirement schema:

  if $\{(ought\ to\ p,\ bel), (p\ is\ up\ to\ me,\ bel)\} \subseteq C$, then $(p,\ int) \in C$

1.4.2 Rational Choice

In this section, I restrict my analysis to the simplest theory of rational choice: rational choice under certainty, others are decision theory under uncertainty and game theory.[9],[10] In the standard non-Broomean versions of the theory, a set of all possible options is given. Let X be the given set of all possible options in the standard theory. Typically a single option, e.g. apple, banana, is denoted x, y, ... and so on.

These are the objects of preference and choice. Preferences are represented as binary relations on pairs of options. For mathematical convenience, economists usually start from the weak preference relation and derive the strict preference relation and the indifference relation. Strict preference relations, indifference relations, and weak preference relations are all binary relations on X, where X is a set of objects of choice and not propositions. Moreover, choices are represented as choice functions that assign the set of 'choiceworthy' objects from the feasible set.

[8] Ought has a normative content here. It is often used in two senses. Oughts that refer to all-things-considered reasons and oughts that refer to specific reasons. Oughts in the first sense outweigh all other reasons the agent might have, whereas oughts in the second sense do not, and are often stated as 'have reason'. Broome leaves open whether rationality is normative, whether it generates normative oughts (2013, p. 146).

[9] For example, the treatment of rational choice under certainty in (Kreps, 1988)

[10] e.g. (Savage, 1954) and (Aumann, 1999) respectively.

Definition 1.11. Let $\geq$ be a weak preference relation on X. Then define the strict preference relation $>$ and the indifference relation $\equiv$ on X as follows:

- $x > y$ if and only if $x \geq y$ and not $y \geq x$, and
- $x \equiv y$ if and only if $x \geq y$ and $y \geq x$.

In the standard theory of choice, transitivity and completeness are usually formulated in terms of weak preference relations as defined above. For any relation $\geq$ on X, the requirement of transitivity requires that $\geq$ is transitive and the requirement of completeness that it is complete.

- The relation $\geq$ is transitive if, for any x, y, z in the set X, if $x \geq y$ and $y \geq z$, then $x \geq z$.
- The relation $\geq$ is complete if, for any x, y in the set X, either $x \geq y$ or $y \geq x$.

In my Broomean model, I identify the set of options, e.g. apple, banana, with a set of propositions of possible consequences of choice, e.g. *I get apple*, *I get banana*, and so on. Formally, X is the set of all possible propositions of consequences of choice. Any such proposition is denoted x. These propositions are the objects of psychological attitudes and not mathematical objects such as relations or functions. Propositions from X form the domain of the attitude of preference, the domain of the attitude of indifference, and the domain of the attitude of intention to choose. Strict preferences, indifferences and intentions to choose are the counterpart of binary relations and choice functions in standard versions of the theory.

Preferences are two-place attitudes, denoted $\succ$ and $\sim$. The following are the two mental states that describe preference and indifference attitudes:

- the mental state $(x, y, \succ)$ describes any preference, e.g the preference (*I get apples*, *I get bananas*, $\succ$) given that apples and bananas are objects of choice.
- the mental state $(x, y, \sim)$ describes any indifference, e.g the indifference (*I get apples*, *I get bananas*, $\sim$) given that apples and bananas are objects of choice.

Strict preferences and indifferences seem to be the natural ways for expressing our comparative attitudes with language. The marked contents of preferences can be understood in two ways; either preferences as comparative desires or as beliefs about betterness. For example, saying to yourself *"Rather I get x than I get y"* is one of the possibles ways of expressing a preference between x and y. The other is saying to yourself *"getting x is better than getting y"*. I say earlier that economists often use weak preference relations for mathematical convenience. In practice, we can

express a weak preference relation with the proposition *"x is at least as good as y"*. This proposition is the object of a belief. If we do this, then we do not need to have a preference attitude as the content of this proposition is not the content of a mental state but that of a disjunction of mental states: a strict preference and indifference. If preferences are interpreted as a comparative desire, then this is an important constraint for expressing weak preferences with language.

Intentions are the Broomean counterpart of choice. Intentions are naturally understood as the final output of a mental process prior to any choice. The following mental state describe these intentions:

- the mental state (x, int) describes an intention to choose something, e.g the intention (*I get apples*, *int*) given that apples are an object of choice.

For example, saying to yourself *"I shall get x"* is normally the way in which you express your intention to get x. Intentions are in many ways different from desires in guiding our actions. One important way in which they differ is that intentions persist while this is not so for desires. So an intention that is caused by a preference gives particular meaning to preferences, one that economists assume.

Usually your preferences and intentions depend on your beliefs about the set of objects which are available to you. I identify each non-empty subset of X with a proposition about the feasibility of a choice set. Formally, Y denotes any nonempty subset of X, i.e. any element of $2^X \setminus \{\emptyset\}$. The following example illustrates this belief:

- the mental state (Y, bel) describes a belief in a feasible set, e.g the belief (*I get either apples or bananas*, *bel*) given that the choice set $Y = \{x, y\}$ contains only apples and bananas as objects of choice.

Beliefs are commonly expressed in the indicative mood (2013, p. 268). For example in the case in which $Y = \{x, y\}$, the agent says to herself *"I can only get either x or y"* which is normally the way in which you express in language your belief that you can only get either x or y.

Note that feasibility beliefs and intentions are absent from standard formulations of rational-choice models. Feasibility beliefs are absent from standard choice models, because of the background assumption that the feasible set is automatically known. Moreover, intentions are absent from these models as these models aim to describe observable behaviour. Feasibility beliefs are needed to explain intentions and belong to any complete description of the agent's mind.

One can now construct the mental states of rational choice from combinations of:

1. the set $\mathcal{L} = X \cup 2^X \setminus \{\emptyset\}$ which contains all choice related propositions, and

2. the set $\mathcal{A} = \{int, bel, \succ, \sim\}$ which contains all choice related attitudes, i.e. the attitude of intention, belief, strict preference and indifference.

For any $x, y \in X$ and $Y \subset X$, the set of *all* rational choice mental states consists of the following four types of mental states:

1. $(x, y, \succ)$ describes a preference, e.g. the agent says to herself *"Rather I get x than I get y"*,

2. $(x, y, \sim)$ describes an indifference, e.g. the agent says to herself *"Just as well I get x as I get y"*,

3. (x, int) describes an intention to choose, e.g. the agent says to herself *"I shall get x"*, and

4. (Y, bel) describes a belief in a feasible set, e.g. in the case that $Y = \{x, y\}$ the agent says to herself *"I can only get either x or y"*.

I can now present my non-standard rational choice axiomatisation on the above set of mental states. Eight requirements (R1-R8) make together the proposed 'Broomean' rational choice. There are two points to made about this axiomatisation. First, binary relations are separate mathematical objects and cannot be in the set $\mathcal{A} = \{int, bel, \succ, \sim\}$ of attitudes. The set of all mental states which can be formed by the pair of attitudes $(\succ, \sim)$ is the set $M_{(\succ,\sim)} = \{(x, y, a) : x, y \in X \text{ and } a \in \{\succ, \sim\}\}$. The second is that a weak preference relation $\geq$ on X cannot be given by a single attitude but by the disjunction of strict preference and indifference attitudes. To this end, I will relate the concept of weak preference to that of constitutions.

Definition 1.12. Given your mental structure $(\mathcal{L}, \mathcal{A})$, a constitution C for the above model is *compatible with* the binary relation $\geq$ if $C \cap M_{(\succ,\sim)} = \{(x, y, \succ) : x, y \in X \text{ and } x > y\} \cup \{(x, y, \sim) : x, y \in X \text{ and } x \equiv y\}$.

Theorem 1.1. *A constitution C is compatible with some weak preference relation iff for all $x, y \in X$, C satisfies the three requirement schemas:*

- *R2* asymmetry of preference: *if* $(x, y, \succ) \in C$ *then* $(y, x, \succ) \notin C$ (a consistency requirement)

- *R3* (incompatibility of preference and indifference): *if* $(x, y, \succ) \in C$ *then* $(x, y, \sim) \notin C$ (a consistency requirement)

- *R4* (symmetry of indifference): *if* $(x, y, \sim) \in C$ *then* $(y, x, \sim) \in C$ (a closedness requirement)

The theorem says that incompatibility of any constitution with weak preference can be attributed to violations of three implicit requirements on strict preferences and indifference attitudes. Given theorem 1.1, one can recast the requirement schemas transitivity and completeness of weak preference relations by the pair $(\succ, \sim)$ of attitudes. Transitivity of weak preferences is a conjunction of four transitivity requirement schemas, i.e. one for each pair of preference and indifference attitudes.

Proposition 1.5. *A constitution* C *is compatible with some transitive weak preference relation iff (i) for all* $x, y \in X$, C *satisfies R2, R3, R4 in Th. 1.1, and (ii) for all* $x, y, z \in X$ *it satisfies the four requirement schemas:*

- *R1*$\succ$ *transitivity of strict preference:if* $\{(x, y, \succ), (y, z, \succ)\} \subseteq C$ *then* $(x, z, \succ) \in C$ (a closedness requirement)
- *R1*$\sim$ *transitivity of indifference: if* $\{(x, y, \sim), (y, z, \sim)\} \subseteq C$ *then* $(x, z, \sim) \in C$ (a closedness requirement)
- *R1*$\succ\sim$ *PI–transitivity: if* $\{(x, y, \succ), (y, z, \sim)\} \subseteq C$ *then* $(x, z, \succ) \in C$ (a closedness requirement)
- *R1*$\sim\succ$ *IP–transitivity: if* $\{(x, y, \sim), (y, z, \succ)\} \subseteq C$ *then* $(x, z, \succ) \in C$ (a closedness requirement)

Proposition 1.6. *A constitution* C *is compatible with some complete weak preference relation iff for all* $x, y \in X$, C *satisfies R2, R3, R4 in Th. 1.1 and the following requirement schema:*

- *R5 completeness of preferences: if* $(x, y, \succ) \notin C$ *and* $(y, x, \succ) \notin C$, *then* $(x, y, \sim) \in C$ (a completeness requirement)

Corollary 1.1. *Define the relation* $\geq$ *as reflexive if, for any* x *in the set* X, $x \geq x$. *A constitution* C *is compatible with some reflexive and complete weak preference relation iff for all* $x \in X$, C *satisfies R2, R3, R4 in Th. 1.1, R5 and the following requirement schema:*

- *R5+ reflexivity of indifference: if* $(x, x, \sim) \in C$ *then* $(x, x, \sim) \in C$ (a closedness requirement)

R1-R5 ensure fully classical preferences. Given Proposition 1.6, one can simplify transitivity of weak preferences to a more economical version of transitivity given completeness with one requirement instead of four.

Proposition 1.7. *A constitution C is compatible with some complete and transitive weak preference relation iff (i) for all $x, y \in X$, C satisfies R2, R3, R4 in Th. 1.1 and R5 of Prop. 1.6, and (ii) for all $x, y, z \in X$ it satisfies the following requirement schema:*

- *R1+ negative transitivity: if $(y, x, \succ) \notin C$ and $(z, y, \succ) \notin C$ then $(z, x, \succ) \notin C$* (a negative closedness requirement)

These results suggest different interpretations of weak preferences. On the first interpretation, strict preference and indifference are primitive, non-composite attitudes and therefore weak preference relations are derived from them. Completeness is not an implicit requirement on preferences. On the second interpretation, strict preferences are truly primitive and given completeness, indifferences are derived from the absence of strict preferences. A weak preference relation is therefore understood as the complement of the primitive strict preference relation. Given completeness, transitivity of weak preferences can be reduced to negative transitivity.[11]

But the schemas of Th. 1.1 seem to be rationality requirements on preferences in a way that completeness is not. Completeness is questionable from both a descriptive and a normative point of view. Comparing propositions from X is not always possible and in some cases it is non-desirable. The absence of a preference could express indifference or non-comparability. By describing preference relations by the pair $(\succ, \sim)$ one accounts for this difference.

The next two requirement schemas, R6 and R7, exclude contradictory intentions or feasibility beliefs:

- For distinct feasible sets $Y, Y' \in 2^X \setminus \{\emptyset\}$:

  R6 *No conflicting feasibility beliefs*: if $(Y, bel) \in C$ then $(Y', bel) \notin C$ (a consistency requirement)

- For distinct options $x, y \in X$:

  R7 *No conflicting intentions*: if $(x, int) \in C$ then $(y, int) \notin C$ (a consistency requirement)

[11] A well known treatment of the second interpretation is that of Savage. His definition (1954, p. 18) of the weak preference relation "x is not preferred to y" is built on a strict preference and two *implicit* requirements on preferences: asymmetry of preferences and completeness. Starting with strict preference as the only primitive, Savage is able to derive its complement strict preference relation "x is not preferred to y", i.e. $(x, y, \succ) \notin C$ assuming that strict preferences are non-contradictory and indifference to be the absence of strict preferences in both directions.

The next requirement schema connects feasibility beliefs and preferences to intentions to choose. Consider this as the counterpart of *enkrasia* in economic theory. Rational choice theory as a theory of practical rationality must contain a kind of requirement which is about the relationship between preferences, as the natural primitive input of the theory, and intention to choose, as the final output of those preferences. Informally this requirement schema states that if you believe that you can get x or y and most prefer to get x, then you intend to get x. In the language of the theory,

- For any feasible set $Y \subseteq X$ and $x \in Y$:

  R8$\succ$ *Simple economic enkrasia*: if $(Y, bel) \in C$, $(x, y, \succ) \in C$ for all $y \in Y \setminus \{x\}$, then $(x, int) \in C$ (a closedness requirement)

Economic enkrasia can be formulated in a more general way to cover the case of ties between different top-ranked feasible options. The following more general requirement schema achieves this by replacing the top option x by a non-empty set of top options Z. In these cases, there can be more than one intention that satisfy the requirement.

- For any feasible set $Y \subseteq X$ and non-empty set $Z \subseteq Y$:

  R8 *economic enkrasia*: if $(Y, bel) \in C$, $(x, y, \succ) \in C$ for all $x \in Z, y \in Y \setminus Z$, and $(x, y, \sim) \in C$ for all distinct $x, y \in Z$, then $(x, int) \in C$ for some $x \in Z$ (a conditional completeness requirement)

Informally, this requirement schema says that if you believe the feasible set to be Y and prefer options in Z to options in $Y \setminus Z$ and view options in Z as mutually indifferent, then you intend some option in Z. Conventional rational choice theory analyses choice in terms of choice acts directly rather than in terms of intentions prior to any action. Choice acts are mathematically described by choice functions which are usually assumed to pick a non-empty set of 'choiceworthy' options. This is to deal with indifference. If two or more options are choiceworthy, rational choice theory doesn't say anything about which should be chosen. In this Broomean version of the theory, however, the agent needs to have an intention to get a single object from X. This issue is discussed in relation to Buridan's ass paradox in Chapter 2.

I presented a 'Broomean' version of rational choice that consists of eight requirements. Requirements R1-R7 correspond to one of the four types of the requirement while R8 is a conditional completeness requirement. R1-R5 ensure fully classical preferences and R6-R7 exclude contradictory intentions or feasibility beliefs. Of them all, R8 is the most 'Broomean' one; the choice-theoretic counterpart

of Broome's ordinary enkrasia. It differs from ordinary enkrasia in that intentions respond to preferences and feasibility beliefs rather than ought-beliefs. R8 reflects the classical preference-maximisation hypothesis: you intend something that you most prefer among what you believe to be feasible.

1.5 Conclusion

This chapter started with an observation about the structural analogy that exists between some requirement types written in *'If ... then ...'* form and rules of reasoning. Requirements whose consequent is a proposition about a single mental state and antecedent is a proposition about a set of mental states are structurally analogous to rules that specify 'correct' ways of deriving a conclusion from a set of premises.[12] I presented a novel taxonomy of all requirement types that have this property. I showed that, except for a particular case, if a requirement is of one type in the four-type taxonomy then it is not of any other type in the taxonomy.

I then used the taxonomy of requirements to discuss rational choice as a main example of a theory of rationality requirements. This simple version of rational choice has been inspired by Broome's philosophical analysis of mental states and rationality. The next chapter uses this taxonomy to show that certain types of requirement cannot be reached by a particular process of reasoning.

1.6 Appendix

Proof of Prop. 1.1. Part 1: Notice:

(i) every completeness requirement forbids $C = \emptyset$ and permits $C = \mathcal{M}$

(ii) every consistency requirement permits $C = \emptyset$ and forbids $C = \mathcal{M}$

(iii) every closedness requirement permits $C = \emptyset$ and permits $C = \mathcal{M}$

(iv) every negative closedness requirement permits $C = \emptyset$ and permits $C = \mathcal{M}$.

From this it follows immediately that the only types than can possibly overlap are closedness and negative closedness requirements.

Part 2: Let R be a closedenss requirement specified for M, m. Let R' be a negative closedness requirement specified for M', m'. R says: if all elements of M

[12] Theories such as the AGM theory of belief revision (Alchourrón et al., 1985; Gärdenfors, 1988) and non-Bayesian models of preference revision Grüne-Yanoff and Hansson (2009) represent different ways of changing by expanding or contracting the current set of mental states in light of new information. My analysis which is on the formal structure of conditional statements can be used to classify logical concepts such as the expansion, contraction, and revision in revision theories.

are in C, then m is in C. R' says: if all elements of M' are not in C, then m' is not in C. The contrapositive form of R' says: if m' is in C, then some element of M' is in C. Given M and m, it's possible to restate R as a negative closedness requirement only if M is a singleton, i.e. $M = \{m^*\}$. In this case I can set $m' = m^*$, $M' = \{m\}$, and then R' and R are equivalent. □

Proof of Prop. 1.2. Part 1: I suppose that $\mathcal{M}$ contains either one or two or three mental states and prove that T is weakly exhaustive.

Claim 1: If $\mathcal{M}$ has one or two or three mental states, then for any constitution C, there is some requirement R in the taxonomy that uniquely *rules out* C.

Proof: By definition, a requirement allows some set of constitutions R. Therefore, it rules out some set of constitutions, i.e. the complement of the allowed constitutions (i.e. $R' = \mathcal{M} \setminus C$). A requirement *uniquely rules out* a constitution C if it rules out C and no other constitution C'. Let $\mathcal{M}$ be the set of all mental states and $\mathcal{C} = 2^{\mathcal{M}}$ be the set of all possible constitutions. Fix the set $\mathcal{M}$ such that it contains either one mental state, or two mental states, or three mental states. Consider any requirement R from the taxonomy T that consists of completeness, consistency, closedness, and negative closedness requirements denoted $\mathcal{R}com$, $\mathcal{R}con$, $\mathcal{R}clo$, and $\mathcal{R}nclo$ respectively:

$\mathcal{R}com$ completeness requirements: Consider a completeness requirement with $M \subset \mathcal{M}$, i.e. there is some $k \in \mathcal{M}$ such that $k \notin M$. This requirement rules out the constitution $\emptyset$, but it also rules out $\{k\}$. So the only kind of completeness requirement that can rule out exactly one constitution is one with $M = \mathcal{M}$. This rules out $\emptyset$ and nothing else.

$\mathcal{R}con$ consistency requirements: By duality with $\mathcal{R}com$, the only kind of consistency requirement that can rule out exactly one constitution is one with $M = \mathcal{M}$. This rules out $\mathcal{M}$ and nothing else.

$\mathcal{R}clo$ closedness requirements: Consider a closedness requirement with $M \cup \{m\} \subset \mathcal{M}$, i.e. there is some $k \in \mathcal{M}$ such that $k \notin M$ and $k \neq m$. This requirement rules out M, but it also rules out $M \cup \{k\}$. So the only kind of closedness requirement that can rule out exactly one constitution is one with $M \cup \{m\} = \mathcal{M}$. This requirement rules out $\mathcal{M} \setminus \{m\}$ and nothing else.

$\mathcal{R}nclo$ negative closedness requirements: By duality with $\mathcal{R}clo$, the only kind of negative closedness requirement that can rule out exactly one constitution is one with $M \cup \{m\} = \mathcal{M}$. This requirement rules out $\{m\}$ and nothing else.

Implications of this for the question of whether every possible requirement is a conjunction of requirements of types $\mathcal{R}com$, $\mathcal{R}con$, $\mathcal{R}clo$, and $\mathcal{R}nclo$: Any require-

ment can be described by the set of mental states that it rules out. So if each mental state in $\mathcal{M}$ is uniquely ruled out by some requirement in the taxonomy, we know that every possible requirement is a conjunction of requirements in the taxonomy.

From the previous results, for any M: (a) $\emptyset$ can be ruled out by a completeness requirement, (b) $\mathcal{M}$ can be ruled out by a consistency requirement, (c) $\mathcal{M} \setminus \{m\}$ can be ruled out by a closedness requirement, and (d) $\{m\}$ can be ruled out by a negative closedness requirement. So each of the eight possible constitutions is uniquely ruled out by *at least one* requirement of the taxonomy.

It follows that if $\mathcal{M}$ contains no more than three mental states, every possible constitution can be uniquely ruled out by some requirement from the taxonomy, and so every possible requirement is equivalent to a conjunction of requirements from the taxonomy.

Part 2: I want to prove that if T is weakly exhaustive, then $\mathcal{M}$ contains at most three mental states. I prove this by contraposition. I suppose that $\mathcal{M}$ contains more than three mental states and prove that T is not weakly exhaustive.

Claim 2: Consider any constitution C. If $\mathcal{M}$ has more than three mental states, then for some constitution C, there is no requirement R in the taxonomy that uniquely rules out C.

Proof: Let $\mathcal{M}$ be the set of all mental states and $\mathcal{C} = 2^{\mathcal{M}}$ be the set of all possible constitutions. Fix the set $\mathcal{M}$ such that it has more than three mental states. Suppose $\mathcal{M} = \{i, j, k, l\}$. Then the constitution $C' = \{k, l\}$ cannot be uniquely ruled out by any single requirement from the taxonomy. Now consider the requirement that rules out only this constitution (in the example, the requirement that the constitution is not $\{k, l\}$). We know that no requirement from the taxonomy uniquely rules out C'. So any conjunction of these requirements must either (a) not rule out C', or (b) rule out C' and something else. So no conjunction of requirements is equivalent to the requirement that rules out only C'. □

Proof of Prop. 1.3. Let the unified requirement be given relative to $M, N \subseteq \mathcal{M}$ with $M \cup N \neq \emptyset$ such that $M \subseteq C \Rightarrow N \cap C \neq \emptyset$. Then a unified requirement generalises completeness, consistency, closedness, and negative closedness requirements.

If $M = \emptyset$ and $N \neq \emptyset$, then the unified requirement with respect to M and N reduces to completeness with respect to N, that is $N \cap C \neq \emptyset$.

If $M \neq \emptyset$ and $N = \emptyset$, then the unified requirement with respect to M and N reduces to consistency with respect to M, that is $\neg M \subseteq C$.

If $M \neq \emptyset$ and $N = \{m\}$, then the unified requirement with respect to M and N reduces to closedness with respect to M and m, that is $M \subseteq C \Rightarrow m \in C$.

If $M = \{m\}$ and $N \neq \emptyset$, then the unified requirement with respect to M and N reduces to negative closedness with respect to N and m, that is $m \in C \Rightarrow N \cap C \neq \emptyset$, which is $N \cap C = \emptyset \Rightarrow m \notin C$.

Thus each of requirements $\mathcal{R}com$, $\mathcal{R}con$, $\mathcal{R}clo$, and $\mathcal{R}nclo$ is a special case of the unified requirement. □

Proof of Prop. 1.4. Suppose that $\mathcal{M}$ contains a finite number of mental states. I prove that T is weakly exhaustive. It suffices to define the unified requirement $M \subseteq C \Rightarrow N \cap C \neq \emptyset$ with $M \cap N = \emptyset$ and $M \cup N = \mathcal{M} \neq \emptyset$. This requirement rules out M and nothing else. Since any requirement can be given by a pair of sets of mental states $M \subseteq \mathcal{M}$, $N \subseteq \mathcal{M}$, each constitution from $\mathcal{M}$ is uniquely ruled out by some unified requirement. By the results in Prop. 1.2, (a) $\emptyset$ can be ruled out by a unified requirement reduced to a completeness requirement, (b) $\mathcal{M}$ can be ruled out by a unified requirement reduced to a consistency requirement, (c) $\mathcal{M} \setminus \{m\}$ can be ruled out by a unified requirement reduced to a closedness requirement, (d) $\{m\}$ can be ruled out by a unified requirement reduced to a negative closedness requirement. Moreover, (e) any constitution $\mathcal{M} \setminus M$ not specified above can be ruled out by at least one of the other unified requirements. □

Proof of Theorem. 1.1. The proof has two parts. First I consider any constitution C that satisfies the three requirements above and show that there exists a binary relation $\geq$ on X such that C is compatible with it. In the second part I suppose that there exits a binary relation $\geq$ on X such that a constitution C is compatible with it and show that C satisfies all the requirements above.

Part 1: Consider a constitution C such that for all $x, y \in X$, C satisfies asymmetry of preference, incompatibility of preference and indifference, and symmetry of indifference.

Claim 1: For any given $x, y \in X$, one of the following is true:

(a) either C contains the strict preference $(x, y, \succ)$ and no other preferences and indifferences between x and y,

(b) or it contains the strict preference $(y, x, \succ)$ and no other preferences and indifferences between x and y,

(b) or it contains the indifferences $(x, y, \sim)$ and $(y, x, \sim)$ and no other preferences between x and y,

(c) or it does not contain any preferences or indifferences between x and y.

Proof: (a) Let $(x, y, \succ)$ be in C. By asymmetry, C cannot contain $(y, x, \succ)$. By incompatibility C cannot contain $(x, y, \sim)$. By symmetry this implies that C cannot contain $(y, x, \sim)$. So neither $(y, x, \succ)$ nor $(y, x, \sim)$ nor $(x, y, \sim)$ are in C.

(b) Let $(y, x, \succ)$ be in C. The proof that neither $(x, y \succ)$ nor $(x, y, \sim)$ nor $(y, x, \sim)$ are in C follows a similar logic.

(c) Let $(x, y, \sim)$ be in C. By symmetry, this implies that C contains $(y, x, \sim)$. By incompatibility neither $(x, y, \succ)$ nor $(y, x, \succ)$ are in C.

(d) If neither (a) nor (b) nor (c) hold, then by previous results neither $(y, x, \succ)$ nor $(y, x, \sim)$ nor $(x, y \succ)$ nor $(x, y, \sim)$ are in C.

Claim 2: For any given $x, y \in X$, the above constitution C is compatible with some binary relation $\geq$.

Proof: I construct the relation $\geq$ from strict preferences and indifferences using the rule: for any $x, y \in X$, $x \geq y$ if and only if either $(x, y, \succ) \in C$ or $(x, y, \sim) \in C$.

(a) Suppose that C contains $(x, y, \succ)$. Then neither $(y, x, \succ)$ nor $(y, x, \sim)$ are in C. Then by construction, $x \geq y$ and not $y \geq x$. By the definition of weak preferences, $x > y$.

(b) Suppose that C contains $(y, x, \succ)$. The proof that C is compatible with $y > x$ follows a similar logic.

(c) Suppose that C contains $(x, y, \sim)$. Then C contains $(y, x, \sim)$. Then by construction, $x \geq y$ and $y \geq x$. By the definition of weak preferences, $x \equiv y$.

(d) Suppose that C contains neither $(x, y, \succ)$ nor $(y, x, \succ)$ nor $(x, y, \sim)$ nor $(y, x, \sim)$. Then by construction, neither $x \geq y$ nor $y \geq x$. Thus neither $x > y$ nor $x \equiv y$ nor $y > x$ nor $y \equiv x$.

Part 2: I suppose that some relation $\geq$ on X exists and that a constitution C is compatible with it. Therefore, for any given $x, y \in X$, exactly one is true: (a) either $x \geq y$ and not $y \geq x$, (b) or $y \geq x$ and not $x \geq y$, (c) or both $x \geq y$ and $y \geq x$, (d) or neither $x \geq y$ nor $y \geq x$. I show that for each case, C satisfies all requirements above.

(a) Suppose $x \geq y$ and not $y \geq x$. Then by the definition of weak preferences, $x > y$ and neither $y > x$ nor $x \equiv y$ nor $y \equiv x$. Therefore, by compatibility of a constitution with it, $C \cap M_{(\succ, \sim)} = \{(x, y, \succ)\}$. Thus neither $(y, x, \succ)$ nor $(y, x, \sim)$ nor $(x, y, \sim)$ are in C. Because $(y, x, \succ)$ is not in C, then asymmetry is satisfied. Because $(x, y, \sim)$ is not in C, then incompatibility is satisfied. Because both $(x, y, \sim)$ and $(y, x, \sim)$ are not in C, then symmetry is satisfied. Hence C satisfies the requirements.

(b) The proof that the compatible constitution $C \cap M_{(\succ, \sim)}$ is the set $\{(y, x, \succ)\}$ and that it satisfies the requirements has a similar logic.

(c) Suppose $x \geq y$ and $y \geq x$. Then by the definition of weak preferences,

$x \equiv y$ and $y \equiv x$. By compatibility of a constitution with it, $C \cap M_{(\succ,\sim)} = \{(x,y,\sim),(y,x,\sim)\}$. Thus neither $(x,y,\succ)$ nor $(y,x,\succ)$ are in C. Because both $(x,y,\succ)$ and $(y,x,\succ)$ are not in C, then asymmetry is satisfied. Because neither $(x,y,\succ)$ nor $(y,x,\succ)$ are in C, then incompatibility is satisfied. Because $(x,y,\sim)$ and $(y,x,\sim)$ are in C, then symmetry is satisfied. Hence C satisfies the requirements.

(d) Suppose neither $x \geq y$ nor $y \geq x$. By the definition of weak preferences, neither $x > y$, nor $y > x$, nor $y \equiv x$, nor $x \equiv y$. By compatibility of a constitution with it, $C \cap M_{(\succ,\sim)} = \emptyset$. Then symmetry, asymmetry, and incompatibility vacuously hold. □

Proof of Prop. 1.5. Part 1: The proof has two parts. In the first part I suppose that a constitution C is compatible with some transitive relation $\geq$, and show that C satisfies all four transitivity requirements and the requirements R2, R3, R4 in Th. 1.1.

Consider any constitution C which is compatible with some weak preference transitive relation $\geq$. By the results of Th. 1.1, C satisfies the requirements R2, R3, R4. I now show that C satisfies all four transitivity requirements. I start with the first requirement.

Let $(x,y,\succ)$ and $(y,z,\succ)$ be in C. By compatibility of C, $x > y$ and $y > z$. By Def. 1.11, $x \geq y$ and not $y \geq x$ and $y \geq z$ and not $z \geq y$. By transitivity of $\geq$, (a) $x \geq z$ and (b) not $z \geq x$. Therefore, $x > z$. Thus, by compatibility of C with it, $(x,z,\succ)$ is in C. Hence C satisfies transitivity of strict preference.

By following this method, I can prove that C satisfies the rest of the requirements.

Part 2: In the second part, I suppose that for all $x,y,z \in X$, a constitution C satisfies all four transitivity requirements and the requirements R2, R3, R4 in Th. 1.1. By Th. 1.1, C is compatible with some transitive weak preference relation $\geq$, i.e. there exists a weak preference relation $\geq$ on X. Suppose $x \geq y$ and $y \geq z$. There are four possible cases: (1) $x > y$ and $y > z$, (2) $x > y$ and $y \equiv z$, (3) $x \equiv y$ and $y > z$, (4) $x \equiv y$ and $y \equiv z$.

Suppose that $x > y$ and $y > z$. By compatibility of C with $\geq$, then $(x,y,\succ) \in C$ and $(y,z,\succ) \in C$. Then by the first requirement, $(x,z,\succ) \in C$. So $x \geq z$.

By following this method, I show that in the next three cases a constitution that satisfies all four transitivity requirements is compatible with the transitive weak preference relation $x \geq y$, $y \geq z$, and $x \geq z$. □

Proof of Prop. 1.6. Part 1: Suppose that a constitution C is compatible with some complete weak preference relation. By the results of Th. 1.1, C satisfies the requirements R2, R3, R4. I now prove that for all $x, y \in X$, C satisfies the completeness of preferences requirement. Let $\geq$ be the relation that C is compatible with. Because $\geq$ is complete, for any $x, y \in X$, either $x \geq y$ or $y \geq x$. By Def. 1.11, either $x > y$ or $y > x$ or $x \equiv y$. By compatibility of C with $\geq$, either $(x, y, \succ) \in C$ or $(x, y, \succ) \in C$ or $(x, y, \sim) \in C$. Thus completeness of preferences.

Part 2: I now prove the "only if" part. I suppose that for all $x, y \in X$, C satisfies the requirements R2, R3, R4 in Th. 1.1 and completeness of preferences, and prove that C is compatible with some complete weak preference relation.

Since C satisfies the requirements R2, R3, R4 in Th. 1.1, C is compatible with some weak preference relation $\geq$. By completeness of preferences, it is the case that either $(x, y, \succ) \in C$ or $(y, x, \succ) \in C$ or $(x, y, \sim) \in C$. By compatibility of C, then either $x > y$ or $y > x$ or $x \equiv y$. By Def. 1.11, either $x \geq y$ and not $y \geq x$ or $y \geq x$ and not $x \geq y$ or $x \geq y$ and $y \geq x$. Thus either $x \geq y$ or $y \geq x$, i.e. the relation is complete. □

Proof of Corollary 1.1. Part 1: Suppose that a constitution C is compatible with some complete weak preference relation. By the results of Th. 1.1, C satisfies the requirements R2, R3, R4, and by the results of Prop. 1.6, C satisfies the requirement R5. I now prove that for all $x \in X$, C satisfies the reflexivity of indifference requirement. Let $\geq$ be the relation that C is compatible with. Because $\geq$ is reflexive, for any $x \in X$, $x \geq x$. By R5, either $(x, x, \succ) \in C$ or $(x, x, \sim) \in C$. By R_1, $(x, x, \succ) \notin C$. Thus $(x, x, \sim) \in C$, i.e. reflexivity of indifference.

Part 2: I now prove the "only if" part. I suppose that for all $x \in X$, C satisfies the requirements R2, R3, R4 in Th. 1.1, R5 of Prop. 1.6, and reflexivity of indifference, and prove that C is compatible with some reflexive and complete weak preference relation.

Since C satisfies the requirements R2, R3, R4 in Th. 1.1, C is compatible with some weak preference relation $\geq$. By completeness of preferences and by reflexivity of indifference, it is the case that $(x, x, \sim) \in C$. By compatibility of C, then $x \equiv x$. By Def. 1.11, $x \geq x$, i.e. the relation is complete. □

Proof of Prop. 1.7. Part 1: Suppose that a constitution C is compatible with some complete and transitive weak preference relation $\geq$. I prove that (i) for all $x, y \in X$,

C satisfies the requirements R2, R3, R4 in Th. 1.1 and R5 of Prop. 1.6, and (ii) for all $x, y, z \in X$ it satisfies the negative transitivity requirement above.

(i) By the results of Th. 1.1 and Prop. 1.6, for all $x, y \in X$, C satisfies the requirements R2, R3, R4 and R5.

(ii) Suppose that neither $(y, x, \succ)$ nor $(z, y, \succ)$ are in C. By compatibility of C, not $y > x$ and not $z > y$. By completeness of $\geq$, $x \geq y$ and $y \geq z$. By transitivity of $\geq$, $x \geq z$. By Def. 1.11, either $x > z$ or $x \equiv z$. By compatibility of C with it, either $(x, z, \succ)$ is in C or $(x, z, \sim)$ is in C. By R2, R3, R4 in Th. 1.1, $(z, x, \succ) \notin C$. By asymmetry, if $(x, z, \succ) \in C$ then $(z, x, \succ) \notin C$. By symmetry, if $(x, z, \sim) \in C$ then $(z, x, \sim) \in C$. And by incompatibility, if $(x, z, \sim) \in C$ then $(x, z, \succ) \notin C$, and if $(z, x, \sim) \in C$ then $(z, x, \succ) \notin C$. Hence $(z, x, \succ) \notin C$. This is negative transitivity.

Part 2: I now prove the "only if" part. I suppose that for all $x, y \in X$, C satisfies the requirements R2, R3, R4, R5 and R1+, and prove that C is compatible with some complete and transitive weak preference relation $\geq$.

By Prop. 1.5 and 1.6, there is some complete relation $\geq$ such that C is compatible with it. I now show that $\geq$ is transitive. Suppose $x \geq y$ and $y \geq z$. By Def. 1.11, either $x > y$ or $x \equiv y$, and either $y > z$ or $y \equiv z$. By compatibility of C with it, either $(x, y, \succ) \in C$ or $(x, y, \sim) \in C$, and either $(y, z, \succ) \in C$ or $(y, z, \sim) \in C$. By the requirements R2, R3, R4, $(y, x, \succ) \notin C$ and $(z, y, \succ) \notin C$. By negative transitivity, $(z, x, \succ) \notin C$. By completeness, the either $(x, z, \succ) \in C$ or $(x, z, \sim) \in C$. By compatibility of C, either $x > z$ or $x \equiv z$. By Def.1.11, $x \geq z$. Hence $\geq$ is transitive. $\square$

Chapter 2

A simple model of reasoning

2.1 Introduction

In Chapter 1, I have defined rational choice as a theory of rationality requirements on preferences, beliefs, and intentions and have described the different types of requirement they belong to. Chapter 2 focuses on *mental processes*, particularly the *mental process of reasoning* by which a person constructs his preferences, beliefs, and intentions. This chapter offers a novel definition of the mental process of reasoning and addresses the main question of this thesis: whether and in what ways one can come to satisfy rational choice requirements by that process. The model is conceptually inspired by Broome's rigorous analysis of reasoning. For Broome, reasoning is a *conscious*, *explicit* or *verbal*, and *rule-guided* mental process with *propositional attitudes* that is distinct from mental processes which are subconscious and automatic. His analysis is very well received among philosophers but is almost unknown to economists and behavioural and cognitive scientists.

Behavioural and cognitive sciences have also suggested the distinction of the mental processes between the slow, conscious, explicit or verbal, and rule-guided processes of "System 2" and the automatic, subpersonal, non-verbal, and associative processes of "System 1" (in particular, Wason and Evans 1974 and Kahneman 2003b, 2003a, and 2011). One central difference between the two systems is that System 2 processes are conducted in a language while System 1 processes, operating on experiences and feelings, are not. In Kahneman's words,

> "the perceptual system and the intuitive operations of System 1 generate impressions of the attributes of objects of perception and thought. These impressions are not voluntary and need not be verbally explicit. In contrast, judgments are always explicit and intentional, whether or not they are overtly expressed. Thus, System 2 is involved in all judgments,

whether they originate in impressions or in deliberate reasoning" (2003a, p. 1452).

My model can be regarded as an explicit description of the reasoning underlying System 2.[1] To capture the difference between the two systems: the conscious, verbal and rule-guided System 2 and the automatic, non-verbal, and associative System 1, I introduce the notion of *propositional attitudes* or mental states as I have defined them in Chapter 1. System 2 mode is propositional – it perceives the external world by describing it syntactically – and System 1 mode is automatic. When the decision maker enters the System 2 mode she processes propositions forming *explicit* attitudes towards them. When she enters the System 1 mode she responds in a non-verbal way to stimuli with her feelings and impressions that inform her explicit attitudes.

Drawing on the classification of the formal structure of requirements developed in Chapter 1, I show in five theorems that System 2 can achieve certain types of requirement, but not all of them. System 2 (i) is always capable of achieving closedness requirements (e.g. transitivity), (ii) cannot achieve consistency requirements (e.g. non-contradiction of preferences), and (iii) can achieve (a) Completeness (e.g. completeness of preferences), or (b) negative closedness requirements (e.g. negative transitivity), or (c) Conditional Completeness (e.g. Economic Enkrasia which is a weak version of WARP with intention rather than choices), but usually only at the cost of creating inconsistencies in the theory; and so, give a partially negative answer to the question this thesis is about.

This model offers an alternative to existing 'System 1/ System 2' models in cognitive sciences and behavioural economics. Behavioural economists often assume that the mental processes by which people achieve the rationality requirements of economic theory involve a "flawless" System 2, and attribute the inability to construct "rational and context-independent" preferences in accordance with these requirements mainly to imperfections of System 1 which cuts in on System 2 reasoning. The theorems, on the contrary, show that limitations of reasoning can also involve System 2. A possible interpretation is that the mental processes (if they exist at all) by which people achieve the rationality requirements of economic theory would involve System 1 as well as System 2.

The chapter proceeds in the usual order: Section 2 presents the main ingredients of the model and some examples of how it works. Subsection 2.4 reviews Broome's philosophical analysis of reasoning and explains how it can be mapped into the

[1] Boghossian (2014) has also related his own philosophical account of theoretical reasoning with existing literature on 'System 1/ System 2'.

model. Section 3 presents the results that answer the main question. Section 4 relates my formal analysis to existing literature, particularly 'System 1/ System 2' models in cognitive sciences and behavioural economics. Section 5 concludes. Proofs omitted from the main text are in an appendix.

2.2 The model

I model a decision maker who has two modes of reasoning: System 1 and System 2 mode. System 2 is conscious, verbal, and rule-guided. System 1 is automatic, non-verbal, and associative. System 1 processes can be described as 'pre-reasoning' or 'codification' by which System 1 impressions are 'codified' as System 2. In the spirit of Broome, I do not try to represent 'pre-reasoning' but to distinguish between 'pre-reasoning', which is non-propositional, and 'reasoning', which is propositional.

2.2.1 Definitions

Consider a non-empty set $\mathcal{L}$ of *propositions* and a non-empty finite set $\mathcal{A}$ of *types of attitude* such as a preference, a belief, or an intention. Let X be a fixed non-empty set of *mutually exclusive* choice options, e.g. goods or political candidates or career plans. Certain options from X are feasible; they form the feasible set, formally a non-empty subset $Y \subseteq X$ from which the agent chooses one element. Set $\mathcal{L} = X \cup 2^X \setminus \{\emptyset\}$ contains all relevant propositions for choice. The decision maker is related to propositions through different attitudes. She has certain preferences between options from X and beliefs that certain options from X are feasible. I assume that the decision maker can only form "simple" attitudes. Weak preferences are composite attitudes; the weak preference "x is weakly preferred to y" is the disjunction of two attitudes; "x is strictly preferred to y" or "x is indifferent to y". The first is a strict preference and the second is an indifference. Her reasoning results in certain choices. Choices are not types of attitudes. Choices are the output of some thinking. Intentions describe thinking prior to any choice. In conventional choice theory, preference is the only type of psychological attitude which is modelled. Preferences are revealed in choices as intentions are absent. Moreover, feasibility beliefs are absent from standard behavioural models, because of the background assumption that the feasible set is automatically known. I choose to describe explicitly the agent's psychology and enrich this framework with beliefs and intentions. The set $\mathcal{A} = \{\succ, \sim, bel, int\}$ contains all relevant *types of attitude* for choice. Strict preferences and indifference attitudes are denoted $\succ$ and $\sim$ respectively, intentions are denoted *int* and feasibility beliefs are denoted *bel*.

A **propositional attitude**, more simply, an **attitude** or a **mental state** is an attitude type a in $\mathcal{A}$ towards some propositions $p_1, p_2, ..., p_k$ in $\mathcal{L}$, a tuple $(p_1, p_2, ..., p_k, a)$. Every attitude type a in $\mathcal{A}$ comes with k number of places of a and a non-empty domain $D_a \subseteq \mathcal{L}$ of propositions. The number of places and the domain tell us that the type of attitude applies to combinations $p_1, p_2, ..., p_k$ of k propositions that belong to D_a. For example, attitude $(x, y, \succ)$ is a preference that applies to pairs of propositions that belong to the domain of preferences $D_\succ = X$. Denote by $\mathcal{M}$ the set of all attitudes. The pair $(\mathcal{L}, \mathcal{A})$ denotes a mental structure. An agent holds in her mind a subset $C \subseteq \mathcal{M}$ of attitudes I call her **mental constitution**. The constitution forms an explicit representation of the agent's psychology at a single point in time. I write $\mathcal{C} = 2^{\mathcal{M}}$ to denote the set of all constitutions.

System 1 which processes experiences and feelings is the main source of the System 2 initial attitudes which together form the initial constitution C_0, the inputs of System 2. I write $\mathcal{M}_0 \subseteq \mathcal{M}$ to denote the set of all System 1 attitudes. I write $\mathcal{C}_0 = 2^{\mathcal{M}_0}$ to denote the set of constitutions made out of attitudes in $\mathcal{M}_0$. These are attitudes that can originate in System 1 impressions and perceptions. System 2 can 'codify' them as attitudes. Set $\mathcal{M}_0 \subseteq \mathcal{M}$ contains all impressions and perceptions which can be codified as attitudes. For example, the perception of sunshine (System 1) is codified as the belief that it is sunny (System 2). Other examples of 'codification' or 'pre-reasoning' include the formation of habits which can be codified as the intention to repeat past behaviour, or the use of analogies which can be codified as the beliefs that one decision situation is the same as another one.

System 2 can change the initial constitution C_0 through rules that create new attitudes such as new preferences, beliefs, and intentions from the existing ones in C_0. Formally, a **rule** is an ordered pair (M, m) where M is the set of *premise attitudes* and m is the *conclusion attitude* which is *deduced from* M. The agent can *revise* a constitution C through applying rules from $\mathcal{S}$ to C. The **revision** of C *through* a rule $s = (M, m)$, denoted $C \mid s$, is given by: either $C \mid s = C \cup \{m\}$ if $M \subseteq C$ or $C \mid s = C$ if $M \nsubseteq C$. Informally, a rule applies to add a new attitude to a constitution. Such rules are restrictive in two ways. First, they create rather than remove attitudes; for instance, no rule removes the preference $(x, y, \succ)$ based on the premise $(y, x, \succ)$. Second, premises of rules are attitudes rather than absences of attitudes; for instance, no rule forms a preference based on the absence of other preferences. Just as requirements, rules typically come in schemas. So, one might alternatively define a rule as a schema (set) of pairs (M, m), where these pairs are the instances of the rule. I adopt the current more convenient terminology in which

schemas of rules become rules and rules become instances of rules.

I represent the System 2 mode of reasoning by a set $\mathcal{S}$ of rules (M, m). Any set $\mathcal{S}$ of rules is called a **System 2**, interpreted as a theoretically possible specification of System 2 reasoning. If nothing can be added to C by applying rules from the set $\mathcal{S}$, then a constitution C is closed under $\mathcal{S}$. That is, C is **closed under** $\mathcal{S}$ if for all s in $\mathcal{S}$, $C|s = C$. I call the so-reached new constitution the *closure* of C through $\mathcal{S}$. So the **closure** of C through $\mathcal{S}$ is the constitution $C|\mathcal{S}$ obtained from C by applying rules from $\mathcal{S}$ until the constitution reached cannot be changed by any of these rules. Formally, $C|\mathcal{S}$ is the minimal extension of C closed under $\mathcal{S}$. I offer a more complete definition of $C|\mathcal{S}$ in a separate appendix.

The decision maker enters the System 1 or the System 2 mode of reasoning at a time as many times as she wants to achieve any of the theory's requirements. Any **requirement** is written as a set of constitutions $R \subseteq \mathcal{C}$. A **theory** of rationality requirements is a requirement, denoted $\mathcal{T} \subseteq \mathcal{C}$.

What is a 'good' System 2, given a theory of rationality? I postulate two desirable conditions.

1. *Desirable condition 1*: System 2 should *achieve* that requirement.

2. *Desirable condition 2*: In so doing, System 2 should *preserve consistency* by not creating inconsistencies in the theory.

The following definitions make the two conditions precise:

Definition 2.1. A System 2 **achieves** a requirement R if for each constitution C, its revision $C \mid \mathcal{S}$ satisfies R.

The next two definitions clarify the Desirable condition 2:

Definition 2.2. Given a theory $\mathcal{T}$, a constitution C is **consistent** if its attitudes can be rationally held together, i.e. if some constitution $C' \in \mathcal{T}$ includes C.

How does consistency of a constitution relate to the notion of consistency requirements from Chapter 1? In Dietrich et al. (2018) we show that a constitution is consistent if and only if it satisfies all consistency requirements of the theory.

Definition 2.3. Given a theory $\mathcal{T}$, a System 2 **preserves** consistency if for every consistent constitution C, the revised constitution $C \mid \mathcal{S}$ is still consistent.

2.2.2 Examples

In Chapter 1, I describe formal rationality like rational choice in terms of rationality requirements. Then I set out an account of rational choice that consists of eight schemas of requirements. Let $\mathcal{T}$ be a set of constitutions which are deemed rational by the eight schemas of requirements. Consider a set of arbitrary rules $\mathcal{S}$ and a constitution C that satisfies schemas of requirements R2 asymmetry of $\succ$, R3 incompatibility of $\succ$ and $\sim$, R4 symmetry of $\sim$, and also R6 non-conflicting feasibility beliefs and R7 non-conflicting intentions. This section provides examples that show how the model of reasoning applies with respect to the rest of the requirements of $\mathcal{T}$.

Example 1

R1 Transitivity of weak preferences: For any $x, y, z \in X$,

if $(x, y, \succ) \in C$ or $(x, y, \sim) \in C$ and $(y, z, \succ) \in C$ or $(y, z, \sim) \in C$, then $(x, z, \succ) \in C$ or $(x, z, \sim) \in C$.

Conditional on asymmetry of $\succ$, symmetry of $\sim$, and incompatibility between $\succ$ and $\sim$, this schema of requirement is equivalent to four subschemas of requirements: strict preference transitivity, indifference transitivity, PI transitivity, and IP transitivity, all of which have a structure analogous to that of *"if... then..."* rules in $\mathcal{S}$. So, I can construct a System 2 that contains one rule to achieve each of these requirements. This demonstrates a fundamental truth about explicit reasoning which is a process by which the presence of certain attitudes causes the presence of another attitude. A process like this is well adapted to achieve requirements whose antecedent and consequent are that certain attitudes are present.

Unlike this example which shows that it is possible that a System 2 could achieve certain requirements of $\mathcal{T}$, the following three examples show *limitations* of System 2 to achieve certain other requirements of $\mathcal{T}$.

Example 2

R5 Completeness of weak preferences: For any $x, y \in X$,

either $(x, y, \succ) \in C$ or $(y, x, \succ) \in C$ or $(x, y, \sim) \in C$.

This schema of requirements requires that a preference or indifference between x and y exists. There is an easy way to achieve R5: fix an attitude and adopt the rule 'always form this attitude with an empty premise set'. There are three rules of

reasoning that can achieve completeness from the empty set: (i) $s = (\emptyset, (x, y, \succ))$, (ii) $s' = (\emptyset, (y, x, \succ))$, and (iii) $s'' = (\emptyset, (x, y, \sim))$. There are two problems with this.

First, from a psychological point of view, such arbitrary rules seem unjustified for System 2. System 2 operates on inputs generated by System 1 to form explicit attitudes such as any specific preference or indifference between options x and y. If, given the preferences and beliefs you actually hold, there is no way of reasoning towards any specific preference or indifference between options x and y, one might doubt that you are justified to hold some preference or indifference, while being silent about which. That is, if inputs are insufficient to allow conscious System 2 reasoning to arrive at particular types of conclusion, this limitation does not seem to be a fault of reasoning. As Gilboa et al. (2012) and Infante et al. (2016) have also argued, there can be cases in which it is not possible or desirable to compare propositions from X.

Second, from a theoretical point of view, designing such arbitrary rules that can achieve a completeness requirement often causes inconsistencies in the theory. Suppose you initially have no preference or indifference between x and y, violating R5, but you prefer y to another option z, and z to x. So your initial constitution is $C_0 = \{(y, z, \succ), (z, x, \succ), ...\}$, where '...' stands for other attitudes. Let System 2 contain the rule s, so that the revised constitution is $C|\mathcal{S} = \{(x, y, \succ), (y, z, \succ), (z, x, \succ), ...\}$. While $C|\mathcal{S}$ satisfies R5, it violates a preference-acyclicity requirement. Completeness of preferences has been achieved at the cost of creating inconsistencies in the theory.

The conclusion is that, by setting up the theory in terms of requirements alone, one does not deal with choice as the output of a process of reasoning. If, however, choice is viewed as the output of a process of reasoning, transitivity appears to be rational in a way that completeness does not. In the course of achieving transitivity, you create a new preference you did not originally have, given your other preferences. The way you achieve transitivity may actually be viewed as a process that helps you to find reasons to infer certain preference from others. If you prefer x to y, and you prefer y to z, then these two preferences give you a reason to prefer x to z. By contrast, completeness provides no help in finding reasons to rank elements of X, say x over y. Gilboa et al. (2012) make a similar point: they call the transitivity axiom a reasoning axiom and conclude that choice that is the output of a process of reasoning need not be complete.

Example 3

R1+ Negative Transitivity: For any $x, y, z \in X$,

if $(y, x, \succ) \notin C$ and $(z, y, \succ) \notin C$, then $(z, x, \succ) \notin C$.

Unlike transitivity which is formulated on preferences, this is a schema of requirements which is formulated on non-preferences. It is questionable whether you can conclude anything from the absence of certain preferences and particularly the absence of a specific preference. So from a psychological point of view, transitivity appears to be rational in a way that negative transitivity is not. Moreover, given the resources of my model, we simply cannot design a rule that concludes in the absence of an attitude. But we can construct a rule that concludes in the presence of a strict preference from the the presence of other strict preferences. Negative transitivity in contrapositive form tells us that for all options $x, y, z \in X$, if $(z, x, \succ) \in C$ then for any y, either $(z, y, \succ) \in C$ or $(y, x, \succ) \in C$. So we can find two rules that achieve negative transitivity: they are (i) $s = (\{(z, x, \succ)\}, (y, x, \succ))$ and (ii) $s' = (\{(z, x, \succ)\}, (z, y, \succ))$. But such rules are arbitrary and unjustified.

Suppose you initially prefer z to x and x to y and prefer neither y to x nor z to y violating negative transitivity. So your initial constitution $C_0 = \{(z, x, \succ), (x, y, \succ), ...\}$, where '...' stands for other attitudes, is consistent. Suppose that System 2 contains rule s. The revised constitution is $C|\mathcal{S} = \{(z, x, \succ), (x, y, \succ), (y, x, \succ), ...\}$. While $C|\mathcal{S}$ satisfies negative transitivity, it violates asymmetry of $\succ$. Negative transitivity has been achieved at the cost of creating inconsistencies in the theory.

A different approach is to assume that non-preferences suggest a particular interpretation of weak preferences as the complement of strict preferences. Given completeness, strict preferences are the only primitive attitudes and weak preferences are derived from the absence of strict preferences. In this case, negative transitivity of strict preferences can be reduced, given completeness of weak preferences, to (positive) transitivity of weak preferences. Note that this interpretation of weak preferences differs from the interpretation of weak preferences as a disjunction of weak preference and indifference attitudes. Moreover, it requires completeness that can be achieved but usually only though arbitrary rules.

Example 4

R8 Economic Enkrasia: This schema of requirements links intention to preference. For any feasible set $Y \subseteq X$ and $Z \subseteq Y$ such that $Z \neq \emptyset$,

1. if $(Y, bel) \in C$ and $(x, y, \succ) \in C$ for all $x \in Z$, $y \in Y \setminus Z$
2. and $(x, z, \sim) \in C$ for all distinct $x, z, \in Z$,
3. then $(x, int) \in C$ for some $x \in Z$.

This requirement reflects the classical preference-maximisation hypothesis resulting in intentions rather than choices. There are two main differences of analysing intentions prior to actions rather than choice acts directly (see for example the weak axiom of revealed preference (Samuelson, 1938)). First, although preferences range over all options, feasible or non-feasible, intentions relate only to what you believe to be the feasible set. Thus, there is no mental analogue to the choice-theoretic concept of a choice function. Second, even if there is more than one top-ranked option in your feasible set, you intend a specific one of them. By allowing the choice function to output non-singleton sets, conventional choice theory evades the question of how you choose between indifferent options. This point can be illustrated with the classic story of Buridan's ass.

In the story, the ass is exactly equidistant from two identical bales of hay, one to its right and one to its left. Let the ass face the feasible set $Y = \{l, r, s\}$, and have the initial constitution $C = \{(Y, bel), (l, r, \sim), (l, s, \succ), (r, s, \succ)\}$ where l, r, and s is 'left', 'right', and 'starve' respectively. Intuitively, the ass choosing to apply either of the rules $s = (C, (l, int))$ or $s' = (C, (r, int))$ gets to a bale of hay and survives. According to rational choice choosing either bale of hay when the ass could have chosen the other appears as if one bale of hay was no worse than the other.

The traditional interpretation of the story is that the ass was indifferent between the two bales of hay. Unable to decide between the two, it fails to form an intention for left or right and starves to death. Viewed through the model, the ass also fails to achieve economic enkrasia. If the ass applies either (or both) of the rules s and s', the ass can form an intention to go to one bale of hay, thereby achieving the economic enkrasia requirement. But this leads to conflicting intentions whenever (against the story) another intention was already present. So while $C|\mathcal{S}$ satisfies R8, it violates non-contradiction of intentions. By explicitly modelling rules that conclude in single attitudes rather than choice functions that output non-singleton sets, I point out a gap in the theory that does not tell us how you choose between indifferent options.

The second, less popular, interpretation of the story is that the ass did not have any preferences or indifferences over the two bales of hay (Sen, 1973). On this interpretation, the ass has failed to achieve completeness of preferences which I have discussed in example 2.

2.2.3 Possible solutions

I tried in the above examples to explore some of the possibilities and limitations of System 2 reasoning, given that System 2 is understood as a set of rules. Now I will

consider different ways in which the automatic processes of System 1 might help to solve these problems. For example, the automatic processes of System 1 might (i) cut in to economise the cognitive costs of ranking elements of X or (ii) activate a stopping rule to eliminate the surplus intentions.

Here is one possibility: the decision maker is assumed to be able to construct context-independent preferences between any relevant pair of options using System 2, but doing so is cognitively costly (e.g. a limited number of elements from X can be evaluated). To economise on these costs, she uses a System 1 heuristic which makes preferences over a shortlist of feasible options easy to retrieve. In terms of the model, the initial set of attitudes $\mathcal{C}_0$ is the outcome of a System 1 heuristic which reduces the complexity of the decision process. Then System 2 maximises a preference relation on a subset $X' \subseteq X$ of "choiceworthy" options, i.e. $\mathcal{L}' \subseteq \mathcal{L}$ where $\mathcal{L}' = X' \cup 2^{X'} \setminus \{\emptyset\}$. Manzini and Mariotti (2012) propose a "Categorise then choose" model in which options are subdivided into categories (e.g. if the set of options X is all cereals in the supermarket, one category would be based on sugar concentration). The decision maker 'simplifies' the problem by eliminating some subsets of the feasible set of options X that belong in least-preferred categories (e.g. sugar-free cereals dominate cereals with sugar). Then he picks the maximal element according to his preference among the surviving alternatives.

In a related model Dietrich and List (2016) assume that each option is described by the agent as a bundle of 'motivationally salient properties'. The agent maximises preferences over such property bundles. Although these preferences are context-independent, choice reversals happen because the context influences which properties are motivationally salient and hence how options are perceived.

An alternative approach is to model choice as the outcome of some mental process that does not (or does not need to) include the concept of preference. The agent might lack any preference or indifference between the options in question, a possibility defended informally by Infante et al. (2016). The automatic processes of System 1 then step in and lead to intentions towards propositions from X without these intentions being the result of some preference-maximization procedure. In terms of the model, the initial set of attitudes $\mathcal{C}_0$ is the outcome of some automatic process which results directly in intentions with no prior preferences linked to them. Simon's model of "satisficing" (1955) offers one such alternative. The agent processes decisions sequentially, and stops when an option is above some fixed reservation level. In example 4 the process stops when one of the attitudes (l, int) or (r, int) is reached, eliminating the surplus intention.

Another alternative is offered by models that introduce some sort of System

1 "thinking by analogies". Cerigioni (2017) presents a model that formalises the activation of automatic choices as the result of non-deliberative processes driven by (analogies with) past experiences. The paper focuses on priming, replicating past behaviour in familiar choice environment, as the main source of automatic choice. Choice is the result of a conscious process if the alternative chosen maximises a preference relation $\succ$ on X otherwise choice is the result of an automatic process that replicates past behaviour. In the language of the dual-system hypothesis, System 1 uses analogies to deal with the choice environment, and System 2 uses a preference relation to choose the top option among the available. Gilboa and Wang (2018) present a dual system model where the automatic processes of following habits and sticking with status quo decisions belong to System 1 and the conscious decision making processes belong to System 2. The model formalises a decision maker who sometimes is better off retaining the status quo rather than making conscious choice. This model shares similarities with Cerigioni's model in that the status quo can be viewed as making the same choice that has been made in similar cases in the past and is not the result of preference maximisation.

2.2.4 Philosophical and cognitive foundations of System 2

This section shows how John Broome's rigorous analysis of reasoning which has been the main theme of his book *"Rationality through Reasoning"* (2013) can be mapped into the model, and discusses some findings in cognitive sciences in support of it. Referring to the process of reasoning, Broome says,

> "Some requirements are too difficult for our automatic processes to cope with [...] when automatic processes let us down, our mortal rational disposition equips us with a further, self-help mechanism. We have another way of improving our score by our own efforts. We can do it through the mental activity of reasoning" (2013, p. 207)

What, according to Broome, distinguishes the mental process of reasoning from the subconscious and automatic mental processes is that it is:

1. Plausibly explicit.
2. Rule-governed.
3. A conscious act.

The first characteristic is that active reasoning is *plausibly explicit*. Explicit reasoning means that you express the content of the attitude to yourself using a *sentence*.

For example, saying to yourself *'It is raining'* is normally the way in which you express your belief that it is raining in language. The pair consisting of the attitude and the attitude's content is a *marked content*. Saying to yourself the marked contents of attitudes explains how you acquire a new attitude you initially did not have from existing attitudes. It is crucial in this account that not all the reasoning you can do is reasoning with beliefs – that not all attitudes are reducible to belief attitudes. For example, rational choice, as a theory of practical rationality, is primarily concerned with attitudes other than beliefs and sometimes with beliefs. So an agent's reasoning can go as follows: *'getting x and getting y are the only feasible options'*, *'Rather I get x than I get y'*, So, *'I shall get x'*. This links preferences, as the natural primitive input of the theory, and intention to choose, as the final output of those preferences. Section 2.1 proposes a simple way to describe these mental states as particular tuples of types of attitudes and their contents.

The second characteristic is that in reasoning you are guided by a *rule*. Broome considers the following example of reasoning to illustrate the sense in which you are guided by a rule: Imagine that you wake up one morning and hear dripping water. You come to believe that it is raining.[2] You recall that last night it was snowing. You combine this with the knowledge that if it is raining the snow will melt and conclude that the snow will melt (2013, pp. 216, 223). In this example, you initially believe two propositions: *'It is raining'* and *'If it is raining the snow will melt'*, and the come to believe that *'the snow will melt'*. A rule, *modus ponens*, allows you to create a new belief you initially did not have from the existing beliefs, saying to your self *'So the snow will melt'*. If reasoning is purely causal then you do not follow a rule; the rule causes something to you. So, what takes place in your mind is your following a rule. You follow a particular rule because it seems right to you; whether you follow correctly the rule or not; and whether the rule is correct or not. I have described these processes as ordered pairs that consist of a set of premise attitudes and the conclusion attitude.

The third characteristic, that reasoning is a *conscious act*, follows from the previous two. It is conscious because you are conscious of the content of the attitude you reason with. The contents you reason with are usually not about your attitudes. So it is a mental operation *with* and not *about* attitudes. In the rain example, you are conscious of the content of your belief attitudes. Moreover, this is an act because reasoning is something you do rather than something that happens to you. You acquire a belief you initially did not have (e.g. that the snow will melt) following

[2]This is an example of codification I have discussed earlier. The perception of hearing water dripping is codified as the belief that it is raining.

rules of reasoning.

Consider the following example: Imagine that you see a spider. You come to believe that there is spider in the room. Suppose you are in Australia where spiders are dangerous and that you want to stay safe. You combine your knowledge with your belief that there is spider in the room and conclude that you shall leave the room. Because reasoning allows you to 'codify' or 'organise' your perceptions and feelings as attitudes you might be able to describe your process that lead to your leaving the room as a process that looks like conscious reasoning: *'there is a spider in the room', 'spiders are dangerous', 'I would like to be safe', 'So, I shall leave the room'*. So many philosophers believe that automatic and subconscious processes involve propositions. See Sugden (2006) for a critical approach to this view.

But suppose that it was not the belief that there is a spider combined with your knowledge that spiders are dangerous and your desire to stay safe that caused the intention but a different mental process. One possibility is that it was fear that caused you to leave – that it was not your reasoning. This causal process is automatic and subconscious. It is a process that is not reasoning. Fear causes you to leave without thinking. Upon fear, you act and do not intend. Intending to leave is not leaving the room. Reasoning is something you do rather than something that happens to you. Codifying and organising our perceptions in terms of beliefs and intentions in the spider example is one of many ways we have to explain to ourselves what caused a particular behaviour. It is a property of our language that these subconscious processes can be redescribed in terms of propositional attitudes as if we were reasoning with them. But this might not be the process in itself – e.g. the process that caused you to leave the room.[3]

Experimental studies in self-justification (for example, Nisbett and Wilson 1977 and Wilson and Bar-Anan 2008) suggest that people who have previously observed their own behaviour are good at coming up with *reasons* that can justify their behaviour to themselves, but relatively poor at finding the "actual" mental process that led to this behaviour. In the spider example, the rationalisation that you acted on the belief that spiders are dangerous is false. Your 'feeling fear' is an empirically accurate description of your state of mind. So if we were to redescribe this state of mind with marked contents, the pair (spider, fear) would be an empirically accurate description of the state of mind you were in.

For Broome, the three characteristics I have described above capture the intuitive notion of everyday human reasoning which is a *conscious act* conducted in a *language*

[3]In Broome's terminology, these automatic processes are treated in the jogging account of reasoning (2013, pp. 225-227).

through which you give rise to new attitudes from existing ones following a reasoning *rule*. So Broome's account, naturally, explores the relations of consequence that hold between attitudes and not between propositions or sentences as it is usually the case in classical and modal logics (2013, p. 254). Language is usually the means you have to express an attitude of yours by bringing in your mind the proposition together with its type of attitude. If reasoning is conducted in a language, then according to Broome,

> "we can only reason about marked contents if we have markers in our language to designate them [and] [t]his means that the reasoning we can do is limited by the contingencies of our language"(Broome, 2010, p. 76)

For example, by saying 'I do not prefer the fruit to the cake' you do not express a non-preference as there is no English markers to express non-preferences but there are markeres to express beliefs, in particular the belief in the absence of a preference. Negative attitudes such as nonbeliefs or nonpreferences and disbeliefs or dispreferences have no role in explicit reasoning as language has no way of expressing them. In reasoning explicitly with your attitudes you cannot conclude the absence of an attitude (2013, p. 278).

I am aware of no formal model (of reasoning) that explores this possibility as limitations of a System 2. The following section explores one way. I prove five theorems that show how far a decision maker can go when reasoning explicitly with her attitudes to achieve rationality requirements. Although I am primarily interested in rational choice requirements these results apply to requirements in general.

2.3 A characterisation of System 2

While Section 2.3 presents some of the limitations of System 2 to achieve rational choice requirements, Section 3 shows what requirements, in general, System 2 can or cannot achieve. I am first going to need a language for requirements that roughly matches the language in which the reasoning rules are described. Most requirements can be written in some *'If ... then ...'* form. So I can roughly match requirements with reasoning rules. In chapter 1 I have identified five requirement types that have this form. I briefly restate them:

1. A requirement R is a completeness requirement if it is of the form 'if absence of all elements of $M \setminus \{m\}$ then presence of m' with respect to a set of mental states $M \subseteq \mathcal{M}$ with $M \neq \emptyset$ and $m \in M$.

2. A requirement R is a consistency requirement if it is of the form 'if presence of all elements of $M \setminus \{m\}$ then absence of m' with respect to a set of mental states $M \subseteq \mathcal{M}$ with $M \neq \emptyset$ and $m \in M$.

3. A requirement R is a closedness requirement if it is of the form 'if presence of all elements of M then presence of m' with respect to a set of mental states $M \subset \mathcal{M}$ with $M \neq \emptyset$ and a mental state $m \in \mathcal{M}$ with $m \notin M$.

4. A requirement R is a negative closedness requirement if it is of the form 'if absence of all elements of M then absence of m' (equivalently, 'if presence of m then presence of at least one element of M') with respect to a set of mental states $M \subset \mathcal{M}$ with $M \neq \emptyset$ and a mental state $m \in \mathcal{M}$ with $m \notin M$.

5. A requirement R is a conditional completeness requirement if it is of the form 'if presence of all elements of M then presence of at least one element of N' with respect to a pair of sets of mental states $M, N \subseteq \mathcal{M}$ with $M \cap N = \emptyset$ and $M, N \neq \emptyset$; formally: $M \subseteq C \Rightarrow N \cap C \neq \emptyset$.

The words "if ... then ..." and "at least one" in the brackets all belong in the language in which the requirement is described, and not the language in which you are thinking and making choices.[4]

Notice two things: First, completeness, consistency, closedness, and negative closedness do not overlap except for the special case that negative closedness is with respect to some singleton set $M = \{m'\}$, in which case negative closedness and closedeness overlap. Conditional completeness does not overlap with any other requirement type given that it is defined with respect to pairs of distinct non-singleton sets M, N. Second, given the first observation, each requirement from R1 to R8 has a single corresponding type of requirement. I can now address my central question: **Can System 2 achieve full rationality? How far can one become rational with System 2?** Five theorems show the (im)possibility to achieve particular types of rationality requirements whilst preserving consistency. They roughly state the following:

1. System 2 can achieve Closedness requirements while also preserving consistency.

2. System 2 cannot achieve Consistency requirements.

[4]It is common to formally represent theories of rationality as propositions in some language.

3. System 2 can achieve (a) Completeness, or (b) Negative Closedness, or (c) Conditional Completeness requirements, but usually only at the cost of creating inconsistencies.

2.3.1 Can System 2 achieve closedeness requirements?

Theorem 2.1. *Given any theory $\mathcal{T}$, there is a system 2 which achieves each closedness requirement of $\mathcal{T}$ and preserves consistency.*

This theorem formalises the fundamental truth that explicit reasoning is well adapted to achieving closedness requirements. The proof establishes this by constructing a System 2 that contains one rule for each closedness requirement of $\mathcal{T}$. As a result of this, each closedness requirement of $\mathcal{T}$ is achieved in a single reasoning step. So $\mathcal{S}$ may contain a huge number of rules. But not all closedness requirements need have a tailor-made rule in $\mathcal{S}$. Often there exists a much smaller $\mathcal{S}$ which still achieves every closedness requirements (and preserves consistency) in more than one reasoning step.

2.3.2 Can System 2 achieve consistency requirements?

Theorem 2.2. *No system 2 achieves any consistency requirement.*

This theorem formalises the idea that in explicit reasoning you cannot conclude the absence of an attitude – that you can only add but cannot remove attitudes. Non-contradiction of preferences for example, concludes the absence of the preference for x over y given your existing preference for y over x.

2.3.3 Can System 2 achieve completeness requirements?

This theorem will draw on the following notion:

Definition 2.4. Given a theory $\mathcal{T}$, an attitude m is **falsifiable** if some consistent constitution C becomes inconsistent after adding m.

In plausible theories of rationality, almost all attitudes are falsifiable. The attitude $(x, y, \succ)$ in Example 2 is falsifiable because it rules out $(y, x, \succ)$ and $(x, y, \sim)$, and the combination of $(y, z, \succ)$ and $(z, x, \succ)$.

Theorem 2.3. *Given any theory $\mathcal{T} \neq \emptyset$,*

1. *some system 2 achieves all completeness requirements of $\mathcal{T}$, but*
2. *no system 2 that preserves consistency achieves any completeness requirement of $\mathcal{T}$ that is given by a set M whose members are falsifiable.*

2.3.4 Can System 2 achieve negative closedness requirements?

As one can check the negative closedness requirements which are defined with respect to a singleton M and attitude m are logically equivalent to closedeness requirements with respect to a singleton M and attitude m. So a system 2 can achieve some special cases of negative closedness requirements of $\mathcal{T}$.

The following theorem about negative closedness is restricted to cases where M is not a singleton (e.g. negative transitivity). This theorem will draw on the notion of "conditional falsifiability":

Definition 2.5. Given any theory of rationality $\mathcal{T}$ and any set $M \subseteq N$ of attitudes, an attitude $m \in N \setminus M$ is **falsifiable given** M if some consistent constitution that includes M becomes inconsistent through adding m.

For negative closedness, an attitude $m \in N \setminus M$ is falsifiable given a singleton $\{m'\}$. The attitude $(y, x, \succ)$ in Example 3 is falsifiable given singleton $\{(z, x, \succ)\}$ because it rules out $(x, y, \succ)$.

Theorem 2.4. *Given any theory $\mathcal{T}$,*

1. *some system 2 achieves all negative closedness requirements of $\mathcal{T}$, but*
2. *no system 2 that preserves consistency achieves any negative closedness requirement whose each attitude in M is falsifiable given $\{m'\}$.*

2.3.5 Can System 2 achieve conditional completeness requirements?

This theorem will also draw on the notion of "conditional falsifiability". For conditional completeness, an attitude $m \in N \setminus M$ is falsifiable given set M. The attitude (l, int) in Example 4 is falsifiable given C because it rules out (r, int).

The notion of falsifiability or conditional falsifiability arises when a requirement contains disjuncts. Completeness, negative closedness, and conditional completeness are the only types of requirement that can contain disjuncts. Examples 2, 3, and 4 are examples of the above types of requirement that contain disjuncts. Completeness of preferences requires that at least one of the three disjuncts $(x, y, \succ)$, $(y, x, \succ)$ and $(x, y, \sim)$ is in the constitution. Negative transitivity requires that at least one of the two disjuncts $(y, x, \succ)$ and $(z, y, \succ)$ is in the constitution. Economic enkrasia requires that at least one of the two disjuncts (x, int) and (z, int) is in the constitution.

In each of these cases, falsifiability or conditional falsifiability arises because each disjunction has two disjuncts, and there is a rule for each disjunct. In example 2, any one of the rules $s = (\emptyset, (x, y, \succ))$, or $s' = (\emptyset, (y, x, \succ))$, or $s'' = (\emptyset, (x, y, \sim))$ can achieve completeness of weak preferences, a completeness requirement. In example 3, either of the rules $s = (\{(z, x, \succ)\}, (y, x, \succ))$, or $s' = (\{(z, x, \succ)\}, (z, y, \succ))$ can achieve negative transitivity, a negative closedness requirement. And in example 4 either of the rules $s = (C, (l, int))$, or $s' = (C, (r, int))$ can achieve economic enkrasia which is a conditional completeness requirement. So it is possible to achieve either completeness, or negative closedness, or conditional completeness by applying at least one of these rules from $\mathcal{S}$. But because it is possible to achieve these types of requirement by applying many of these rules, doing so can result in the conclusion of a certain attitude m that is (conditionally) falsifiable.

The next theorem formalises the idea that reasoning explicitly towards an "either... or..." conclusion can result in a certain attitude m that is (conditionally) falsifiable. So Theorems 2.3 and 2.4 are special cases of this theorem. It would be convenient for the proof of this theorem to consider a single type of requirement that generalises completeness, negative closedness, and conditional completeness types of requirement. This requirement can be written as a conditional-completeness-like type of requirement. I call this type of requirement a *conditional completeness** requirement. Formally, a requirement R is a **conditional completeness*** requirement if there is a pair of sets of mental states $M, N \subseteq \mathcal{M}$ with $M \cap N = \emptyset$ and $N \neq \emptyset$ such that if $M \subseteq C$ then $\exists m \in N : m \in C$.[5]

Theorem 2.5. *Given any theory $\mathcal{T}$,*

1. *some system 2 achieves all conditional completeness* requirements of $\mathcal{T}$, but*
2. *no system 2 that preserves consistency achieves any conditional completeness* requirement whose attitudes in N are all falsifiable given M.*

The negative findings in part (2) of the theorems concern the notions of (conditional) falifiability. Note the special cases of completeness with respect to a singleton $M = \{m\}$, and conditional completeness (negative closedness) with respect to singleton sets $M = \{m\}, N = \{m'\}$ (singleton set $M = \{m'\}$). For completeness, the attitude m is by definition non-falsifiable because any plausible theory of rationality that is consistent would already contain m. For conditional completeness (negative closedness), attitude m' given attitude m in M (attitude m given attitude

[5]In fact, Chapter 1 identifies a single type of requirement that unifies all types of requirement considered above, the *unified requirement*. The unified requirement is written as a conditional-completeness-like type of requirement imposing $M \cup N \neq \emptyset$ rather than $N \neq \emptyset$.

m' in M) is by definition non-falsifiable because any plausible theory of rationality that is consistent would already contain m' given M.

Summing up, Theorems 2-5 show that there are formal limitations of a System 2 to achieve certain types of requirement. The limitations are related solely to the formal structure of requirements and reasoning and not to any account of what qualifies as correct reasoning. This conclusion however hinges on accepting the rule-following and language-based account of System 2 I have offered, which assumes that you cannot start from the *absence* of an attitude, a thesis that Kolodny agrees with (2005, pp. 527-8), or conclude in the *absence* of an attitude, a thesis that Broome agrees with (2013, p. 278).

2.4 Discussion

This section relates the formal analysis to existing 'System 1/ System 2' models in cognitive sciences and behavioural economics, particularly the dual self model. Since Wason and Evans 1974 and Kahneman 2003b, 2003a, and 2011, a growing number of behavioural economic models have incorporated the dual-system hypothesis to distinguish the choice which is prone to contextual cues and is made by the fast System 1 from the choice which is not affected by environmental cues and is made by the slow System 2. Neuroeconomically inspired models have gone as far as to use the dual-system hypothesis as a useful analogy to describe the neurophysiology of the brain. My intention is to identify the main modelling approaches that adopt the dual-system hypothesis in relation to the thesis's main question, and not to write a comprehensive literature review.

I explore how these three approaches answer my question with the means of an example. There are two options x and y that can be faced in two alternative contexts, K and K'. I focus on contexts that have been the subject of experimental studies of dynamic choice that have inspired dual process economic models of choice. A typical dynamic choice experiment, an experiment that studies decision-making over time, involves choices between outcomes at different points in time; for example, the choice between a monetary prize p today and a bigger monetary prize q at a later point in time made today and the same choice made at an earlier point in time, or the choice between a fruit and a cake for dinner dessert made in the morning and the same choice made at the dinner time. In relation to my choice example above, the two options x and y that can be faced are the fruit and the cake or prize p and prize q, and the two alternative contexts, K and K', is the the time in which the choice is made.

Dynamically consistent rational choice requires that choices planned for a given future point in time do not differ from the actual choices made at that time. If (fruit, this afternoon) trumps (cake, this afternoon) when the options are compared in the morning, then (fruit, this afternoon) trumps (cake, this afternoon) when the options are compared in the afternoon. So for rational choice, these contexts include information that is irrelevant for the evaluation of the two options. For all aspects that are relevant for ranking x and y, x chosen in context K is identical to x chosen in K', and the same is true for y.

However, we observe that the agent chooses x in K and y in K'. Experiments will typically show that, contrary to what this principle requires, people do not make stable choices; they do not stick to their plans (see for example, Ainslie 1992, Horowitz 1991, Loewenstein 1988, and Loewenstein and Prelec 1992). They will plan to have the fruit but will order the cake at dinner; or they will prefer to get the smaller prize on 1 September rather than get the bigger prize on 8 September if they compare the options on 1 September, but will prefer to get the bigger prize on 8 September if they compare the options on 1 August. In short, people have different attitudes towards short-run and long-run payments; they will be more sensitive towards the time of payment when payments are in the present or near future, and will become more sensitive towards the size of the gain when payments are in the more distant future. And they will focus on different attributes (calories, sugar, price) depending on how far ahead the time of choice is; so, they will have different attitudes towards short-run and long-run options. So it is uncontroversial that such patterns of behaviour violate the rationality requirements imposed by standard choice theory. But is this evidence of failure of System 2 or System 1?

I investigate three possible approaches to answering this question that all, in one way or another, incorporate some sort of dual-process model of reasoning for choice. The first two modelling approaches are known in the economics literature as "dual self models" and have been (naturally) used to study time preferences and as the example above. To explain deviations from economic theory, these models have enriched the conventional model of choice with additional information regarding the frame or the context in which choice is made, and have assigned one self the ability to make decisions like an economist and made the other self or selves prone to contextual cues. The idea of a model in which two or more selves interact with each other dates back at least to Strotz (1955). Strotz's model explored how an agent who experiences changes in preferences during a course of action would maximise expected utility. To answer this question, Strotz, introduces different levels of sophistication breaking down the maximisation problem to "smaller" ones each cor-

responding to one self. There are two types of Strotzian models: one that assumes that selves are cooperative and the other that they are conflicting.

2.4.1 The cooperative selves approach

McClennen (1990) was one of the first to offer a cooperative selves model. There is a single continuing or atemporal self who lives and makes a sequence of decisions at different points in time. Temporal selves live at different points in time and have their own goals and desires. The continuing self does what is best for her devising a plan of action. Each of these temporal selves anticipates what the next self wants and knows what the continuing self also wants. She will do what is best for all selves provided that the other self will do what is best for all selves. So each temporal self can contribute to the resolution of the plan minimising the cost of deviating from it. In this framework, *being resolute* is a form of being rational. Formally the decision maker, the continuing self, plays a coordination game working backwards through each period. Temporal selves contribute to the execution of the plan assuming perfect knowledge.

In relation to my example above, the agent is assumed to have and is always conscious of a context-independent preference in favour of one of the options, say x. Knowing that in context K' and not in K, as it is the case of the examples above, there are psychological 'cues' that activate a temptation to take y instead of the optimal x, the agent devises a plan of action to choose x avoiding these cues. Preferences and choices that do not conform to the optimal plan are excluded as non-pertinent. So on this approach, rationality will be assigned to one self, the continuing self. Temporal selves reach a resolution employing some sort of team reasoning which is now assumed to be some sort of System 2 thinking.

The benefit this approach has is that it allows us to think of dynamic choice without assuming the deviations from decision theory are lack of self-control. Since there is one self capable of deciding a plan of action, not many, no conflicting selves cut in to obstruct this process and the final plan will follow a normative standard that satisfies all selves involved.

2.4.2 The conflicting selves approach

The other approach models an agent who faces an internal tension between the deliberative System 2 and the impulsive System 1. In context K' (and not in K) the psychological 'cues' will activate a temptation to take y despite preferring x. Choice contexts in which a consumer faces a choice that involves the consumption

of addictive substances can be modelled as self-control problems. Lack of self-control will then naturally be seen as a failure to activate the deliberative System 2 and constrain the impulsive System 1. In such contexts it might be useful to think of each system as a "self" that has preferences and is motivated by her own interest, and describe the reasoning process as an internal conflict between a far-sighted, deliberative self and a short-sighted, impulsive self or selves.

Thaler and Sunstein (2008) model a decision maker who is assigned two selves, a 'planner' and a 'doer' that each has a preference relation. The 'planner' has the ability to construct rational and context-independent preferences but psychological cues which are irrelevant to the evaluation of options in question such as which option is designated as the default activate the 'doer'. Ideally, the 'planner' who is the more rational self exercises control over the the other self. Thaler and Sunstein use this hypothesis to motivate a behavioural analysis of welfare that promotes public policies such as *nudging* that are said to help individuals avoid reasoning errors (2008, pp. 40-1). Sunstein (2014, p. 150) writes referring to the practices of the nudger

> "In all these cases, the goal is not to encourage conscious deliberation or to activate System 2. It is to produce certain outcomes by influencing or appealing to System 1."

In these cases, the behavioural welfare economist, knowing what your System 2 preferences and your System 1 preferences are, can design a choice context that will activate the latter to achieve what is best for you that is, for System 2. According to the nudger, the architect of the choice context, these decisions are in people's own interest, and people have been informed about them before this policy takes place.

Bernheim and Rangel (2004) propose a model that describes the processes of drug addicts. Two selves describe the interactions between a "hot" mode of automatic responses to cues – the short-sighted self – and a "cold" mode of forward-looking reasoning – the far-sighted self. When the decision maker enters a "hot" mode she consumes addictive substances irrespective of the consequences of this choice. When the decision maker enters a "cold" mode she constructs preferences considering all possible implications of her choices, including the effects of cues on entering the "hot" mode in the future. Fudenberg and Levine (2006) present a game theoretic model where a sequence of short-sighted selves interact with a far-sighted self to construct their preferences. Fudenberg and Levine (2012); Thaler and Shefrin (1981); Brocas and Carrillo (2008) also model one self who is capable of making far-sighted economic decisions constraining the impulsive, short-sighted self. The setting these models propose is basically the same: there is an individual who makes a consumption-

Choices under Uncertainty			
Gamble 1	Gamble 2	Gamble 3	Gamble 4
'certain'	'likely'	'unlikely'	'very unlikely'
'small prize'	'medium prize'	'large prize'	'very large prize'

Time choices			
Reward 1	Reward 2	Reward 3	Reward 4
'today'	'tomorrow'	'in the future'	'in the distant future'
'small prize'	'medium prize'	'large prize'	'very large prize'

Table 2.1: Representation of different choice situations, the first from the perspective of how risky they feel and the second from the perspective of how distant they feel.

savings decision at different points in time. A decision is eventually made when one self overtakes the other in some specific sense defined in the model. These models differ essentially at the level of conflict between the two selves.

This tension between a short-run self and a long-run self has also been used to explain people's commonly observed attitudes towards risk. For example, Fudenberg et al. (2014) derive a simple version of their benchmark dual self model (2006) to interpret different attitudes towards certain and uncertain prizes as a tension between a short-run risk-averse self and a long-run risk-neutral self. The idea is that there is a relationship between the different attitudes that people have towards short-run and long-run rewards and the different attitudes that people have towards certain and uncertain outcomes. Table 1 shows one way in which decisions over time are related to decisions that involve gambles. Distance in time is treated as a source of uncertainty (i.e. the more distant the outcome is said to be, the less probable it is that you will get it, e.g. because of death).

Psychologically, attitudes to distance in time are similar to attitudes to probability: both impose 'temptation' between you and the outcome, and so reduce the force of positive and negative cues about the outcome. The short-run self who is 'tempted' to spend wins of the lottery immediately is very risk averse with small prizes relative to the long-run self who, in this model, thinks through the problem in a risk-neutral way. The authors identify the conditions which activate each self. Large stakes activate the long-run self who considers the additional amount saved from accepting the gamble, and small stakes activate the short run self. So this dual-self model explains different attitudes towards risk in the Allais's and the common ratio paradoxes as some sort of temptation to spend immediately the wins. Hammond and Zank (2014) offer a comprehensive review of the literature on this topic, which with the rise of dual self models has recently received more attention in economics.

All the dual-self models presented in this section have conflicting selves with conflicting preferences. That is, both selves are preference maximisers. When they agree, the common preference relation is maximised. But the one self can influence decisions by constraining, informing, or imposing costs on the first. If so, the deliberative self resolves the conflict by correcting the mistakes of the impulsive self. The interaction between the two systems is construed in such a way that System 2 is by design the outcome of a process according to standard economic theory.

To sum up, System 2 can be used in dual self models in many ways. I identified two of them: the two selves are conflicting or cooperative. What these types of models have in common is that all belong to a class of models in which an existing economic model can be retrieved. The advantage of the cooperative approach over the conflicting approach is that the optimal decision is calculated on System 2 alone. System 1 thinking is excluded, and with it the cost of constraining it. On the conflicting approach, the System 2 far-sighted self would have to constrain, often at some cost, the System 1 short-sighted self to reach a resolution. The underlying assumption is that the agent possesses some sort of 'meta-preferences' or 'meta-rationality' and has the ability to exercise self-control to decide which mode of thinking *should* prevail. A corollary of this is that every suboptimal solution is understood as a failure of self-control.

2.4.3 The complementary systems approach

A third modelling approach is to describe two processes, not their outcome while staying silent about what these processes are and do. One of the main insights of dividing psychology into two systems is that we can model the two systems together or in isolation, and better understand what the reasoning process of System 2 and the automatic process of System 1 can do. Cognitive sciences have, mostly informally, described models of this type. These models do not assume or exclude that existing economic models can be generated by some set of processes. Such a model will have the potential of telling us whether the brain cannot or simply does not produce certain reasoning patterns; if rational behaviour *can* be the output of reasoning or of automatic processes.

This approach, unlike the ones discussed previously, treats choice as the outcome of information processing carried out by different systems. Systems differ in their ability to process different types of information, and not necessarily in their rationality. This alternative, is closer to Kahneman's original idea of the two systems as

ways to describe the complementary functions of the brain processes with imperfect communication. Complementary functions might result in conflict or cooperation defined in the model in some specific sense. But the main point is that decision making takes into account the set of systems as a whole to make a choice.

Neurobiology inspired models have naturally adopted this approach. In Brocas and Carrillo (2012) there is a single decision maker, which they call 'Central Executive System'. Its role is to coordinate the systems that are involved in carrying out different tasks. Each system cares only about transmitting information to perform its own function which together describe the physiological constraints faced by the brain in the process of decision-making. Behaviour is the result of the interaction between systems with different objectives. The objective of the coordinator is to maximise the overall performance in the tasks. Her optimal decision depends on the physiological constraints of the systems that contribute towards the process carried out.

Yet another model by Brocas and Carrillo (2011) models the two systems as different ways in which our brain can retrieve information from memory to solve different kinds of (choice) problems. Different memory systems solve different kinds of problems. The authors distinguish systems of memory between the declarative and non-declarative or procedural. Declarative memory refers to recollection of historic events and facts while non-declarative memory refers to a simple way to retrieve information. Choice of one system over another is the result of an optimisation process between the effort and precision required of remembering an experience.

2.4.4 Taking stock

I identified three main approaches to answering my question. In the first two approaches, choice is the result of two types of selves. In the first approach, the two selves are allies that work together to solve a maximisation problem. In the second approach, choice is the result of two types of conflicting selves, the System 1 'irrational' self and the System 2 'rational' self that solve the maximisation problem against each other. This way of incorporated the System 1/ System 2 into behavioural economic models has been criticised by Kahneman:

> "The rational agent of economic theory would be described, in the language of the present treatment, as endowed with a single cognitive system that has the logical ability of a flawless System 2 and the low computing costs of System 1. Theories in behavioral economics have generally retained the basic architecture of the rational model, adding

> assumptions about cognitive limitations designed to account for specific anomalies." (Kahneman, 2003a, p. 1469)

Kahneman's own take on people's often non-economic reasoning is that this does not show that agents 'reason poorly but that they often act intuitively'. The implications derived in Section 3 make a case for an alternative way of modelling the agent's reasoning capabilities, according to which System 2 too has its own limitations or equivalently, that System 1 is not the only source of failures of rationality. At face value, this might seem wrong. The subconscious, implicit or non-verbal, associative reasoning processes are the source of some limitations of rationality. And the reflective and time-taking System 2 is activated to correct them. Although Kahneman does not say this explicitly, he seems to suggest it:

> "When we think of ourselves, we identify with System 2, the conscious, reasoning self that has beliefs, makes choices, and decides what to think about and what to do [...] I describe System 1 as effortlessly originating impressions and feelings that are the main sources of the explicit beliefs and deliberate choices of System 2 [...] System 1 has biases, however, systematic errors that it is prone to make in specific circumstances." (Kahneman, 2011, pp. 21 and 25)

But if conscious, explicit or verbal, rule-guided reasoning is part of what is involved in being rational, the limitations of our conscious, explicit or verbal, rule-guided reasoning should also be limitations of rationality.[6] If both systems have their own limitations, then both can be the source of systematic errors. Put it in another way, if a system can be the source of systematic errors, these errors must be errors of something. They must be errors of failing to reason in some *correct* way; where this correct way is the result of an interplay between the two systems. Sugden (2018, p. 68) puts this nicely,

> "One is not entitled simply to assume that the mental processes of System 2 can generate preferences and modes of strategic reasoning that are consistent with conventional decision and game theory. Indeed, that assumption does not fit easily with the logic of dual-process theory. One of the fundamental insights of that theory is that the automatic processing mechanisms of System 1 are evolutionarily older than the conscious mechanisms of System 2. Thus, except in so far as its original features have atrophied, we should expect System 1 to be capable of

[6] This is something that Kahneman (2011, p. 21) has acknowledged *"You will be invited to think of the two systems as agents with their individual abilities, limitations, and functions."*

> generating reasonably coherent and successful actions without assistance from other processes. But if System 2 processes are later add-ons, there is no obvious reason to expect them to be able to work independently of the processes to which they have been added."

So if one has not investigated the inner workings of our System 2, one is not entitled to assume that System 2 is by design capable of creating those mental states such as preferences, beliefs, and intentions that a far-sighted self is assumed to have. For example, Chen (2013) studies future-oriented behaviours such as the stylised examples of dynamic consistency studied above. This study is an empirical investigation of the time-honoured hypothesis that language affects a decision maker's view of the world. Chen shows how speaking and thinking in a different language affects people's future-oriented behaviours such as saving, exercising, abstaining from smoking and long-run health. In some languages the future is not separated from the present. The future appears closer to speakers of these languages, and more distant to speakers of languages where there is a sharp distinction between present and future. The speakers of languages where future appears closer tend to save and adopt a healthy lifestyle more than the speakers of language where future appears more distant. This, it seems to me, is the outcome of some conscious, explicit, and verbal process in which we think and make plans. But, as I have discussed economists have traditionally attributed the behaviour of the short-sighted self as the outcome of the less rational and impatient System 1, and the behaviour of the far-sighted self as the outcome of the infinitely patient System 2.

Although Chen's findings are impressive, his analysis has attracted the criticism of both economists and linguists. Roberts et al. (2015) have shown that the correlation between languages that grammatically mark future events and their speakers' propensity to save in Chen's paper is weaker when controlling for links between other cultural traits. This is despite that in the original paper a set of controls is designed to address many of these concerns. Another kind of criticism comes from the linguist Dahl. His criticism is about the way in which Chen classifies languages according to how strongly they grammatically separate the future and the present (Dahl, 2013). More favourable is the experimental study conducted by Sutter et al. (2015) in which time preferences of either German-speaking or Italian-speaking primary school children are examined. Their results provide evidence that is, according to the authors, 'markedly consistent with the linguistic-savings hypothesis proposed by Chen (2013)'. Becker et al. (2018); Galor et al. (2017); Tabellini (2008) empirically investigated the effect of language on economic decisions suggesting further evidence in support of this effect.

The moral from this section is that on a correct understanding of the dual system hypothesis, there is no reason to qualify one self as rational and the other as irrational. Without an explicit model of the reasoning process by which people are alleged to construct their rational preferences, we cannot really be sure whether human behaviour is the output of our reasoned or automatic processes. And, we cannot really know whether they are mistakes in reasoning or unrealistic assumptions about reasoning.

2.5 Conclusion

Rational choice often assumes that System 2 is capable of constructing rational and context-independent preferences according to economic theory. Failure to achieve them is often attributed to the fast System 1 which often cuts in the decision maker's System 2 reasoning.

Philosophy and cognitive sciences have highlighted the main features of a cognitive system with the logical ability of System 2. It is conscious, explicit, and rule-guided. This chapter presented a novel model of the conscious, explicit, and rule-guided System 2. I have created the formal language that allows us to address the limitations of System 2. To my knowledge, this is the first attempt to formalise these processes and with it its limitations. Some of these limitations may result from the fact that System 2 reasoning is conducted in a language. In reasoning explicitly you cannot start from the *absence* of an attitude, a thesis that Kolodny agrees with (2005, pp. 527-8), or conclude in the *absence* of an attitude, a thesis that Broome agrees with (2013, p. 278), or conclude in *either* the presence of one attitude *or* of another one.

Theorem 2.2 shows that reasoning that concludes in the absence of an attitude is impossible, and Theorem 2.3 shows that reasoning that starts from the empty set although possible is prone to inconsistencies. I also show in theorems 2.5 and 2.4 that reasoning that concludes in an *either* ... *or* ... form is also prone to inconsistencies.

Broome's own interpretation is that when reasoning fails 'automatic processes will normally prevent you from having contradictory beliefs' to achieve consistency (2013, pp. 279-280). I discussed certain ways in which System 1 can overcome the limitations of System 2; Manzini and Mariotti's "categorise then choose" model (2012) and Simon's model of "satisficing" (1955) among others.

2.6 Appendix

2.6.1 Definitions of closure

I give two equivalent definitions of $C|\mathcal{S}$: the top-down and the bottom-up closure. Def. 2.7 and 2.6 give both formal definitions of $C|\mathcal{S}$. Prop. 2.1 shows their equivalence.

Definition 2.6. For any constitution C and any set $\mathcal{S}$ of reasoning rules:

1. Define a constitution $C \mid_1 \mathcal{S} = C \cup \{m : (M, m) \in \mathcal{S} \text{ and } M \subseteq C\}$ that is produced by maximally applying rules from $\mathcal{S}$ to C itself, i.e. 'one-step' applications of rules.

2. Define a constitution $C \mid_2 \mathcal{S} = (C \mid_1 \mathcal{S}) \mid_1 \mathcal{S}$, i.e. by maximally applying rules from $\mathcal{S}$ to C two-times.

3. Define a constitution $C \mid_n \mathcal{S} = ((C \mid_1 \mathcal{S}) ...) \mid_1 \mathcal{S}$, i.e. by maximally applying rules from $\mathcal{S}$ to C n-times.

4. Let n be the smallest number at which $C \mid_{n+1} \mathcal{S} = C \mid_n \mathcal{S}$.

   Then define the bottom-up closure of C under $\mathcal{S}$ by $C \mid_n \mathcal{S}$.

Definition 2.7. For any constitution C and any set $\mathcal{S}$ of reasoning rules the revision $C|\mathcal{S}$ is:

(a) The smallest expansion of C that is closed under S, and

(b) The intersection of all expansions of C that are closed under S.

Then define $C|\mathcal{S}$ as the top-down closure of C under $\mathcal{S}$.

I now show their equivalence.

Proposition 2.1. *For any constitution C and any set $\mathcal{S}$ of reasoning rules, the two definitions of bottom-up and top-down closure are logically equivalent.*

Proof of Prop. 2.1. Let C be a constitution, $\mathcal{S}$ a system 2, and n the smallest number at which C is closed under $\mathcal{S}$. Consider that $C|\mathcal{S}$ is defined in the bottom-up way. I need to prove the following two: $C|\mathcal{S}$ is (a) the smallest expansion of C that is closed under $\mathcal{S}$, and (b) the intersection of all expansions of C that are closed under $\mathcal{S}$.

Proof of (a): (1) By construction, the following are true of $C|\mathcal{S}$: (1a) $C|\mathcal{S}$ is an expansion of C that is closed under $\mathcal{S}$ because $(C|\mathcal{S})\,|_1\mathcal{S} = C|\mathcal{S}$. (1b) $C|\mathcal{S}$ is an expansion of C that contains every attitude that can be derived from C by successive application of rules from $\mathcal{S}$.

(2) From (1b), every expansion of C that is closed under S is a weak superset of $C|\mathcal{S}$.

Proof: Let C' be any weak superset of C which is not a weak superset of $C|\mathcal{S}$. Then there is some attitude m such that $m \in C|\mathcal{S}$, $m \notin C'$. By (1b) m can be derived from C, and hence also from C', by successive application of rules from $\mathcal{S}$. So C' is not closed under $\mathcal{S}$.

(3) From (1a) and (2): $C|\mathcal{S}$ is the smallest expansion of C that is closed under $\mathcal{S}$.

Proof of (b): From (2): $C|\mathcal{S}$ is a weak subset of the intersection of all expansions of C that are closed under $\mathcal{S}$. But from (1a), $C|\mathcal{S}$ is itself an expansion of C that is closed under $\mathcal{S}$. So the intersection of all expansions of C that are closed under $\mathcal{S}$ must be $C|\mathcal{S}$. □

2.6.2 Proof of the characterisation results

Proof of Theorem 2.1. Let $\mathcal{T}$ be a theory. Define the system 2 $\mathcal{S}$ as to contain all reasoning rules corresponding to closedness requirements of $\mathcal{T}$. So $\mathcal{S} = \{(M, m)$: the closedeness requirement given by (M, m) is a requirement of $\mathcal{T}\}$. Now consider any initial constitution C_0 and any closedness requirement R of $\mathcal{T}$, given by a pair (M, m).

Claim 1. $C_0 \mid \mathcal{S}$ satisfies the requirement R.

Proof: This is true because if $M \subseteq C_0 \mid \mathcal{S}$, then $m \in C_0 \mid \mathcal{S}$ because $C_0 \mid \mathcal{S}$ is closed under $\mathcal{S}$ which contains (M, m).

Claim 2. $C_0 \mid \mathcal{S}$ preserves consistency.

Proof: Assume C_0 is consistent, hence a subset of a rational constitution in C^* in $\mathcal{T}$. I show that $C_0 \mid \mathcal{S} \subseteq C^*$. This follows from two facts. The first is that C^* is closed under $\mathcal{S}$, because it is rational and hence, in particular satisfies all closedness requirements of $\mathcal{T}$. The second is that $C_0 \mid \mathcal{S}$ is by definition the smallest expansion of C_0 under $\mathcal{S}$. □

Proof of Theorem 2.2. Consider a system 2 $\mathcal{S}$, and a consistency requirement R given by a set of attitudes M. So $R = \{C : M \not\subseteq C\}$. It suffices to specify a

constitution C_0 such that $C_0 \mid \mathcal{S}$ violates R. Simply let C_0 be any constitution that includes M. Since $C_0 \mid \mathcal{S}$ includes C_0, it also includes M, hence violates the requirement R. □

Proof of Theorem 2.3. Consider a theory $\mathcal{T}$.

1. For each completeness requirement R of $\mathcal{T}$, fix an arbitraray member m_R of the set M defining R. This is possible because the set M of any completeness requirement is by definition non-empty, or so I need to assume. Define the system 2 as $\mathcal{S} = \{(\emptyset, m_R)$: R is a completeness requirement of $\mathcal{T}\}$. Now consider any constitution C_0. I must show that $C_0 \mid \mathcal{S}$ satisfies all completeness requirements R of $\mathcal{T}$. As one easily checks, $C_0 \mid \mathcal{S}$=$C_0 \cup \{m_R$: R is a completeness requirement of $\mathcal{T}\}$. Clearly, this constitution $C_0 \mid \mathcal{S}$ satisfies all completeness requirements of $\mathcal{T}$.

2. Let $\mathcal{S}$ be a system 2 which achieves a completeness requirement R of $\mathcal{T}$ given by a set M consisting of attitudes whose members are falsifiable. I must find a consistent constitution whose revision is inconsistent. I first show the following:

Claim. $\mathcal{S}$ contains a rule $s = \{(M', m) : m \in M\}$.

Proof: Consider any constitution C_0 disjoint from M (e.g. $C_0 = \emptyset$). By the definition of "achieving requirement", $\mathcal{S}$ achieves R if $\emptyset \mid \mathcal{S}$ satisfies R. So there is a $m \in M$ such that $m \in \emptyset \mid \mathcal{S}$. This implies the claim, by definition of revision through $\mathcal{S}$.

Now let m be as in the above claim. As m is falsifiable, we may pick a consistent constitution C_0 such that $C_0 \cup \{m\}$ is inconsistent. By the definition of "preserving consistency", every superset of $C_0 \cup \{m\}$ is inconsistent. I need to show that there is some consistent constitution C' such that $C'|\mathcal{S}$ is inconsistent. The way to do this is to set $C' = C_m$ and to show that $C_m|\mathcal{S}$ is inconsistent. So I need to show that $m \in C_m|\mathcal{S}$ and hence that, $C_m|\mathcal{S}$ is a superset of $C_m \cup \{m\}$. It suffices to note that $\emptyset|\mathcal{S} \subseteq C_m|\mathcal{S}$ because anything that can be derived from $\emptyset$ can be derived from C_m. Since $m \in \emptyset|\mathcal{S}$, we must have $m \in C_m|\mathcal{S}$. So $\mathcal{S}$ does not preserve consistency. □

Proof of Theorem 2.4 and 2.5. Since Theorem 2.4 is the special case of Theorem 2.5 in which N is a singleton, it suffices to prove Theorem 2.5. Consider a theory $\mathcal{T}$.

1. By definition, each conditional completeness* requirement of $\mathcal{T}$ is conditional on some set M of attitudes, and has at least one attitude from N. For each such requirement, fix an arbitrary member m_N of N, and define the rule $s = (M, m_N)$. Let $\mathcal{S}$ be any System 2 containing one such rule for each conditional completeness* requirement. Clearly, $\mathcal{S}$ achieves all conditional completeness* requirements of $\mathcal{T}$.

2. Consider any conditional completeness* requirement of $\mathcal{T}$, defined by some M and N where $N \neq \emptyset$. Suppose some consistency-preserving System 2 $\mathcal{S}$ achieves this requirement. Since the requirement is achieved, there must be some $m' \in N$ such that $m' \in M|\mathcal{S}$. To complete the proof, it suffices to show that m' is not falsifiable given M. To that end, I consider a consistent $C_0 \supseteq M$, and must show consistency of $C_0 \cup \{m'\}$. Because $\mathcal{S}$ is consistency-preserving, $C_0|\mathcal{S}$ is consistent. But $m' \in M|\mathcal{S}$ and $C_0 \supseteq M$ imply $m' \in C_0|\mathcal{S}$. As $C_0|\mathcal{S}$ $(= C_0|\mathcal{S} \cup \{m'\})$ is consistent, so is its subset $C_0 \cup \{m'\}$. □

Chapter 3

"Preaching" rationality

3.1 Introduction

Chapter 2 presented a simple language-based and rule-following model of reasoning and investigated the extent to which rational choice can be reached by it. I argued that this account of reasoning can be understood as an explicit description of System 2 reasoning, and showed that if reasoning is conducted in language, it cannot achieve consistency requirements. In plain English, this result says that, in reasoning explicitly you cannot conclude in the *absence* of an attitude, a thesis that Broome agrees with (2013, p. 278). A legitimate critique against this account of reasoning is that reasoning that does not allow you to remove an attitude of yours, even if doing so would prevent you from being inconsistent, is incomplete reasoning. [1] Broome's own interpretation is that when explicit reasoning fails 'automatic processes will normally prevent you from having contradictory beliefs' to achieve consistency (2013, p. 278). But there is not much discussion of this. Chapter 3 gives a possible interpretation.

I re-examine the famous case of Savage's response to the Allais Paradox (1954) in which Savage explains how he resolved his personal problem of discovering that his preferences over Allais's gambles were inconsistent with the sure-thing principle; a basic requirement of rational choice under uncertainty. The interest of reconstructing Savage's response is that on my Broome-inspired account of reasoning as a rule-following mental process by which you form new attitudes based on existing ones, a rule applies only to add a new attitude. For instance, no rule removes the belief in a proposition p based on the premise belief in not p. So it is unclear how Savage can reason without the use of such rules that allow the removal of inconsistent

[1] I would like to thank many people for pointing out to me that this aspect of Broome's account of reasoning should not remain unaddressed and in particular Ben McQuillin.

preferences. A possible solution is to start from an ought-belief that you ought to satisfy the axioms of expected utility to derive the intention to satisfy them. But on this account of reasoning, this solution is not possible: although you can conclude through active reasoning that you ought to give up a preference, this adds an ought-belief rather than removing the preference in question. This ought-belief may thereafter cause disappearance of this preference, but no longer through explicit reasoning.

Section 2 gives the background of the debate between Allais and Savage and Savage's response in his book (1954, pp. 101-103). I conclude that his response draws on a process of reasoning that goes beyond what I consider as formally 'correct' reasoning and discuss what are the other psychological causal processes that help Savage to become consistent. In section 3 I use the formal analysis in Chapter 2 to reconstruct Savage's response. Section 4 discusses an alternative response that starts from the normative belief that a person ought to satisfy the axioms of expected utility. I argue that this type of reasoning is ineffective. Section 5 offers my concluding remarks of the chapter and Section 6 my concluding remarks of my thesis.

3.2 Savage's discussion of the Allais paradox

The 1952 Paris symposium on the 'Foundations and applications of the theory of risk-bearing' was the scene of an important debate in the history of behavioural economics. Savage presented his axiomatization of subjective expected utility that would later become the core of his book *The Foundations of Statistics*. Maurice Allais was among the main objectors to the use of the expected utility axioms as requirements of rational choice. In an encounter on the fringes of the colloquium Allais presented Savage with a choice problem that trapped him into violating his expected utility axioms – what has now become known as the Allais paradox.[2]

In (1954, pp. 101-103), Savage discusses the paradox and explains how he reversed his preferences and become consistent. The problem consists of two decision situations. Each situation asks for a choice between two gambles. In Situation 1, the choice is between Gamble 1, which gives \$500,000 with probability 1, and Gamble 2, which gives \$2,500,000 with probability 10/100, \$500,000 with probability 89/100, and nothing with probability 1/100. In Situation 2, the choice is between Gamble 3, which gives \$500,000 with probability 11/100 and nothing with probability 89/100, and Gamble 4, which gives \$2,500,000 with probability 10/100 and nothing with

[2]See in particular, (Mongin, 2018).

probability 90/100. Savage has initially expressed a preference for Gamble 1 to Gamble 2 and a preference for Gamble 4 to Gamble 3. This pair of preferences violated the sure-thing principle which is an implication of his own axioms.[3]

Savage states, rather informally, the sure-thing principle in (1954, pp. 21-22):

> "Let me give a relatively formal statement thus: If the person [...] would definitely prefer g to f, knowing that E obtained, and, if he would not prefer f to g, knowing that E did not obtain, then he definitely prefers g to f."

In Savage's framework, a preference can be understood as some kind of conditional intention to choose. Of two gambles, x and y, the preference for x to y is the attitude that would typically cause x to be chosen from the two gambles given no other gamble was available. But as he acknowledges (1954, p. 22), the notion of 'f preferred to g, *knowing* the event E obtains' cannot be expressed in terms of his primitives. So Savage uses the following example to interpret his principle:

> "A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he knew that the Republican candidate were going to win, and again finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say." (Savage, 1954, p. 21)

For Savage, the sure-thing principle is, along with P1 (completeness and transitivity), one of the two 'extralogical principle[s] governing decisions'. So Savage develops 'a sense of discomfort' when he finds out that his initial preferences violate it. In Savage's own words:

> "When the two situations were first presented, I immediately expressed [the preferences Allais predicted], and I still feel an intuitive attraction to those preferences. But I have since accepted the following way of looking at the two situations which amounts to repeated use of the sure-thing

[3]There is a large literature in behavioural economics and psychology on choice experiments showing that people have difficulty with hypothetical thinking, and particularly the type of hypothetical reasoning that is related to the sure-thing principle (see in particular, (Slovic and Tversky, 1974; Esponda and Vespa, 2014, 2017; Sugden, 1991; Berg and Gigerenzer, 2010; Allais, 1953; Shafir and Tversky, 1992; Cubitt and Sugden, 2003, 2014; Cubitt et al., 1998)).

Situation 1			
Gamble 1	'certain'		
	'large prize'		
Gamble 2	'unlikely'	'likely'	'very unlikely'
	'very large prize'	'large prize'	'no prize'

Situation 2		
Gamble 3	'unlikely'	'likely'
	'large prize'	'no prize'
Gamble 4	'unlikely'	'likely'
	'very large prize'	'no prize'

Table 3.1: Savage's initial mental representation of the two decision situations.

> principle [...] It seems to me that in reversing my preference between Gambles 3 and 4 I have corrected an error" (Savage, 1954, p. 103)

The passage in which Savage spells out his own way of looking at the two situations which amounts to repeated use of the sure-thing principle is two pages long (1954, pp. 101-103), which is relatively short given that Savage's response to Allais depends on it. It will be useful to divide his response into smaller parts or steps and when possible, represent his 'reasoning' in them.

In **Step 1** Savage expresses his initial preferences for Gamble 1 to Gamble 2 in Situation 1 and for Gamble 4 to Gamble 3 in Situation 2. He says that these preferences report his 'initial impression of the situation' as one between the gift of a large prize and the chance of winning a very large prize in Situation 1, and one between a large prize and a very large prize at nearly the same chance in Situation 2 (1954, p. 102). In this spirit, Rubinstein (1988) has proposed a model in which people check similarities between outcomes and between probabilities in gambles. If only one of the two dimensions, probabilities or outcomes, are similar, then the other dimension becomes the decisive factor. Table 1 is one possible way to present Savage's initial *mental representation* of the two situations. From this table it can be seen that, you compare the certainty of a large prize in Gamble 1 with the chance to win a very large prize in Gamble 2, and the chance to win a very large prize in Gamble 4 with the chance to win a large prize in Gamble 3; which justifies why you may end up expressing 'a strong intuitive appeal' for Gamble 1 to Gamble 2 and for Gamble 4 to Gamble 3.

The language in which you compare the gambles (in terms of their sizes and the chances of winning them) is vague and coarse but intuitive, allowing you to immediately express your inclinations. But in the explicit language of expected utilities in which you multiply exact probabilities and utilities, the inequalities U(\$500,000) >

	100 Tickets	1	2 to 11	12 to 100
Situation 1	Gamble 1	0.5	0.5	0.5
	Gamble 2	0	2.5	0.5
Situation 2	Gamble 3	0.5	0.5	0
	Gamble 4	0	2.5	0

Table 3.2: Savage's mental representation of the two decision situations in Step 2 (prizes are in units of $1,000,000).

10/100 U($2,500,000) + 89/100 U($500,000) + 1/100 U(0), and 11/100 U($500,000) + 89/100 U(0) < 10/100 U($2,500,000) + 90/100 U(0) are inconsistent as no function U satisfies both inequalities. Savage realises that his preferences are inconsistent with expected utility, and indeed with the sure-thing principle, so, in **Step 2** he presents his reader with a table in which the two situations can be reconstructed with the sure-thing principle. In it, uncertainty is not described by probabilities but by a set of states of the world which in this case represent a lottery with a hundred tickets numbered 1–100. The set of all states of the world are all possible tickets in the lottery. Table 2 adopts Savage's representation of the two situations.

Notice that, Savage's reconstruction of the two situations does not make any direct reference to his own theoretical framework. A person that does not know expected utility theory might be able to think of the four gambles in the two situations in an explicit way in terms of a lottery or another notion that represents probabilities. One possibility is to sketch a table in which there are four rows corresponding to the Gambles 1, 2, 3, 4 and a hundred columns each corresponding to the probability 1/100; and then assign the outcome of a given gamble in the corresponding row and column(s). One way of doing this, say for Gamble 2, is to assign 0 in column one, $2,500,000 in columns 2 to 11, and $500,000 in all other columns. If we first group the columns with the same outcome for any given gamble and then align the outcomes in each column, we will obtain Table 3 which is the same as Table 2 with only exception that the 100 tickets of Savage's lottery have been replaced by 100 columns.

From the previous step, Savage infers in **Step 3** that the two choices between the 'original' pairs of gambles in the situations depends on the choice between a 'new' pair of gambles given that one of the tickets numbered from 1 to 11 is drawn. In Savage's words (1954, p. 103):

> "if one of the tickets numbered from 12 through 100 is drawn, it will not matter, in either situation, which gamble I choose. I therefore focus on the possibility that one of the tickets numbered from 1 through 11 will be drawn, in which case Situations 1 and 2 are exactly parallel"

	Columns	1	2 ... 11	12 ... 100
Situation 1	Gamble 1	0.5	0.5	0.5
	Gamble 2	0	2.5	0.5
Situation 2	Gamble 3	0.5	0.5	0
	Gamble 4	0	2.5	0

Table 3.3: Explicit representation of the decision situations (prizes are in units of \$1,000,000).

	100 Tickets	1	2 to 11	2 to 11
	Events	E		$\bar{E}$
Situation 1	Gamble 1	0.5	0.5	*
	Gamble 2	0	2.5	*
Situation 2	Gamble 3	0.5	0.5	#
	Gamble 4	0	2.5	#

Table 3.4: Savage's mental representation of the two decision situations in Step 3 (prizes are in units of \$1,000,000).

In Table 4, I present a possible way of representing Savage's reasoning in this step in which he *focuses* on the possibility that one of the tickets numbered from 1 through 11 will be drawn. For representational convenience, I write E and $\bar{E}$ (the complement of E) to demarcate the possibility that a ticket numbered from 1 to 11 is drawn from the lottery, from the possibility that a ticket numbered from 12 to 100 is drawn from the lottery. In each situation, the outcomes of the two gambles are the same in $\bar{E}$ but differ in E. So Savage focuses on event E and ignores the event $\bar{E}$ where in each situation outcomes are the same. I label with '*' the same outcomes in $\bar{E}$ in Situation 1 and with '#' the same outcomes in $\bar{E}$ in Situation 2. Note that in Savage's framework, the notion of an *event* is defined as a collection of states of the world, but his analysis does not make any explicit use of this notion.

The upshot is that Savage can now alter his initial preferences by forming a preference between the gambles conditional on a ticket numbered from 1 to 11 being drawn from the lottery rather than between the acts themselves. So the next step, **Step 4**, is to *reconsider* his preferences focusing on event E:

> "The subsidiary decision depends in both situations on whether I would sell an outright gift of \$500,000 for a 10-to-1 chance to win \$2,500,000 – a conclusion that I think has a claim to universality, or objectivity." (Savage, 1954, p. 103)

	Events	E		$\bar{E}$
Situation 3	Gamble 1 or 3	'certain'		*
		'large prize'		
	Gamble 2 or 4	'unlikely'	'very likely'	*
		'no prize'	'very large prize'	

Table 3.5: Savage's mental representation of the decision situations after repeated use of the sure-thing principle.

So in Step 4 Savage calls back his intuitive and non-explicit language he used in Step 1 to justify his preference for Gamble 1 to Gamble 2 in E (equivalently for Gamble 3 to Gamble 4 in E), thus expressing a preference for a large and certain prize to a very large but uncertain prize. Table 5 illustrates the natural way in which Savage justifies his preferences between the new pair of gambles in E, call this a new situation 'Situation 3'. On this 'way of looking at' the decision problem, the decision in Situation 1 between the certainty of a large prize (Gamble 1) and a very large but uncertain prize (Gamble 2) and the decision in Situation 2 between an uncertain but very large prize (Gamble 4) and an uncertain but large prize (Gamble 3), have been reconstructed into a decision between a new pair of gambles: the certainty of a large prize (Gamble 1 or 3 in E) and a very large but uncertain prize (Gamble 2 or 4 in E). What is crucial, the uncertain Gamble 2 has been transformed into a less uncertain Gamble 2 in E and the uncertain Gamble 3 has transformed into a certain Gamble 3 in E. In Savage's words, "... consulting my purely personal taste, I find that I would prefer the gift of \$500,000 ... " (Savage, 1954, p. 103).

In **Step 5** Savage moves back from Situation 3 to situations 1 and 2. He uses the sure-thing principle in two instances: once to express a preference for Gamble 1 to Gamble 2 in Situation 1 given his preference for Gamble 1 or 3 in E to Gamble 2 or 4 in E in Situation 3, and a second time to express a preference for Gamble 3 to Gamble 4 in Situation 2 given his preference for Gamble 1 or 3 in E to Gamble 2 or 4 in E in Situation 3. Now Savage, has two conflicting preferences: a preference for Gamble 3 to Gamble 4 and a preference for Gamble 4 to Gamble 3.

Savage says that, "... accordingly, [I find] that I prefer Gamble 1 to Gamble 2 and (contrary to my initial reaction) Gamble 3 to Gamble 4." (Savage, 1954, p. 103). The final step in Svage's response to the problem, **Step 6,** is to find that he no longer has a preference for Gamble 4 to Gamble 3.

Discussing an analogous case, Savage says:

> "When it is explicitly brought to my attention [that my preferences are non-transitive] I feel uncomfortable in much the same way as when it is brought to my attention that some of my beliefs are logically contradictory. Whenever I examine such a triple of preferences on my own part, I find that it is not at all difficult to reverse one of them. In fact, I find on contemplating the three alleged preferences side by side that at least one of them is not a preference at all, at any rate not any more." (Savage, 1954, p. 21)

It seems that Savage's reasoning does not involve an intention to remove the preference for Gamble 4 to Gamble 3 to become consistent. But Savage's reasoning to become consistent draws on some sort of 'implicit' or 'automatic' process that replaces his old preference with his new preference for Gamble 3 to Gamble 4 that is derived by reasoning from premises he feels confident about.

The moral from this section is twofold: First, Savage develops 'a sense of discomfort' when he finds out that his initial preferences violate the sure-thing principle, and subsequently, adopts a 'way of looking at' the decision problem [represented in Tables 2-5] that helps him reach a new set of preferences, about which he feels more confident. Second, Savage's newly-reached preferences satisfy the sure-thing principle, and Savage reaches these preferences drawing on a process that combines System 1 with System 2. In particular, Savage draws on System 1 not only to express his initial preferences in Step 1 [represented in Table 1] but also to express a preference between the new pair of gambles in E in Step 4 [represented in Table 5]. Notice that Table 5 is a natural representation of the decision in Situation 3 which justifies why you may end up producing Savage's revised preferences (Gamble 3 preferred to Gamble 4) while Table 1 is a natural representation of the decisions in situations 1 and 2 which justifies why you may end up producing Allais's preferences (Gamble 4 preferred to Gamble 3).

The section above re-examined a particular case: Savage's own account of how he 'corrected' his 'error' of having preferences that violate the axioms of expected utility theory. Many decision theorists and behavioural economists also characterise behaviour that violates these axioms as 'error' and assume that individuals will somehow be able to correct this error. One particular way is to think that "preaching" the axioms of expected utility theory as normative will (eventually) help people make more rational choices and less biased judgments. I call this, the "preaching" approach. The alternative is to think that people can construct rational preferences without knowing, for example without being preached, what the axioms of expected

utility are. If you have a 'legitimate' way of discovering that your preferences are in some sense in 'error' with you, because if you apply a legitimate rules of reasoning on these preferences you end up in a contradiction, then you can revise these preferences. I call this the rule-following approach. The next two sub-sections are about these two approaches and how Savage's response is related to them.

3.3 The "preaching" approach

Advocates of the preaching approach include Gilboa (2010, p. 4), Gilboa (2009, pp. 200-201), and Bleichrodt et al. (2001), among others. Gilboa discusses two approaches in face of experimental evidence showing that in practice people often violate the axioms of rational choice:[4]

> "One approach is to incorporate them [the violations] into our descriptive theories, to make the latter more accurate. This is, to a large extent, the road taken by behavioral economics. Another approach is to go out and preach our classical theories, that is, to use them as normative ones. For example, if we teach more probability calculus in highschool, future generations might make less mistakes in probability judgments. [...] [I]f we find that, when we explain the theory to decision makers, they are convinced and wish to change their choices, (that is, if their choices were irrational to them), we may declare a success of the classical theory as a normative one. It would indeed be reasonable to preach the classical theory and help decision makers make better decisions (as judged by themselves)" (Gilboa, 2010, p. 4)

Against the preaching approach, there are two main lines of criticism. Those, Broome (2013), Infante et al. (2016), and Sugden (2018) among them, who think that preaching is not effective and do not assume that failing to respond to preaching is a reasoning error; but argue that we need to understand the reasoning process by which people are supposed to satisfy the rationality requirements. And those, Thaler and Sunstein (2008) among them as proponents of the nudging approach, who think that preaching is not effective but assume that failing to respond to preaching is a reasoning error. This chapter is about the former. Chapter 2 has commented on the latter. Broome criticises theorists that seek to describe theories of rationality such as rational choice theory in terms of requirements of rationality (e.g. the sure-thing

[4]Gilboa and Schmeidler (1995); Gilboa et al. (2009, 2012, p. 630) acknowledge that forming preferences according to expected utility is not always "rational" or even possible. They do not necessarily accept expected utility as normative criterion.

principle), and often neglect to explain the 'reasoning' by which one comes to satisfy these requirements. In the words of Broome:

> "[they] seem to think they have finished their job when they have described the requirements of rationality. [...] I think these authors must believe that, once you know what requirements there are, that knowledge directly supplies you with premises you can use in active reasoning. They must believe that, starting from knowledge of a particular requirement, you can reason your way actively to satisfying that requirement" (2013, pp. 208-9)

According to Broome, second-order reasoning – reasoning about propositions about your own attitudes – cannot actively bring you to satisfy a particular requirement. For instance, you can conclude through active reasoning that you ought to give up your belief in p, but this adds an ought-belief rather than removing the belief in p. This ought-belief may thereafter cause disappearance of the belief in p, but no longer through explicit reasoning. Broome thus concludes that having second-order attitudes is not necessary for reasoning (2013, p. 236). Rather, for him, reasoning is mainly *with* and not *about* your attitudes because this is the fundamental kind of reasoning that is done using language (2013, pp. 268-86).

Infante et al. (2016) and Sugden (2018) have criticised the preaching approach as psychologically problematic. According to Sugden, behavioural welfare economics explains the inability to construct rational (complete and transitive) and context-independent preferences as *reasoning imperfections* but,

> "does not try to explain the reasoning by which individuals construct their preferences. Implicitly, rational-choice theory assumes the existence of a mode of reasoning that generates preferences that satisfy the consistency axioms, but it treats that reasoning as a black box" (2018, p. 63).

For Infante et al. and Sugden, these economists assume the existence of an 'inner rational agent' who is isolated from the world by 'a psychological shell', and who can construct rational (complete and transitive) and context-independent preferences. Their reply is that the 'preachers' ought to tell us what they think this process that leads to rational and context-independent preferences is.

Slovic and Tversky (1974) point out another problem of thinking that preaching the theory's axioms as normative (even if they are rationality axioms) will make people more rational. Their study looks experimentally at whether preaching the sure-thing principle is enough to convince people to accept it. This study can be

classified as an empirical test for the "preaching" approach. In a first experiment, subjects were asked to report their preferences between two pairs of gambles in two decision problems; the Allais and the Ellsberg decision problems which are two classical examples that show that people's choices violate the sure-thing principle. Having reported their preferences, subjects were asked to read the competing arguments of experts Dr. S, advocating Savage's sure-thing principle, and Dr. A., advocating Allais' position. Those who made a choice in accordance with the principle were asked to read Dr. A's position and those who made a choice violating the axiom were asked to read Dr. S's position. Exposed to arguments that countered their choices, subjects were asked whether they would switch them. The authors found that subjects' choices survived the counterarguments with that of Dr. S being even less effective in influencing the subjects' choices. In a second experiment, subjects were given both arguments for and against the sure-thing principle and were asked to rate how persuasive these arguments were. The authors found that subjects rated Dr A's arguments as more persuasive. Slovic and Tversky conclude their paper with a hypothetical dialogue between Dr. S and Dr. A. The authors attribute preaching to Dr. S and the psychological mechanism that experimental economists have used to explain observed violations to Dr. A. Dr. S, who tries to defend his axiom against Dr. A's critique, says:

> "In my experience, it often takes a long time for people to appreciate the normative impact of axioms. They have to be educated before they are willing to live by the axioms of rational choice".

And Dr. A's hypothetical answer is,

> "You seem to be saying that [the sure-thing principle] enjoys normative status because [...] some people could convince other people that they should accept it. Even if I could accept the axiom, I certainly could not accept this criterion. Your ability to convince people to accept an axiom is not a sufficient basis for establishing its normative appeal. What you call education, others may call brainwashing. Why do you not simply accept the fact that, unlike transitivity, [the sure-thing principle] lacks general appeal as a normative principle of choice?" (Slovic and Tversky, 1974, p. 372)

The authors' conclusion section aims to stress that success in preaching an axiom is not evidence of the rationality of the axiom. Since Dr. S's and Dr. A's arguments are contradictory but both had persuasive power as a matter of empirical fact, it would be a mistake to assume that preaching rationality is effective.

So Slovic and Tversky attribute the preaching approach to Savage who believed his axioms to be normative and famously said: "the main use I would make of [completeness and transitivity] and its successors is normative, to police my own decisions for consistency and, where possible, to make complicated decisions depend on simpler ones" (1954, p. 20). Unfortunately, Savage never explains in his response to Allais in (1954, pp. 101-103) how he became aware of his violation. Savage might have learned, by hunch, private information or reasoning, that the sure-thing principle is violated. Most plausibly, Savage has been told that his initial preferences violate the sure-thing principle which explains his initial 'sense of discomfort'. One way to interpret 'to police [his] own decisions for consistency' is that he relies on some sort of second-order reasoning as he knows that his preferences violate the sure-thing principle. Then Savage plausibly forms the belief that 'I ought to satisfy the the sure-thing principle' and then creates an intention to satisfy it.

Let us investigate this approach considering Savage's response from the previous section. I illustrated that Savage can police his decisions without using any type of second-order normative beliefs. Having second-order beliefs that you ought to be consistent is not sufficient for you to become consistent. Savage knows that his initial preferences are inconsistent but does not know which of the two preference to throw out. Expected utility does not tell you which of the initial preferences to throw out. It can help you work out that if you prefer Gamble 1 to Gamble 2 then you prefer Gamble 3 to Gamble 4, but cannot help you work out which preference to throw out, the preference for Gamble 1 over Gamble 2 or the preference for Gamble 4 over Gamble 3 on the basis of your belief that they are inconsistent. Moreover, although Savage has most probably been told that his initial preferences violate the sure-thing principle, this might not always be the case: we can plausibly violate the axioms of expected utility without being aware that we do so. The belief that we ought to satisfy them is of no help if we do not have a legitimate way of discovering the error to which, we are are supposed to be subjected to. The implication is that having the second-order belief that you ought be consistent is not a sufficient condition to become consistent.

Neither is it necessary. Savage believed that his initial preferences cannot be held together after finding that they violate the sure-thing principle and has possibly formed the belief that 'My preference for Gamble 1 to Gamble 2 and my preference for Gamble 4 to Gamble 3 are inconsistent', but then he found a *legitimate* way of looking at the two situations, which resulted in a new set of preferences. Savage derives this new set of preferences from premises that he feels more confident about – the preference knowing that E obtains – and so reverses one of his initial preferences

that Allais predicted. So my reading of 'make complicated decisions depend on simpler ones' is that Savage made the complicated decisions between Gamble 1 and Gamble 2 in Situation 1 and between Gamble 3 and Gamble 4 in Situation 2 depend on the simpler one between two new gambles knowing that E obtains in Situation 3 and did not have to rely on any type of second-order beliefs.

3.4 The rule-following approach

The legitimate way in which Savage reaches his new preferences justifies why Savage feels more confident about them and willing to revise with them his initial preferences. This way of interpreting Savage's response reminds of Gilbert's argument in his seminar article *"How mental systems believe."* in which he argues that the main role of conscious reasoning is to help us *doubt* or *unbelieve* what we initially think or judge to be true (Gilbert, 1991). Kahneman who wants to relate Gilbert's work to his own analysis in terms of the intuitive System 1 and the rule-following System 2 notes:

> "The initial attempt to believe is an automatic operation of System 1, which involves the construction of the best possible interpretation of the situation. Even a nonsensical statement, Gilbert argues, will evoke initial belief [...] The moral is significant: when System 2 is otherwise engaged, we will believe almost anything. System 1 is gullible and biased to believe, System 2 is in charge of doubting and unbelieving" (Kahneman, 2011, p. 81)

But Savage was aware that his preferences violated the sure-thing principle. The interest of this case is to check whether Savage *did* or *could* reason his way out of paradox without relying on any form of second-order reasoning that is, if he could *discover* that the preferences for Gamble 1 and for Gamble 4 are inconsistent without knowing that the sure-thing principle is violated. To do so, I use the formal analysis in Chapter 2 to develop an account of reasoning in which Savage can: (i) *discover* that his initial preferences are, in some sense in 'error', because if he applies a legitimate rules of reasoning on these preferences he ends up in a contradiction; (ii) *derive* a new preference between Gambles 3 and 4 by reasoning from a pre-existing preference 'knowing event E obtains'; and (iii) *use* these preferences to reason his way out of the paradox.

To do so, I adopt Savage's primitive into my model and reconstruct his response. Let Z be a non-empty set of consequences and W a non-empty set of exhaustive

states of the world. We assume Z and W are finite for simplicity. In this case, the gambles' outcomes form the set of consequences and the tickets numbered 1–100 in the lottery describe the set of all states of the world. An act $f : W \to Z$ maps states into consequences. So $f(w)$ is the consequence of choosing act f if the state of the world is w. An act is constant if the consequence of choosing act f is the same in all states of the world, i.e. $f(w) = z$ for all $w \in W$. I identify constant acts with elements of X in Chapter 2. Let F denote the set of all acts. The decision maker has intentions and preferences (strict preferences and indifferences) over acts, and forms beliefs that certain acts from F are feasible. Let $Y \subseteq F$ be the non-empty subset of choice acts that are feasible. Sets $\mathcal{L} = F \cup 2^F \setminus \{\emptyset\}$ and $\mathcal{A} = \{\succ, \sim, bel, int\}$ are all I need to represent Savage's framework. I enrich this framework allowing the decision maker to form comparative beliefs over events and form 'a preference knowing event E obtains' which are derived notions in Savage's framework. I take them as primitive attitudes to formalise the sure-thing principle which in Savage's original framework is an informal principle. Let $E \subseteq W$ denote a set of states. I denote the complement of a set E by $\bar{E}$. The belief that one event is more probable than (or as probable as) the other is denoted $>$ $(=)$. The preference (indifference) between a pair of acts knowing that some event E obtains is denoted $\succ_E$ $(\sim_E)$. Let a constitution C be a set of attitudes at any given point in time. I can now state formally the sure thing principle.[5][6]

- For any $f, g \in F$ and $E \subseteq W$, if $[(f, g, \succ_E) \in C$ or $(f, g, \sim_E) \in C]$ and $[(f, g, \succ_{\bar{E}}) \in C$ or $(f, g, \sim_{\bar{E}}) \in C]$ then $[(f, g, \succ) \in C$ or $(f, g, \sim) \in C]$

Moreover,

- For any $f, g \in F$ and $E \subseteq W$, if $[(f, g, \succ_E) \in C]$ and $[(f, g, \succ_{\bar{E}}) \in C$ or $(f, g, \sim_{\bar{E}}) \in C]$ then $(f, g, \succ) \in C$

The sure-thing principle recommends to ignore the events where outcomes are the same. Outcomes are the same where acts agree. Following Savage, f agrees with g in event E if $f(w) = g(w)$ for every $w \in W$. I write this proposition $f = g$ in E. Let $V = F \times F \times 2^W$ denote the collection of all such propositions. The agent can form

[5] The sure-thing principle, has meaning provided that E is non-null. Given null event, acts play no role for the final decision. According to Savage, event E is null iff you are indifferent between act f and g conditional on E for every f, g. In our framework, we allow (E might occur, *bel*) to be analogous with 'E is non-null'.

[6] The following four subschemas imply the sure-thing principle:

1. if $(f, g, \succ_E) \in C$ and $(f, g, \succ_{\bar{E}}) \in C$ then $(f, g, \succ) \in C$.
2. if $(f, g, \succ_E) \in C$ and $(f, g, \sim_{\bar{E}}) \in C$ then $(f, g, \succ) \in C$
3. if $(f, g, \sim_E) \in C$ and $(f, g, \succ_{\bar{E}}) \in C$ then $(f, g, \succ) \in C$
4. if $(f, g, \sim_E) \in C$ and $(f, g, \sim_{\bar{E}}) \in C$ then $(f, g, \sim) \in C$

beliefs towards elements from V. Savage has informally argued that Separability P2 and P3, his second and third axioms, support the sure-thing principle. Savage's initial preferences for Gamble 1 in Situation 1 and for Gamble 4 in Situation 2 violate P2. I can now state separability:

- Separability P2. For any $f, g, f', g' \in F$, $E \subseteq W$, if:

  1. $(f = f' \text{ in } E, bel) \in C$ and $(g = g' \text{ in } E, bel) \in C$
  2. $(f = g \text{ in } \bar{E}, bel) \in C$ and $(f' = g' \text{ in } \bar{E}, bel) \in C$
  3. $(f, g, \succ) \in C$

  Then $(f', g', \succ) \in C$.

Having mapped the main ingredients of Savage's framework into the model, I can now reconstruct Savage's response. To do so, I need to define the concept of a rule, or of a revision following a rule. A rule applies to add a new attitude, the conclusion. Formally, a rule is a pair $s = (M, m)$ of a set of premise attitudes M and a conclusion attitude m. So the revision $C|s$ of a person's constitution C is achieved through a rule s by adding the conclusion provided all premise attitudes are present. For simplicity, the labels f, g, f', and g' will be used for the gambles 1, 2, 3, and 4 respectively. The labels E and $\bar{E}$ will be used for the two events. In Step 1 Savage expresses his initial preferences. According to Separability, a preference for Gamble 1 in Situation 1 implies a preference for Gamble 3 in Situation 2, and similarly, a preference for Gamble 4 in Situation 2 implies a preference for Gamble 2 in Situation 1. Savage's initial preferences for Gamble 1 in Situation 1 and for Gamble 4 in Situation 2 violate the conjunction of Separability and Asymmetry of Preference. Based on the typology of requirements presented in Chapter 1, Separability is a closedness requirement. Theorem 1 of Chapter 2 shows that there is some rule of reasoning which achieves Separability. Given that there exists some rule of reasoning which achieves Separability, if Savage uses this rule (or set of rules) of reasoning, given his initial preferences, he will arrive at a constitution which violates a consistency condition (Gamble 1 will be both preferred and less preferred to Gamble 2, and similarly for Gambles 3 and 4).

Savage starts his reasoning in Step 1 expressing his initial preferences for f over g and for g' over f'. These preferences are not the result of explicit reasoning but of intuition. They enter in our model as inputs to System 2 reasoning. Formally, they form the initial constitution $C_0 = \{(f, g, \succ), (g', f', \succ), ...\}$ before any rule from set of System 2 rules $\mathcal{S}$ has been applied, where '...' stands for all other mental states that might be in the constitution. Then Savage notices that the

two acts f and g agree in event $\bar{E}$. So, the belief attitude $(f = g \text{ in } \bar{E}, bel)$ is now in the constitution $C_1 = \{(f, g, \succ), (g', f', \succ), (f = g \text{ in } \bar{E}, bel), ...\}$. A natural rule for deriving the new preference knowing that event E obtains $(f, g, \succ_E)$ from the original preference $(f, g, \succ)$ is $s = (\{(f, g, \succ), (f = g \text{ in } \bar{E}, bel)\}, (f, g, \succ_E))$. Thus Savage can apply rule s to C_1, expanding the constitution to $C_2 = C_1|s = \{(f, g, \succ), (g', f', \succ), (f = g \text{ in } \bar{E}, bel), (f, g, \succ_E), ...\}$. Then he notices that f and f', and g and g' agree in E. So C_2 expands to include $(f = f' \text{ in } E, bel)$ and $(g = g' \text{ in } E, bel)$, giving $C_3 = \{(f, g, \succ), (g', f', \succ), (f = g \text{ in } \bar{E}, bel), (f, g, \succ_E), (f = f' \text{ in } E, bel), (g = g' \text{ in } E, bel), ...\}$. Savage then can apply another rule $s' = (\{(f, g, \succ_E), (f = f' \text{ in } E, bel), (g = g' \text{ in } E, bel)\}, (f', g', \succ_E))$ to C_3 which derives a new preference knowing that event E obtains from the old one. So, $C_4 = C_3|s'$. Then after noticing that f' and g' agree in $\bar{E}$, constitution C_4 is expanded with $(f' = g' \text{ in } \bar{E}, bel)$. Savage can now derive a new preference from the old preference knowing that event E obtains applying rule $s'' = (\{(f' = g' \text{ in } \bar{E}, bel), (f', g', \succ_E)\}, (f', g', \succ))$ to the expanded C_5, $C_6 = C_5|s''$. So by applying this set of rules to the initial C_0 we obtain the constitution $C_6 = \{(f, g, \succ), (g', f', \succ), ..., (f', g', \succ)\}$. Preferences $(f', g', \succ)$ and $(g', f', \succ)$ violate asymmetry of $\succ$. Savage can apply a different set of Separability rules S' to the initial constitution C_0. Savage can first derive the preference $(g', f', \succ_E)$ from the preference $(g', f', \succ)$ after noticing that f' and g' agree in event $\bar{E}$; then derive a new preference $(g, f, \succ_E)$ from the old preference $(g', f', \succ_E)$ after noticing that f and f' and g and g' agree in E; and finally derive the new preference $(g, f, \succ)$ from the old preference $(g, f, \succ_E)$ after noticing that f' and g' agree in $\bar{E}$. So applying the set of rule S' to C_0 Savage obtains the constitution $C' = \{(f, g, \succ), (g', f', \succ), ..., (g, f, \succ)\}$. Now preferences $(f, g, \succ)$ and $(g, f, \succ)$ violate asymmetry of $\succ$. The point is that for any set of Separability rules S, the closure of the initial constitution $C_0|S$ will violate asymmetry of $\succ$.

The next thing to do is to infer that the initial constitution C_0 is inconsistent because if you expand it by legitimate rules of reasoning you end up in a contradiction. So Savage concludes in the belief that C_0 is inconsistent. The reasoning described above is first-order, and concludes with a constitution that is inconsistent. Savage now needs to use second order reasoning, possibly, with a rule like: $\{(f, g, \succ), (g, f, \succ)\}$, (my constitution is inconsistent, *bel*). If Savage's recognition of his inconsistency is a second-order attitude, arriving at it has to involve second-order reasoning. Alternatively, if Savage's recognition of his inconsistency is an unconscious attitude, arriving at it can but does not have to involve second-order reasoning. In any case, having both $(f, g, \succ)$ and $(g, f, \succ)$ creates a sense of dicomfort to Savage. The problem is with his reasoning before applying any rules from $\mathcal{S}$. It is the System

1 reasoning that led to the initial constitution $C_0 = \{(f, g, \succ), (g', f', \succ), ...\}$. C_0 violates a consistency requirement which is derived by Separability and asymmetry of $\succ$: that if $(f, g, \succ) \in C$ and '...' then $(g', f', \succ) \notin C$ where '...' stands for the mental states in part 1 and part 2 of the Separability requirement.[7]

If Savage is to remove the inconsistency, he must remove one of the initial preferences. Does reasoning which starts from the belief that C_0 is inconsistent and the belief that one ought to be consistent helps Savage to remove one of these preferences? Savage knows from expected utility that his initial preferences are inconsistent but does not know which preference to throw out, and as he self-reports: '[he] still feel[s] an intuitive attraction to those preferences'.

Then in Step 4 Savage adopts a way to look again at the two decision situations focusing on event E. He calls back System 1 to express the preference $(f, g, \succ_E)$ and thus the preference $(f', g', \succ_E)$. Savage notices that f and g, and f' and g' agree in $\bar{E}$, and that in either situation it will not matter which gamble is chosen. So Savage is indifferent between f and g in $\bar{E}$ and f' and g' in $\bar{E}$. The indifferences $(f, g, \sim_{\bar{E}})$ and $(f', g', \sim_{\bar{E}})$ are now in his constitution.

The implication is that in Step 5 Savage uses a rule to derive a new preference $\succ$ from $\succ_E$. Savage can apply rule $r' = (\{(f, g, \succ_E), (f, g, \sim_{\bar{E}})\}, (f, g, \succ))$ which derives the preference $(f, g, \succ)$ from the preference $(f, g, \succ_E)$ and rule $r'' = (\{(f', g', \succ_E), (f', g', \sim_{\bar{E}})\}, (f', g', \succ_E))$ which derives the preference $(f', g', \succ)$ from the preference $(f', g', \succ_E)$. The last the step is the step at which Savage removes $(g', f', \succ)$ from his constitution and replaced it it by $(f', g', \succ)$. I explain in Section 2 that Savage replaces his old preference with his new preference for Gamble 3 to Gamble 4 that is derived by reasoning from premises he feels confident about drawing on some sort of 'implicit' or 'automatic' process.

To sum up, as the above analysis illustrates, Savage applies a set of legitimate rules to his constitution that contained his initial preferences, and ends up with an inconsistent pair of preferences. He then concludes that the initial constitution is inconsistent because if you expand it by legitimate rules of reasoning you end up in a contradiction. But in Savage's response to his problem of inconsistency both System 2 and System 1 participate. Savage concludes that the initial constitution is inconsistent because by expanding it by legitimate rules of reasoning he end up in a contradiction, but he does not know which of the two preferences to revise as he still feels they are intuitively appealing. So he brings back System 1 to reconsider his preferences knowing that E obtains. I conclude that Savage's response draws

[7]Indeed, given the sure-thing principle, asymmetry of $\succ$, symmetry of $\sim$, and exclusiveness of $\succ$ and $\sim$, these preferences also violate another consistency requirement. This is for any $f, g \in F$ and $E \subseteq W$, if $(f, g, \succ_E) \in C$ and $(f, g, \sim_{\bar{E}}) \in C$ then $(g, f, \succ) \notin C$.

necessarily on a process of reasoning that goes beyond what I describe as formally correct System 2 reasoning in Chapter 2.

3.5 Conclusion

This chapter re-examined the famous case of Savage's response to the Allais paradox in (1954, pp. 101-103), in which Savage discusses how his initial preferences violated the sure-thing principle and explains how he reversed his preferences and became consistent. I discussed Savage's response in this text and tried, when possible, to reconstruct it with my Broomean model of reasoning. The upshot is that Savage's response draws on a mental process that combines first-order reasoning (System 2) and automatic processes or intuitive thinking (System 1). I then explored an alternative response according to which knowing what rationality axioms there are can help you become more rational, e.g. by creating the intention to remove the preferences that violate these axioms. I argued that this interpretation would not be justified by Savage's own discussion of how he resolved his problem.

3.6 Conclusion of the thesis

I have started this thesis with a question: "can rational choice be reached by System 2?" and the subquestion "how do we, as economists, model it?" I investigated three possible approaches to answering this question that all, in one way or another, incorporate some sort of dual-process model of reasoning for choice. First, there are choice theorists who have implicitly given an affirmative answer by considering expected utility axioms as normative without giving a psychological explanation of how they come to be satisfied. Then there are the dual self models. Choice is the result of the interaction of two types of selves, the System 1 'irrational' self and the System 2 'rational' self that solve a maximisation problem. Their answer is also affirmative identifying System 2 reasoning with economic reasoning.

The novelty of the dual-systems hypothesis, however, is that it explicitly models two possible *complementary* ways-of-thinking, not two possible ways of producing *conflicting* preference schemas. This approach is more faithful to Kahneman's view of the two systems. I presented a new model that is inspired by Broome's own answer to this questions. With it, I have shown that System 2 understood as a conscious, explicit, and rule-guided mental process can achieve many but not all types of rational choice requirements, and particularly cannot achieve consistency requirements. To illustrate my point further, I applied the model to show how

Savage resolved his personal problem of discovering that his preferences over Allais's gambles are inconsistent with the sure-thing principle.

Bibliography

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person.* Cambridge University Press.

Alchourrón, C. E., P. Gärdenfors, and D. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic 50*(2), 510–530.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica 21*(4), 503–546.

Aumann, R. J. (1999). Interactive epistemology i: knowledge. *International Journal of Game Theory 28*(3), 263–300.

Becker, A., T. Dohmen, B. Enke, A. Falk, D. Huffman, U. Sunde, et al. (2018). Global evidence on economic preferences. Technical report, CRC TRR 190 Rationality and Competition.

Berg, N. and G. Gigerenzer (2010). As-if behavioural economics: Neoclassical economics in disguise? *History of Economic Ideas 18*(1), 133–165.

Bernheim, B. D. and A. Rangel (2004). Addiction and cue-triggered decision processes. *American economic review 94*(5), 1558–1590.

Bleichrodt, H., J. L. Pinto, and P. P. Wakker (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management science 47*(11), 1498–1514.

Boghossian, P. (2014). What is inference? *Philosophical Studies 169*(1), 1–18.

Bratman, M. (1987). *Intention, plans, and practical reason.* Harvard University Press.

Brocas, I. and J. D. Carrillo (2008). The brain as a hierarchical organization. *American Economic Review 98*(4), 1312–46.

Brocas, I. and J. D. Carrillo (2011). A neuroeconomic theory of memory retrieval. Technical report.

Brocas, I. and J. D. Carrillo (2012). From perception to action: an economic model of brain processes. *Games and Economic Behavior 75*(1), 81–103.

Broome, J. (2007). Wide or narrow scope? *Mind 116*(462), 359–370.

Broome, J. (2010). The unity of reasoning? In *Spheres of Reason: New Essays in the Philosophy of Normativity*. Oxford University Press.

Broome, J. (2013). *Rationality through reasoning*. John Wiley & Sons.

Cerigioni, F. (2017). Dual decision processes: retrieving preferences when some choices are intuitive. Technical report.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *The American Economic Review 103*(2), 690–731.

Cubitt, R., C. Starmer, and R. Sugden (1998). Dynamic choice and the common ratio effect: An experimental investigation. *The Economic Journal 108*(450), 1362–1380.

Cubitt, R. P. and R. Sugden (2003). Common knowledge, salience and convention: a reconstruction of david lewis's game theory. *Economics and Philosophy 19*(02), 175–210.

Cubitt, R. P. and R. Sugden (2014). Common reasoning in games: A lewisian analysis of common knowledge of rationality. *Economics and Philosophy 30*(3), 285–329.

Dahl, O. (2013). Stuck in the futureless zone. *Diversity Linguistics comment Posted 03/09/2013 http://dlchypothesesorg/360*.

Dietrich, F. and C. List (2016). Reason-based choice and context-dependence: An explanatory framework. *Economics & Philosophy 32*(2), 175–229.

Dietrich, F., A. Staras, and R. Sugden (2018). Beyond belief: logic in multiple attitudes. *working paper*.

Dietrich, F., A. Staras, and R. Sugden (2019). A broomean model of rationality and reasoning. *forthcoming in The Journal of Philosophy*.

Esponda, I. and E. Vespa (2014). Hypothetical thinking and information extraction in the laboratory. *American Economic Journal: Microeconomics 6*(4), 180–202.

Esponda, I. and E. Vespa (2017). Contingent preferences and the sure-thing principle: Revisiting classic anomalies in the laboratory. *working paper*.

Fudenberg, D. and D. K. Levine (2006). A dual-self model of impulse control. *American economic review 96*(5), 1449–1476.

Fudenberg, D. and D. K. Levine (2012). Timing and self-control. *Econometrica 80*(1), 1–42.

Fudenberg, D., D. K. Levine, and Z. Maniadis (2014). An approximate dual-self model and paradoxes of choice under risk. *Journal of Economic Psychology 41*, 55–67.

Galor, O., Ö. Özak, and A. Sarid (2017). Geographical origins and economic consequences of language structures.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. The MIT press.

Gilbert, D. T. (1991). How mental systems believe. *American psychologist 46*(2), 107.

Gilboa, I. (2009). *Theory of decision under uncertainty*, Volume 1. Cambridge university press Cambridge.

Gilboa, I. (2010). Questions in decision theory. *Annu. Rev. Econ. 2*(1), 1–19.

Gilboa, I., A. Postlewaite, and D. Schmeidler (2009). Is it always rational to satisfy savage's axioms? *Economics and Philosophy 25*(03), 285–296.

Gilboa, I., A. Postlewaite, and D. Schmeidler (2012). Rationality of belief or: why savage's axioms are neither necessary nor sufficient for rationality. *Synthese 187*(1), 11–31.

Gilboa, I. and D. Schmeidler (1995). Case-based decision theory. *The Quarterly Journal of Economics 110*(3), 605–639.

Gilboa, I. and F. Wang (2018). On deciding when to decide. *working paper*.

Grüne-Yanoff, T. and S. O. Hansson (2009). *Preference change: Approaches from philosophy, economics and psychology*, Volume 42. Springer Science & Business Media.

Hammond, P. J. and H. Zank (2014). Rationality and dynamic consistency under risk and uncertainty. In *Handbook of the Economics of Risk and Uncertainty*, Volume 1, pp. 41–97. Elsevier.

Horowitz, J. K. (1991). Discounting money payoffs: An experimental analysis. *Handbook of behavioral economics 2*, 309–324.

Infante, G., G. Lecouteux, and R. Sugden (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology 23*(1), 1–25.

Kahneman, D. (2003a). Maps of bounded rationality: Psychology for behavioral economics. *American economic review 93*(5), 1449–1475.

Kahneman, D. (2003b). A perspective on judgment and choice: mapping bounded rationality. *American psychologist 58*(9), 697.

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

Kolodny, N. (2005). Why be rational? *Mind 114*(455), 509–563.

Kolodny, N. (2007). State or process requirements? *Mind 116*(462), 371–385.

Kreps, D. (1988). *Notes on the Theory of Choice*. Westview press.

Loewenstein, G. and D. Prelec (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics 107*(2), 573–597.

Loewenstein, G. F. (1988). Frames of mind in intertemporal choice. *Management Science 34*(2), 200–214.

Manzini, P. and M. Mariotti (2012). Categorize then choose: Boundedly rational choice and welfare. *Journal of the European Economic Association 10*(5), 1141–1165.

McClennen, E. F. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge university press.

Mongin, P. (2018). The allais paradox: What it became, what it really was, what it now suggests to us.

Nisbett, R. E. and T. D. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review 84*(3), 231.

Roberts, S. G., J. Winters, and K. Chen (2015). Future tense and economic decisions: controlling for cultural evolution. *PloS one 10*(7), e0132145.

Rubinstein, A. (1988). Similarity and decision-making under risk (is there a utility theory resolution to the allais paradox?). *Journal of economic theory 46*(1), 145–153.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica 5*(17), 61–71.

Savage, L. J. (1954). *The foundations of statistics*. Wiley.

Sen, A. (1973). Behaviour and the concept of preference. *Economica 40*(159), 241–259.

Shafir, E. and A. Tversky (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive psychology 24*(4), 449–474.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics 69*(1), 99–118.

Slovic, P. and A. Tversky (1974). Who accepts savage's axiom? *Systems Research and Behavioral Science 19*(6), 368–373.

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

Sugden, R. (1991). Rational choice: a survey of contributions from economics and philosophy. *The Economic Journal 101*(407), 751–785.

Sugden, R. (2006). Hume's non-instrumental and non-propositional decision theory. *Economics & Philosophy 22*(3), 365–391.

Sugden, R. (2018). *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford University Press.

Sunstein, C. R. (2014). *Why nudge?: The politics of libertarian paternalism*. Yale University Press.

Sutter, M., S. Angerer, D. Rützler, and P. Lergetporer (2015). The effect of language on economic behavior: Experimental evidence from children's intertemporal choices.

Tabellini, G. (2008). Vpresidential address institutions and culturev. *Journal of the European Economic Association, MIT Press 6*, 255–294.

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

Thaler, R. H. and C. R. Sunstein (2008). *Nudge: Improving decisions about health, wealth, and happiness*. HeinOnline.

Wason, P. C. and J. S. B. Evans (1974). Dual processes in reasoning? *Cognition 3*(2), 141–154.

Wilson, T. D. and Y. Bar-Anan (2008). The unseen mind. *Science 321*(5892), 1046–1047.