A Broomean model of rationality and reasoning

final version[1]
June 2019

Franz Dietrich — Paris School of Economics & CNRS
Antonios Staras — U. of East Anglia
Robert Sugden — U. of East Anglia

John Broome's (2007, 2013) widely discussed philosophical theory of rationality and reasoning is strikingly different from ordinary rational choice theory. Firstly, it focuses on psychology and mental states – against the behaviourist revealed-preference tradition, but in line with the recent trend towards 'behavioural economics' which takes psychology more seriously.[2] Secondly, it explicitly addresses the process of reasoning by which mental states change, the blind spot of choice theory. Thirdly, it is largely informal, which explains why choice theory has so far managed to ignore it, despite its strong interest to philosophers (for example, Parfit and Broome 1997, Kolodny 2005, 2007, Wright 2014, Boghossian 2016, Cullity 2016, Pettit 2016, Southwood 2016).

In this paper, we formalize Broome's framework, in the hope to open it towards choice theorists, behavioural economists, and logicians; and we present sharp results about Broome's central question of whether we can become more rational through reasoning. First though we should give some context.

Broome understands rationality as a static concept, and reasoning as a dynamic concept. The rationality of a human agent ('you' in his language) is taken to consist in 'order' or 'coherence' in your mind (2013: 152). You are rational if your current mental states or attitudes – your beliefs, preferences, intentions, and so on – cohere with one another. This understanding of rationality corresponds with what many others would call 'formal' as opposed to 'substantive' rationality. Reasoning, by contrast, is regarded as a rule-governed mental process of forming new attitudes from existing ones, through consciously operating on the contents of your attitudes. Reasoning is

[1] We thank John Broome for extensive feedback. We also thank an anonymous referee for valuable advice. This paper was presented widely, for instance in the seminar *Decision, Rationality, Interaction* in Paris, France (May 2018), the *Philosophy Seminar* at University of Milano, Italy (May, 2018), the *Economics and Philosophy Congress 'Norms and Normativity'*, Lyon, France (June 2018), the conference *Algebraic Proof Theory For Social Behaviour* at Delft University of Technology, The Netherlands (July 2018), and the *Economics Seminar* at Glasgow University, Adam Smith Business School, U.K. (October 2018). We thank the audiences for their comments. Franz Dietrich acknowledges support by the French National Research Agency through the grants "Coping With Heterogeneous Opinions" (ANR-17-CE26-0003) and "Collective Attitude Formation"(ANR-16-FRAL-0010) and through an EUR grant.

[2] The debate between behaviourism and mentalism is ongoing; e.g., Hausman (2007, 2012), Dietrich and List (2016a) and Okasha (2016).

seen as a mental 'act', as opposed to other causal psychological processes which are automatic and subpersonal.

To what extent can you become rational through active reasoning? This is Broome's central question. The question poses a problem for those "who write on rationality [and] seem to think that they have finished their job when they have described requirements of rationality" (2013: 208). Why do many people, including choice theorists and philosophers, think describing rationality is enough? Broome suggests – and we agree – that they take for granted that rationality is achievable through reasoning: they "must believe that, starting from knowledge of a particular requirement, you can reason your way actively to satisfying that requirement" (2013: 208–209). So they simply presuppose a positive answer to Broome's open question.

Broome's question is of existential importance for choice theory and behavioural social sciences. In choice theory, rationality is usually formulated in terms of axioms that impose restrictions on preferences and beliefs.[3] But choice theorists rarely ask how you come to satisfy these axioms.[4] When choice theory is interpreted descriptively, there seems to be an implicit assumption that if agents conform to the axioms, then this conformity is not in need of psychological explanation. But that cannot be right: if rationality is in fact a systematic property of agents, there must be some process of mental operations by which rationality is brought about. When choice theory is interpreted normatively, the principle of 'ought implies can' seems to commit theorists to assuming that rationality is achievable, presumably through reasoning.[5]

Broome's question is also crucial for behavioural social sciences. Behavioural research is discovering many systematic patterns in real choices that violate rationality axioms of choice theory. Many of these patterns can be classified as *context-dependent choice reversals*. The same person, facing (what can plausibly be described as) the same two options in different choice contexts, systematically chooses one option in one context and the other in the other. A common explanatory strategy is to attribute these reversals to 'reasoning errors' or 'absence of reasoning'. Such claims are increasingly being used to justify public policies such as 'nudging' that manipulate contextual features of decision environments with the aim of helping individuals avoid reasoning errors (e.g. Thaler and Sunstein, 2008). Many explanations of context-dependent choice effectively assume the existence of stable preferences defined on some domain, and then invoke context-dependent psychological mechanisms (such the use of rules of thumb, changing perceptions, or lack of self-control) which interpose between those

[3] Choice theorists use attitudes such as preferences and beliefs to explain observed choice behaviour. They work not only with axioms on attitudes (such as transitivity), but sometimes also with axioms on choice behaviour itself (such as contraction constituency). We here focus on axioms on attitudes. Choice theorists usually treat some attitudes as effectively observable (typically preferences) and others as more remote and possibly behaviourally underdetermined (e.g., beliefs in some models of choice under uncertainty).

[4] Lewis's (1969) theory of common knowledge is an important exception. In contrast to what is now standard practice in game theory, his analysis treats players' common knowledge as a product of specific modes of reasoning (see Cubitt and Sugden 2003).

[5] In a canonical statement of rational-choice theory, Savage (1954: 20) says that the main use he would make of its rationality axioms 'is normative, to police my own decision for consistency and, where possible, to make complicated decisions to depend on simpler ones'. This concept of 'policing' seems to presuppose that reasoning can in principle achieve rationality.

preferences and choice.[6] These theories again implicitly presuppose that reasoning can in principle achieve rationality.

Despite all this, Broome's theory has so far had little impact on social science. This comes not only from its informal nature, but perhaps also from Broome's parallel pursuit of two orthogonal lines of investigation: a 'structural' investigation into the concepts of requirements and reasoning, and an (often only tentative) investigation into what precisely is rationally required and when reasoning is intuitively 'correct'. Both dimensions are important, but should be disentangled and addressed separately, we believe.

In this paper, we address the *structural* dimension of Broome's programme. We formalize the central features of Broome's approach, including 'requirements of rationality', 'theories of rationality', and 'reasoning rules' **for rational choice**, while setting aside substantive questions of which theory and reasoning rules are 'correct'. Our model differs from ordinary rational-choice models through focusing on psychology and through modelling the reasoning process, including theoretical reasoning (reasoning with beliefs) and practical reasoning (reasoning towards intentions). In line with Broome and rational choice theory, we shall use the term 'rational' to mean 'formally rational'.

Psychologists and thereafter behavioural social scientists have often classified mental processes into two systems: the fast and automatic 'System 1' which generates impressions, intuitions, feelings, and impulses, and the slow, conscious, and deliberative 'System 2' which, operating on the outputs of System 1, constructs explicit thoughts in an orderly way (Wason and Evans, 1975; Kahneman, 2011: 19–30). Behavioural economists often presuppose without question that System 2 is able to create fully rational preferences and beliefs, and surprisingly do not explicitly describe the reasoning process by which this is supposed to happen.[7] Our Broomean model can be regarded as an explicit description of the reasoning underlying System 2.

Within our model, we derive sharp results about whether specific types of rationality requirements can or cannot be achieved through reasoning. As it turns out, some common types of requirement can be achieved by reasoning, but others can not. Our negative results cast doubts on the mentioned idea that System 2 is able to achieve full rationality. Using a simple version of conventional choice theory as our leading example, we discuss how far our negative results reveal deficiencies in received ideas about rationality, and or gaps in Broome's theory.

1 Key features of Broome's structural theory

Let us give some background that will guide our model. Broome fundamentally distinguishes between 'rationality' and 'reasoning'. Rationality is understood as one of many possible sources of requirements on your mental states. Other sources of requirements might be morality, prudence, fashion, or Catholicism (2013: 26, 116). Broome

[6] See, e.g., Thaler and Sunstein (2008: 40–41), Manzini and Mariotti (2012), and Dietrich and List (2016, 2017). Infante et al. (2016) document the widespread use of the concept of 'reasoning error' in behavioural economics.

[7] See Infante et al. (2016) for a documentation of this claim.

leaves open whether rationality is normative, that is, generates normative 'oughts' (2013: 146). A distinctive feature of rationality is that it requires coherence between mental states: "your mental states [are] properly related to each other" (2013: 152). Broome is primarily concerned with synchronic requirements: requirements on mental states held simultaneously. He takes ordinary rationality requirements to have wide scope: for example, for Broome 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", rather than "If you believe you ought to do F, rationality requires of you that you intend to do F" in which the requirement has a narrow scope (2013: 31–32). Our model is thus designed to represent wide-scope synchronic requirements. Broome's contention that ordinary rationality requirements have wide scope has been questioned, in particular by Kolodny (2005, 2007); we need not take a position in this debate. It is however uncontroversial that the principal rationality requirements found in choice theory are both synchronic and wide-scope, which establishes (at least) the relevance of Broome's notion of rationality and our model thereof.

Unconventionally, Broomean reasoning is reasoning with multiple attitudes. This goes beyond reasoning with beliefs ('theoretical reasoning') and reasoning towards intentions ('practical reasoning'), although Broome gives special importance to these two classic types of reasoning (2013: 250). Broome describes reasoning as "a rule-governed operation on the contents of your conscious attitudes" (2013: 234); as a causal psychological process through which "some of your attitudes cause you to acquire a new attitude" (2013: 225); and as a conscious activity under your control, to be distinguished from automatic psychological processes which also affect attitudes. Structurally, a mental state or attitude is given by (i) its content, a proposition, and (ii) the type of attitude held towards that content, for example belief or intention (2013: 251–252). Your mental experience of reasoning consists in bringing to mind a set of premise-attitudes and then finding (in a way that can be expressed as 'So. . . ') that some conclusion-attitude follows from them. The sense of 'following from' is that of (implicitly or explicitly) being guided by some rule, in a way that "seems right to you"; whether that rule is correct in some external sense is beside the point (2013: 237–238). Crucially, reasoning is an operation on the *contents* of your attitudes: you bring to your consciousness the contents of attitudes (for example, that it rains, or that you take an umbrella), not the attitudes towards these contents (for example, that you believe that it rains, or intend that you take an umbrella). Nonetheless, the attitudes held towards contents do not get lost in reasoning: they determine 'how' you consciously entertain the contents (as beliefs? as intentions? and so on), the exact meaning of which belongs to the philosophically trickiest elements of Broome's theory.

Broome distinguishes higher-order from first-order attitudes; the former are attitudes towards propositions about your own attitudes, such as your intention that you believe something (2013: 236). He treats reasoning with first-order attitudes as the fundamental kind of reasoning, insisting that "higher-order beliefs are not necessary for reasoning" (2013: 236). Thus, for example, a child can carry out modus ponens reasoning on the contents of his beliefs without any awareness of having beliefs, indeed without any concept of 'belief' (2013: 229). Although our model does not rule out higher-order attitudes, our applications concentrate on first-order attitudes.

Against this background, we now present our Broomean model.

2 Mental states formalized

An agent has many mental states, like beliefs, intentions, or preferences. A mental state is an attitude towards something (the 'object' or 'content' of that attitude). An example is the belief that it rains: the attitude is belief, the content is that it rains. Another example is the intention that I swim: the attitude is intention, the content is that I swim.

Each attitude has (i) a number of places and (ii) a domain. We explain both in turn. Belief and intention are one-place attitudes: their object is a single thing. Preference is a two-place attitude, since something is being preferred to something else. A mental state with a one-place attitude is represented by an object-attitude pair (p, a), for example $(I\ swim, intention)$. In general, a state with an n-place attitude is represented by a tuple $(p_1, ..., p_n, a)$ where a is the attitude and $(p_1, ..., p_n)$ its n-ary object. For example, $(I\ swim, I\ eat, preference)$ represents preference of swimming over eating. We use the term 'object' of attitudes in two ways: the state $(p_1, ..., p_n, a)$ is said to have n objects $p_1, ..., p_n$, or to have just one object $(p_1, ..., p_n)$. Both usages are equivalent for one-place attitudes.

The domain of an attitude is its set of possible objects: the belief domain contains whatever can be believed, the intention domain contains whatever can be intended, the preference domain contains the things between which preferences can be held, and so on. Some might claim that any attitude can be held towards anything: you can believe anything, intend anything, and so on. Others might insist that you can intend only things you can bring about.

Formally, our model has two primitives:

- a fixed non-empty set L of *objects*. Philosophers might think of them as propositions, choice-theorists as options, events, or other things. More on this soon.
- a fixed non-empty set A of *attitudes* or more exactly *types of attitudes*, each coming with a domain $D \subseteq L$ of possible objects and a number of places n in $\{1, 2, \ldots\}$. A might contain one-place attitudes of belief *bel* and intention *int* and a two-place preference attitude $\succ$, each with some domain of possible objects. Some or all attitudes might have universal domain of applicability $D = L$. Choice-theoretic models (when recast in our framework) typically involve attitudes with restricted domains: in game theory, players hold preferences w.r.t. final outcomes (or outcome lotteries), intentions w.r.t. own actions, and beliefs w.r.t. actions of opponents.

Definition 1 *A* ***(mental) state or attitude*** *is a tuple* $(p_1, ..., p_n, a)$ *– called the 'attitude of type* a *towards* $p_1, ..., p_n$*' – where* a *is an attitude type in* A*,* n *is its number of places, and* $p_1, ..., p_n$ *belong to its domain. Let* M *be the set of all mental states.*

Note that we say 'attitude' both for 'mental states' in M and 'attitude types' in A. No ambiguity will arise. Mental states whose attitude type is belief (intention, preference, ...) are called belief states (intention states, preference states, ...), or simply beliefs (intentions, preferences, ...).

We call the totality of an agent's mental states at a given time his current 'constitution':

Definition 2 *A* ***(mental) constitution*** *is a set $C \subseteq M$ of mental states; it contains the mental states held by the agent.*

We shall often interpret objects in L as propositions. One might indeed argue that all attitudes are held fundamentally towards propositions: that desiring apples ultimately means desiring that one eats apples, that intending to swim ultimately means intending that one swims, and so on. But one may alternatively take L to consist of sentences, properties, actions, events, states of affairs, or even a mixture of things such as actions (the domain of intention) and events (the domain of belief).

Choice theorists, by contrast, might interpret objects in L as choice options, goods, Savage acts, nature events, strategies, or a mixture thereof.

We now give a philosophical or logical application, followed by a choice-theoretic application.

Application 1: agent-internal language for practical reasoning. Following Broome, think of L as containing all propositions relevant to practical reasoning, such as *it rains* or *I take an umbrella*. Let A contain at least the one-place attitudes of belief *bel* and intention *int*, the two central attitudes in Broome's analysis of practical reasoning. Your constitution might contain mental states such as (*it rains*, *bel*), (*I stay dry*, *int*), (*carrying an umbrella is a means implied by staying dry*, *bel*), and (*I carry an umbrella*, *int*).

Following Broome (2013: 251, 260–261), one might not formalize L: one might let L be an abstract set of primitive objects called 'propositions'. But if one needs a model of propositions – to capture their logical relations and combine them through operations like 'and' – then one can go one of two possible ways (defined precisely in Dietrich et al. 2019). A *set-theoretic* or *extensional* model of propositions lets propositions be sets of possible worlds; so L contains subsets of a given set of possible worlds. A *syntactic* or *intensional* model of propositions lets propositions be sentences; so L contains the sentences of a suitable formal language. This is not to claim that sets of worlds or sentences metaphysically 'are' the contents of attitudes, but that they can formally 'represent' these contents. The intensional model is more Broomean in our view: Broome needs an intensional notion of proposition, as agents might believe (intend, ...) some proposition without believing (intending, ...) an equivalent one.

Application 2: choice under certainty. It is easy to recast the classic choice-theoretic model of choice under certainty within our Broomean model. Instead of recasting the classic model *as such*, we choose a slightly revisionary representation of choice under certainty, which strikes us as being more mentalistic, that is, closer to the actual psychology of (conscious and deliberate) choice. We shall indicate where our representation is revisionary, and how a classical representation would have worked. Consider a fixed non-empty set X of mutually exclusive choice options, such as goods or political candidates. The agent has attitudes of preference and indifference towards options. In each choice context, 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. Choice theory implicitly assumes the feasible set to be known. Our mentalistic model rather takes the agent to have beliefs about what the feasible set is; the model is silent about whether those beliefs are correct. Following Broome, acts of choice are not mental states, but non-mental occurrences caused by particular mental states, namely intentions. So our model contains not acts, but intentions as their mental counterparts. Overall, as objects of attitudes we need:

- options x in X, the objects of intention, preference, and indifference,
- feasible sets Y in $2^X \backslash \{\varnothing\}$, the objects of feasibility beliefs.

Formally, $L = X \cup (2^X \backslash \{\varnothing\})$. Whoever wants attitudes to have propositional content can re-interpret an x as the proposition that x is chosen, and a Y as the proposition that Y is the feasible set.

Next, let $A = \{int, bel, \succ, \sim\}$, where:

- int is a one-place attitude with domain X; (x, int) represents intention to choose x;
- $\succ$ and $\sim$ are two-place attitudes with domain X; $(x, y, \succ)$ represents (strict) preference of x to y, $(x, y, \sim)$ represents indifference between x and y;
- bel is a one-place attitude with domain $2^X \backslash \{\varnothing\}$; (Y, bel) represents belief that the feasible set is Y.

Aside from differences in formalism, this model departs in three substantive ways from classical choice theory (and these differences will later allow for a richer spectrum of interesting rationality requirements):

Departure 1: We work with two attitudes $\succ$ and $\sim$, although choice theorists usually start from a single relation of weak preference rather than two relations of strict preference and indifference. We could have used a single weak-preference attitude $\succsim$ and defined A as $\{int, bel, \succsim\}$. However, within our mentalist framework we prefer working with $\succ$ and $\sim$, because a weak-preference attitude $\succsim$ seems artificial and composite in nature. Mentalistically, the weak preference relation of choice theory is best interpreted as a mathematically convenient representation of the disjunction of two attitudes. Technically, the difference between working with $\succ$ and $\sim$ and working with $\succsim$ is small: using $\succ$ and $\sim$ one can express weak preference (as preference-or-indifference), and conversely, using $\succsim$ one can express both strict preference (as weak preference without indifference) and indifference (as weak preference in both directions). Nonetheless, working with $\succ$ and $\sim$ is more general, because it allows violation of three very basic rationality assumptions, namely asymmetry of preference, symmetry of indifference, and mutual exclusiveness between preference and indifference, defined in R2-R4 below.

Departure 2: We model the agent's feasibility beliefs, whereas choice theorists assume that the feasible set is given and known. We could have followed the same route as choice theory, by removing the attitude bel from A, and thus removing feasibility beliefs from constitutions, with an implicit assumption that the feasible set Y is exogenously given and automatically known. But this would suppress an important element of practical reasoning, and under-describe the psychology of choice.

Departure 3: We take the agent to intend a single option, whereas choice theorists work with a 'choice correspondence' C that specifies for each possible feasible set Y a (non-empty) set $C(Y) \subseteq Y$ of chosen options. A non-singleton choice set $C(Y)$ means that different choices are observed in different occurrences of the same context Y. Such non-uniqueness is usually rationalized by indifference between the options in $C(Y)$. Choice theorists do not ask how an agent arrives at a specific choice if $|C(Y)| \geq 2$; nor do they ask how agents break the tie in case of indifference. By contrast, we care about the formation of intentions to choose, given our mentalistic agenda. Again, we could have matched choice theory if we had wanted to: we could have mimicked non-unique choice by re-defining the object of intention *int* as being, not an option, but a non-empty set of options (a 'coarse option'). Here, intending a set means intending that some of its members be chosen (for example, intending $\{x, y\}$ means intending 'x or y'). Why do we not do this? First, choice theorists would not usually explain non-unique choice by lack of specific intention. Choice theorists are notoriously agnostic about psychological matters, but, if forced to link choice to intentions, many would explain a non-singleton choice set $C(Y)$ in terms of different possible (specific) intentions, not one coarse intention. Second, although it might well happen that we make a specific choice without specific intention – that we eat a red apple while intending to eat *some* apple – such partially conscious action is not our main focus. We focus on fully conscious agents. More on this topic in Section 7.

3 Requirements on mental states

We now formalize the notion of requirement, in its most standard sense aligned with choice theory and endorsed by Broome (that is, the 'synchronic' and 'wide-scope' sense mentioned in Section 1). In our terminology, such requirements are restrictions on your constitution. In this section, we define 'requirements' generically, without yet caring about whether they are imposed by rationality.

Since a requirement classifies constitutions into those satisfying and those violating it, we identify each requirement with the set of constitutions satisfying it:

Definition 3 *A **requirement** is a set R of constitutions; constitutions in R 'satisfy' the requirement, others 'violate' it.*

Among the numerous potential requirements (sets of constitutions), the tautological requirement allows all constitutions ($R = 2^M$) and the contradictory one allows no constitution ($R = \varnothing$).

We now give examples of requirements, more precisely requirement schemas since the requirements involve parameters:

Examples of requirements within Application 1: Broome treats the following properties, or versions thereof, as requirements of rationality:[8]

[8] Broome's non-contradiction requirement is exactly ours (2013: 155). His modus ponens require-

- **Non-contradiction**: you do not believe both p and *not* p, formally $R = \{C : (p, bel) \in C) \Rightarrow (not\ p, bel) \notin C\}$. Parameter: $p \in L$.
- **Modus ponens:** believing p and *if* p *then* q implies believing q, formally $R = \{C : (p, bel), (if\ p\ then\ q, bel) \in C) \Rightarrow (q, bel) \in C\}$. Parameters: $p, q \in L$.
- **Enkrasia (non-akrasia):** believing *obligatorily* p implies intending p, formally $R = \{C : (obligatorily\ p, bel) \in C \Rightarrow (p, int) \in C\}$. Parameter: $p \in L$.
- **Instrumental rationality**: intending p and believing *q is a means implied by p* implies intending q, formally $R = \{C : (p, int), (q\ is\ a\ means\ implied\ by\ p, bel) \in C \Rightarrow (q, int) \in C\}$. Parameters: $p, q \in L$.

If propositions have no formal structure, these definitions of requirements are informal, since then we cannot give formal meaning to composite propositions such as *not* p, *if* p *then* q and *obligatorily* p (where $p, q \in L$). But these definitions can, if wished, be turned into formal definitions by adopting the syntactic (intensional) or set-theoretic (extensional) model of propositions sketched above (see Dietrich et al. 2019 for details).

Examples of requirements within Application 2:

R1: Transitivity. The transitivity schema consists of four subschemas, namely transitivity of strict preference, transitivity of indifference, and two mixed transitivities:

R1$_{\succ}$: Preferring x to y and y to z implies preferring x to z, formally $R = \{C :$ if $(x, y, \succ), (y, z, \succ) \in C$ then $(x, z, \succ) \in C\}$. Parameters: $x, y, z \in X$.

R1$_{\sim}$: Indifference between x and y and between y and z implies indifference between x and z, formally $R = \{C :$ if $(x, y, \sim), (y, z, \sim) \in C$ then $(x, z, \sim) \in C\}$. Parameters: $x, y, z \in X$.

R1$_{\succ,\sim}$: Preference of x to y and indifference between y and z imply preference of x to z, formally $R = \{C :$ if $(x, y, \succ), (y, z, \sim) \in C$ then $(x, z, \succ) \in C\}$. Parameters: $x, y, z \in X$.

R1$_{\sim,\succ}$: Indifference between x and y and preference of y to z imply preference of x to z, formally $R = \{C :$ if $(x, y, \sim), (y, z, \succ) \in C$ then $(x, z, \succ) \in C\}$. Parameters: $x, y, z \in X$.

R2: Asymmetry of $\succ$. Preferring x to y excludes preferring y to x, formally $R = \{C : (x, y, \succ) \in C \Rightarrow (y, x, \succ) \notin C\}$. Parameters: $x, y \in X$.

R3: Symmetry of $\sim$. Indifference between x and y implies indifference between y and x, formally $R = \{C : (x, y, \sim) \in C \Rightarrow (y, x, \sim) \in C\}$. Parameters: $x, y \in X$.

R4: Exclusiveness between $\succ$ and $\sim$. Preference of x to y excludes indifference between x and y, formally $R = \{C : (x, y, \succ) \in C \Rightarrow (x, y, \sim) \notin C\}$. Parameters: $x, y \in X$.

R5: Preference completeness. The options x and y are compared, formally $R = \{C : (x, y, \succ) \in C$ or $(y, x, \succ) \in C$ or $(x, y, \sim) \in C\}$. Parameters: $x, y \in X$.

R6: No conflicting feasibility beliefs. You do not believe both Y and Y' to be the feasible set, formally $R = \{C : (Y, bel) \in C \Rightarrow (Y', bel) \notin C\}$. Parameters: distinct feasible sets $Y, Y' \in 2^X \backslash \{\varnothing\}$.

ment includes as a third premise that *I care whether q* (2013: 157). His instrumental rationality and enkrasia requirements include the additional premise that *I believe that q is up to me* (2013: 169, 171).

R7: No conflicting intentions. You do not intend both x and y, formally $R = \{C : (x, int) \in C \Rightarrow (y, int) \notin C\}$. Parameters: distinct options $x, y \in X$.

R8: Economic enkrasia. If you believe that Y is the feasible set and rank the options in its subset Z top within Y then you intend some option in Z, formally $R = \{C :$ if $(Y, bel) \in C$, $(x, y, \succ) \in C$ for all $x \in Z, y \in Y \backslash Z$, and $(x, z, \sim) \in C$ for all distinct $x, z \in Z$, then $(x, int) \in C$ for some $x \in Z\}$. Parameters: $Y \subseteq X$ and $Z \subseteq Y$ $(Z \neq \varnothing)$.

In conjunction, R1-R8 are interpretable as representing the conventional theory of rational choice under certainty, re-expressed mentalistically. To see this, suppose your constitution satisfies the requirements R1-R8. Conditions R1-R5 ensure fully classical preferences; our non-standard axiomatization R1-R5 comes from using two attitudes, strict preference and indifference, rather than a single weak-preference attitude.[9]

Conditions R6-R7 exclude contradictory intentions or feasibility beliefs. Condition R8 reflects the classical preference-maximization hypothesis: you intend something that you most prefer among what you believe to be feasible. R8 also constitutes a choice-theoretic counterpart of ordinary enkrasia (Application 1), which regulates the formation of intentions. R8 differs from ordinary enkrasia in that intentions respond to preferences, not ought-beliefs.

Notice a crucial consequence of working with intentions rather than (as choice theory does) with choice acts: even if you rank more than one option top in your feasible set (so that $|Z| \geq 2$), you intend a specific option (from Z). Conventional choice theory evades the question of how you manage to choose between top-ranked options; see 'Departure 3' in Section 2. We return to this issue in Section 7.[10]

As has become clear, requirements naturally come in schemas. One might therefore have used the name 'requirement' for *sets (schemas)* of sets $R \subseteq M$, while calling each member an 'instance' of the requirement. This would have turned 'our' requirements into instances of requirements. Nothing hinges on our terminological choice.

Whether reasoning can achieve a requirement depends on the type of requirement. Before we can show this, we first introduce our typology of requirements.

Definition 4 *A* ***consistency requirement*** *is a requirement R that forbids holding certain mental states simultaneously; formally, $R = \{C :$ not $F \subseteq C\}$ for some non-empty set F of mental states, the 'forbidden set'.*

In Application 1, the non-contradiction requirement is a consistency requirement $(F = \{(p, bel), (not\ p, bel)\})$. Application 2 contains consistency requirements in R2

[9] See 'Departure 1' in Section 2. R1 is our counterpart of classic choice-theoretic transitivity, which requires for all $x, y, z \in X$ that weak preference of x to y and of y to z implies weak preference of x to z (formally $R = \{C :$ if $[(x, y, \succ) \in C$ or $(x, y, \sim) \in C]$ & $[(y, z, \succ) \in C$ or $(y, z, \sim) \in C]$ then $[(x, z, \succ) \in C$ or $(x, z, \sim) \in C]\}$). R5 is our counterpart of classic choice-theoretic completeness, which requires for all $x, y \in X$ that x be weakly preferred to y or y to x (formally, $R = \{C : [(x, y, \succ) \in C$ or $(x, y, \sim) \in C]$ or $[(y, x, \succ) \in C$ or $(y, x, \sim) \in C]\}$). R2-R4 have no classic counterparts, as they correspond to properties which the classic approach treats as true by definition (through the way it defines strict preference and indifference from weak preferences). In fact, each of R1 and R5 implies its classic counterpart, and is equivalent to it given R2-R4.

[10] Had we allowed coarse intentions (as discussed under 'Departure 3' in Section 2), we could also have stated a weaker version of R8, by replacing the intention of some element of Z by the intention of some non-empty subset of Z.

$(F = \{(x, y, \succ), (y, x, \succ)\})$, R4 $(F = \{(x, y, \succ), (x, y, \sim)\})$, R6 $(F = \{(Y, bel), (Y', bel)\})$, and R7 $(F = \{(x, int), (y, int)\})$. Our requirements R1-R4 imply preference acyclicity, another schema of consistency requirements.[11]

Definition 5 *A* ***completeness requirement*** *is a requirement R that forbids holding none of certain mental states; formally, $R = \{C : C \cap U \neq \varnothing\}$ for some non-empty set U, the 'unavoidable set'.*

Completeness requirements are non-abstention requirements. Examples are the preference-completeness requirements in R5 $(U = \{(x, y, \succ), (y, x, \succ), (x, y, \sim)\})$. In Application 1 one might require 'belief completeness' relative to certain propositions p; here $U = \{(p, bel), (not\ p, bel)\}$.

Finally, a closedness requirement demands that holding certain (premise) states implies holding a certain (conclusion) state:

Definition 6 *A* ***closedness requirement*** *is a requirement R demanding that if certain mental states are held then a certain mental state is held; formally, $R = \{C : P \subseteq C \Rightarrow c \in C\}$ for some set of ('premise') states P and some ('conclusion') state c.*

Modus ponens, enkrasia and instrumental rationality in Application 1 are schemas of closedness requirements, as are transitivity R1 and symmetry of indifference R3 in Application 2.

Our three types of requirement are mutually exclusive, except for one special case: requirements of the form $R = \{C : m \in C\}$ ('you must hold state m', for given $m \in M$) are expressible both as degenerate completeness requirements $R = \{C : C \cap \{m\} \neq \varnothing\}$ and as degenerate closedness requirements $R = \{C : \varnothing \subseteq C \Rightarrow m \in C\}$. All requirements listed for Applications 1 and 2 fall under our typology, with the important exception of economic enkrasia, which we address and 'categorize' in Section 7.

4 Theories of rationality and their requirements

A theory of rationality expresses a specific conception of what rationality requires of you. We now formalize the notion of 'theory of rationality', without attempting to adjudicate between alternative theories. A brute-force approach to specifying a theory would be to list all its requirements. But formally defining a 'theory of rationality' as a set of requirements would be both unparsimonious and too permissive.[12] We thus identify a theory, not with its set of requirements, but with the set of constitutions it deems rational:

Definition 7 *A* ***notion or theory of rationality*** *is a set T of constitutions; constitutions in T are 'rational' under T, others are 'irrational' under T.*

[11] Preference acyclicity forbids a preference cycle over options $x_1, \ldots, x_n$; so $F = \{(x_1, x_2, \succ), (x_2, x_3, \succ), \ldots, (x_{n-1}, x_n, \succ), (x_n, x_1, \succ)\}$. Parameters: $n \geq 2$ and $x_1, ..., x_n \in X$.

[12] *Unparsimonious*: theories usually have enormously many requirements, most of which are artificial (e.g., whenever R and R' are requirements of the theory, so are their conjunction $R \cap R'$ and disjunction $R \cup R'$). *Too permissive*: viewing any set of requirements as a theory allows for awkward theories, since something can then be required without something logically weaker being required.

A theory of rationality is formally the same object as a requirement: it is a set of constitutions. But that set is interpreted differently: it contains only the fully rational constitutions, not all constitutions satisfying one particular requirement. A theory of rationality implies certain requirements, representing the requirements *of rationality* and defined as follows:

Definition 8 *Given a theory of rationality T, the* ***requirements of*** T *are those requirements R which follow from T, that is, for which $T \subseteq R$.*

The strongest requirement of a theory T is the theory itself $R = T$; the weakest is the tautological requirement, i.e., the set of all constitutions $R = 2^M$. Incidentally, a theory T implies not just requirements, but also permissions. It might permit rather than require transitive preferences.[13]

Our definitional setup has proceeded in a top-down direction, by starting with a notion of rationality and deriving individual requirements. In practice, one proceeds in a bottom-up direction: one first comes up with requirements of rationality, whose conjunction then defines one's theory of rationality. We have already argued that a mentalistic version of the standard theory of rational choice under certainty can be axiomatized by R1–R8. Formally, the intersection of those requirements defines a theory of rationality T. Notice that exactly the same theory could be axiomatized by a different set of requirements. For example, we might replace transitivity R1 by Savage's (1954: 17–21) 'negative transitivity'.[14]

5 Reasoning and becoming rational

Can you become rational through reasoning? This is Broome's key question. We now formalize the notions of 'reasoning' and 'becoming rational'. There are many pre-existing models of changes in beliefs or (more rarely) other attitudes; examples are the 'AGM theory' of belief revision (Alchourrón et al. 1985, Gardenfors 1988), non-Bayesian models of preference revision (for example, Hansson 2001, Grüne-Yanoff and Hansson 2009, Dietrich and List 2012), and models of revising degrees of belief or desire (Jeffrey 1957, Dietrich et al. 2016, Bradley 2017). The machinery introduced here is tailored to Broome's notion of revision, while taking inspiration from logic. As explained in Section 1, reasoning for Broome is a conscious, rule-governed mental process by which you form new attitudes based on existing ones. Broome focuses on active reasoning: reasoning as an act, of which you are the agent. Active reasoning is plausibly explicit, and hence expressed in language, whether loud or in our mind (Broome 2013: 267). A 'reasoning rule' encodes this idea. For example, if you have the intention to swim and the belief that swimming necessitates undressing, you might then reason that 'So I shall undress' (where 'shall' is a linguistic marker of intention

[13] T 'permits' the constitution to be of a given kind (formally, to fall into a given set P of constitutions) if some constitution in T is of that kind (formally, if $P \cap T \neq \varnothing$).

[14] Negative transitivity consists in the requirements $R = \{C : [(x, y, \succ) \notin C \& (y, z, \succ) \notin C] \Rightarrow (x, z, \succ) \notin C)\}$ for all options $x, y, z \in X$. It is equivalent to transitivity R1, given R2-R5. Negative transitivity requirements are requirements of a fourth type: they are 'negative closedness' requirements, given by $R = \{C : C \cap P = \varnothing \Rightarrow c \notin C\}$ for certain $P \subseteq M$ and $c \in M$.

and ‘So’ expresses the internal sense of rule-following). In terms of our model, you apply a rule to (certain attitudes in) your constitution, which adds a new attitude to your constitution. Formally:

Definition 9 *A* ***reasoning rule*** *is a pair $r = (P, c)$ of a set of (premise) states $P \subseteq M$ and a (conclusion) state $c \in M$. The* ***revision of a constitution*** *C* ***through a rule*** *$r = (P, c)$ is the constitution $C|r$ obtained by adding the conclusion state provided all premise states are held, formally*

$$C|r = \begin{cases} C \cup \{c\} & \text{if } P \subseteq C \text{ (the rule `applies' to } C\text{)} \\ C & \text{if } P \nsubseteq C \text{ (the rule `does not apply' to } C\text{).} \end{cases}$$

The rule in the swimming example is $r = (P, c)$, where P contains $(I\ swim, int)$ and $(swimming\ necessitates\ undressing, bel)$, and c is $(I\ undress, int)$. This rule naturally belongs to a schema (set) of rules: all rules which form an intention to ϕ whenever one intends to ψ and believes that ϕ-ing requires ψ-ing, for some pair of acts (ϕ, ψ). Just as requirements, rules typically come in schemas. So, like for requirements, we might alternatively have defined a ‘rule’ as a *schema (set)* of pairs (P, r), where these pairs are the ‘instances’ of the rule. This would turn rules into instances of rules, and schemas of rules into rules. Our current terminology is more convenient, though nothing hinges on it.

Rules in this Broomean sense are restrictive in two ways. First, they create rather than remove attitudes; for instance, no rule removes the belief in a proposition p based on the premise belief in *not* p. Second, premises of rules are attitudes rather than absences of attitudes; for instance, no rule forms an intention based on the absence of other intentions. Both principles follow from Broome’s explicit and conscious account of reasoning.[15] For instance, you can conclude through explicit 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. Readers who prefer using the term ‘reasoning’ in a broader sense that covers mental processes other than explicit reasoning should regard our Broomean model of reasoning as a model of explicit reasoning, and should generally replace our term ‘reasoning’ by ‘explicit reasoning’. By contrast, many existing accounts of revision do allow for removal of attitudes. For example, AGM belief revision (Alchourrón et al. 1985) asks which beliefs to give up in the face of information, and Bradley (2017) analyses the restriction or expansion of degrees of belief or desire.

Your way to reason is defined by the set of rules you follow, to be called your ‘reasoning system’.

Definition 10 *A* ***reasoning system*** *is a set S of reasoning rules. A constitution C is* ***closed under*** *S if for each rule $r = (P, c)$ in S, possession of the premise states implies possession of the conclusion state, that is, $P \subseteq C \Rightarrow c \in C$ (equivalently, $C|r = C$).*

[15] Broome (2013: 278) explicitly denies that you can (correctly) reason towards absences. He is less explicit about reasoning from absences, but it is fair to conclude that such reasoning is also excluded by his account of reasoning. Kolodny (2005: 527–528), in turn, is explicit in rejecting reasoning from absences.

A reasoning system is our counterpart for multiple attitudes of a deductive system for beliefs. If your constitution is not yet closed under your reasoning system S, then reasoning leads to the addition ('formation') of new attitudes, until the constitution is finally closed under S, that is, until each rule in S which applies has been applied. We call the so-reached new constitution the 'revision of C through S':

Definition 11 *The* ***revision (or closure) of*** C ***through a reasoning system*** S *is the constitution* $C|S$ *obtained from* C *by applying rules from* S *until the constitution is closed under* S*. Formally,* $C|S$ *is the minimal extension of* C *closed under* S*.*[16]

Revision through a reasoning system is our counterpart of the familiar operation of deductively closing your belief set, but it applies to your constitution (your full 'psychology') rather than just your beliefs. What is a 'good' reasoning system, given a theory of rationality? We postulate two criteria or desiderata:

- *Desideratum 1:* Reasoning should achieve many of the theory's requirements of rationality.
- *Desideratum 2:* Reasoning should never destroy the consistency of a constitution, a minimal demand defined shortly.

So, loosely speaking, reasoning should improve rationality by Desideratum 1, without elsewhere compromising rationality by Desideratum 2. The following definition makes Desideratum 1 precise:

Definition 12 *A reasoning system* S ***achieves*** *a requirement* R *if for each constitution* C *its revision* $C|S$ *satisfies* R*.*

The next two definitions clarify Desideratum 2:

Definition 13 *Given a theory of rationality* T*, a constitution* C *is* ***consistent*** *if its states can be rationally held together, that is, if some expanded constitution* $C' \supseteq C$ *is rational under* T*.*

How does consistency of a constitution relate to our notion of consistency requirements? As one can show, a constitution is consistent if and only if it satisfies all consistency requirements of the theory.[17]

Definition 14 *Given a theory of rationality, a reasoning system* S ***preserves consistency*** *if for each consistent constitution* C *its revision* $C|S$ *is still consistent.*

6 Achieving (or not achieving) rationality through reasoning

We are now in a position to provide answers to Broome's question. Specifically, we address three subquestions: can reasoning help you satisfy consistency requirements, completeness requirements, and closedness requirements, respectively?

[16] This minimal extension exists and is unique. It is the intersection of all extensions of C closed under S.

[17] First, assume C is inconsistent. Then no rational constitution includes C. So C is the forbidden set of a consistency requirement of T. C violates this requirement. Conversely, if C violates some consistency requirement of T, every extension $C' \supseteq C$ also violates it and is thus irrational, implying inconsistency of C.

6.1 Can you achieve closedness requirements?

Fortunately, a general possibility result holds about closedness requirements:

Theorem 1 *Given any theory of rationality T, there exists a reasoning system which achieves each closedness requirement of T and preserves consistency.*

So, whatever theory of rationality T we adopt, all its closedness requirements are achievable through reasoning. For instance, you can achieve the transitivity requirements R1 in Application 2. The achievability of closedness requirements would be trivial if we did not require consistency-preservation, since then you could use the reasoning system S containing all rules: this would transform each constitution C into the maximal constitution $C|S = M$, which trivially satisfies all closedness requirements, but is inconsistent under any non-degenerate theory.

Typically, many alternative reasoning systems S achieve all closedness requirements of a given theory T (and preserve consistency). The theorem's proof introduces a tailor-made rule for each closedness requirement of T: S contains a rule (P, c) whenever T makes the corresponding requirement $R = \{C : P \subseteq C \Rightarrow c \in C\}$. This reasoning system is unnecessarily rich in rules and unrealistic for agents with limited cognitive skills. It is also peculiar, as it achieves each closedness requirement of T in a single reasoning step, by applying a tailor-made rule. Slimmer and psychologically more natural reasoning systems need more reasoning steps, but still achieve all closedness requirements of T (and preserve consistency). This suggests a fundamental trade-off: the richer the reasoning system is, the more rules the agent must internalize or 'store', but the faster he can form some attitudes.

Theorem 1 formalizes the fundamental truth that reasoning, as understood by Broome, is well adapted to achieving closedness requirements. Intuitively, this comes from the structural analogy between reasoning rules and closedness requirements: closedness requirements have an if-then structure whose antecedent and consequent are that certain attitudes are present, and reasoning is a process by which the presence of certain attitudes causes the presence of another attitude.

6.2 Can you achieve consistency requirements?

The picture reverses as we consider consistency requirements:

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

So reasoning can for instance not achieve requirements of non-contradictory beliefs (Application 1), or of asymmetry of preference R2, no conflicting intentions R7, and preference acyclicity (Application 2).

The proof is simple. Informally, if a constitution violates a consistency requirement, then that requirement is achievable only through removing certain attitudes. Yet Broomean reasoning can only add, not remove attitudes.

Theorem 2's negative result can be traced back to the impossibility to reason towards absences of attitudes, a central Broomean thesis. To see why, notice that a consistency requirement forbids simultaneously holding all attitudes from some set

F, which can be expressed as the if-then requirement that if certain F-states (all F-states except a fixed one) are present then the remaining F-state is absent. One could have achieved this requirement if one could reason from the presence of the former states towards the absence of the latter state. Yet reasoning "cannot conclude in an absence", following Broome.

Broome (2013: 278–280) offers a cautious defence of an apparently similar claim to Theorem 2: through 'correct reasoning', you cannot achieve a requirement which forces you to give up some attitude (a 'consistency requirement', in our terminology). Specifically, he argues that if you hold beliefs in p and also in *not* p, then 'correct reasoning' cannot conclude in non-belief in p. You might derive from your belief in *not* p a positive belief in *I do not believe* p, or in *I ought not to believe* p, but these beliefs crucially differ from non-belief in p, and they come from 'incorrect reasoning', following Broome. Theorem 2 says something simpler and more general: reasoning cannot achieve any consistency requirements simpliciter, regardless of correctness considerations.

Broome argues that where reasoning fails "automatic processes will normally prevent you from having contradictory beliefs" (2013: 279–280). We believe that reasoning can nonetheless play a crucial role in achieving consistency requirements, that is, in removing some attitude from a given set F of mutually inconsistent attitudes within your constitution. We see three routes:

1. Reasoning might remove the grounds on which you hold or had formed some attitude in F. In a second step, the 'ground-less' attitude disappears automatically. There might be many ways to remove the grounds of an attitude. One might be to form the belief that you ought not to hold that attitude. Another might be to form some attitude on whose absence that attitude used to be premised.
2. Plausibly, some attitudes in F are not hard-wired in your constitution. They disappear after a while – they 'expire' or fail to be 'refreshed' – should the conditions for their survival no longer be met. In particular, some attitudes (beliefs, intentions, preferences and so on) might ultimately disappear should you develop certain other attitudes which conflict with them. Reasoning can thus block the survival of attitudes through generating attitudes in conflict with them. The belief that it rains might not be refreshed if you meanwhile form a belief in sunshine.
3. Psychologists often assume that your attitudes are of two sorts, 'explicit' and 'implicit' ones (for example, Wilson et al. 2000). Implicit attitudes might be described as unconscious, never actively formed, or the result of automatic processes or 'System 1'. Explicit attitudes might be described as conscious, actively formed, or the result of explicit reasoning or 'System 2'. Explicit attitudes can sometimes crowd out implicit attitudes if there is some direct conflict between the two (Wilson et al. 2000: 102). By producing explicit attitudes, reasoning can thus make some implicit attitudes disappear, thereby restoring consistency. For example, suppose you hold inconsistent beliefs. By reasoning you come to believe a proposition p, which directly conflicts with your existing belief of *not* p. Your explicit belief of p then makes your (supposedly implicit) belief of *not* p disappear.

6.3 Can you achieve completeness requirements?

Consider a completeness requirement: you should hold at least one mental state from U, formally $R = \{C : C \cap U \neq \varnothing\}$. There is an easy (but, as will turn out, unsatisfactory) way to achieve R: fix an attitude $c \in U$, and adopt the rule to 'always form attitude c', formally the rule $r = (\varnothing, c)$ with an empty premise set. For example, let T be the theory of rationality defined by R1–R8 in Application 2, and consider the preference-completeness requirement in R5 for options x, y, formally $R^* = \{C : (x, y, \succ) \in C$ or $(y, x, \succ) \in C$ or $(x, y, \sim) \in C\}$. Any of the rules $r_1 = (\varnothing, (x, y, \succ))$ ('always come to prefer x to y'), $r_2 = (\varnothing, (y, x, \succ))$ ('always come to prefer y to x') or $r_3 = (\varnothing, (x, y, \sim))$ ('always become indifferent between x and y') would achieve that requirement R^*. But this solution is unsatisfactory for two reasons. Firstly, in many contexts, such arbitrary rules seem unjustified: why for example should one systematically come to prefer x to y in the name of preference completeness? But there is a second problem: while achieving a completeness requirement, such a rule often causes violation of a consistency requirement. Suppose you initially have no preference or indifference between x and y, violating R^*, but you prefer y to another option z, and z to x. So your initial constitution is $C = \{(y, z, \succ), (z, x, \succ), \ldots\}$, where '...' stands for other states. As in the putative solution, let your reasoning system S contain the rule r_1, so that after reasoning the revised constitution is $C|S = \{(x, y, \succ), (y, z, \succ), (z, x, \succ), \ldots\}$. While $C|S$ satisfies the completeness requirement R^*, it violates a preference-acyclicity requirement of the theory (defined in footnote 11). A completeness requirement has been achieved at the cost of a consistency requirement. The rules r_2 and r_3 lead to analogous problems.

In general, when can a rule of type $(\varnothing, m)$ cause a consistency violation? To answer this question, we need the concept of falsifiability:

Definition 15 *Given any theory of rationality T, a mental state m is **falsifiable** if some consistent constitution becomes inconsistent through adding m.*

In plausible theories of rationality, almost all states are falsifiable. For example, (p, bel) in Application 1 is falsifiable because it rules out $(\mathit{not}\ p, \mathit{bel})$; and $(x, y, \succ)$ in Application 2 is falsifiable because it rules out $(y, x, \succ)$, and $(x, y, \sim)$, and the combination of $(y, z, \succ)$ and $(z, x, \succ)$, and so on.

Our general result about completeness requirements states as follows:

Theorem 3 *Given any theory of rationality T,*

(a) some reasoning system achieves all completeness requirements of T, but

(b) no consistency-preserving reasoning system achieves any completeness requirement of T whose U-states are all falsifiable.

The negative result (b) can be traced back to the impossibility to (explicitly) reason from absences of attitudes towards attitudes; see footnote 15. Reasoning from absences would have helped us achieve completeness requirements because such requirements are expressible as if-then requirements where the antecedent is the absence of certain states (all but one of the U-states) and the consequent is the presence of an attitude (the remaining U-state). In sum, the difficulty to achieve consistency requirements

(Theorem 2) and completeness requirements (Theorem 3(b)) stems from the Broomean impossibility to reason towards absences and from absences, respectively.

Why does our negative finding in (b) not contradict our positive finding in Theorem 1 about closedness requirements, although certain degenerate completeness requirements (those with a single U-state) are also special closedness requirements? Part (b) rules out such degenerate completeness requirements because their single U-state is non-falsifiable, as one easily checks.

Although completeness requirements share with consistency requirements the unachievability through reasoning, completeness requirements seem even less under your conscious control. In terms of the dual-system model, the conscious reasoning of System 2 operates on inputs generated by the unconscious processes of System 1. If these inputs are insufficient to allow conscious reasoning to arrive at particular types of conclusion, this does not seem to be a fault of reasoning – nor, one might think, a contravention of rationality. Take the case of the preference completeness requirement of choice theory. 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 that rationality requires you to hold some such preference or indifference, while being silent about which (Hausman 2012: 19).[18] However, banishing completeness requirements altogether from theories of rationality may go too far. As we now argue, certain completeness-like requirements have a strong claim to count as rationality requirements.

7 Economic enkrasia and Buridan's ass

The economic enkrasia requirement in R8 is neither a closedness requirement, nor a consistency requirement, nor a completeness requirement. Whether it is achievable through reasoning is thus not settled by Theorems 1-3. This question is however crucial for forming intentions and making choices. To address it, consider the famous problem 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. Unable to decide between the two, it starves to death. Representing this situation in Application 2 of our model, let the ass face the feasible set $Y = \{left, right, starve\}$, and have the initial constitution $C = \{(left, right, \sim), (left, starve, \succ), (right, starve, \succ), (Y, bel)\}$. In our interpretation, the ass fails to form an intention for $left$ or $right$, which leads to the outcome $starve$, the only feasible outcome requiring no intention.

Could the ass have solved its problem by reasoning? Broome (2013: 189–190, 263–4) argues that the answer is 'Yes'. His solution is strikingly simple: in our terminology, the ass's reasoning system can contain both of the rules $r_1 = (C, (left, int))$ and $r_2 = (C, (right, int))$; and (something we do not engage with) both rules are correct. According to Broome, "there is room for choice in reasoning" (2013: 264), and so the ass is free to choose which of the two rules to apply. Having chosen one, it gets to a bale of hay and survives.

[18] It is sometimes claimed that, if your preferences are not complete and transitive, you are necessarily vulnerable to 'money pumps'. This is incorrect (e.g., Cubitt and Sugden 2001).

Viewed through our model, this solution is problematic, because there is nothing in the domain of reasoning to stop the ass from forming both intentions and, if this happens, reasoning cannot ex-post eliminate one of them. Psychologically, conflicting intentions might have much the same unfortunate effects as the absence of intentions.

In having no intention, the ass effectively violates economic enkrasia R8. Economic enkrasia requirements are completeness requirements of the following generalized type:

Definition 16 *A* ***conditional-completeness requirement*** *is a requirement R whereby possession of certain (premise) states implies possession of at least one of certain possible (conclusion) states; formally, $R = \{C : P \subseteq C \Rightarrow C \cap U \neq \varnothing\}$ for some sets of states P and $U \neq \varnothing$.*

Conditional-completeness requirements indeed generalize completeness requirements, obtained when $P = \varnothing$. Are conditional-completeness requirements achievable through reasoning? Theorem 3 generalizes as follows.

Definition 17 *Given any theory of rationality T and any set $P \subseteq M$ of mental states, a mental state $m \in M \backslash P$ is* ***falsifiable given*** *P if some consistent constitution that includes P becomes inconsistent through adding m.*

Theorem 4 *Given any theory of rationality T,*

(a) some reasoning system achieves all conditional-completeness requirements of T, but

(b) no consistency-preserving reasoning system achieves any conditional-completeness requirement whose U-states are all falsifiable given P.

Illustrating part (a), if the ass's reasoning system contains either (or both) of the rules r_1 and r_2, 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 – which illustrates part (b).

Could different reasoning rules solve this problem? Generalizing the notion of 'rule', might one replace r_1 and r_2 with a single 'rule' that derives $(right, int)$ from the presence of the states in C and the absence of $(left, int)$? But that would be reasoning from an absence, which is un-Broomean. Or (making the strong assumption of negative introspection), one might assume that whenever the ass lacks an intention it believes that it lacks it, and then replace r_1 and r_2 with a single rule that derives $(right, int)$ from the union of C and $\{(\textit{no option is currently intended}, \textit{bel})\}$. But that would require second-order reasoning, which seems contrary to the spirit of Broome's approach, although compatible with our Broomean notion of a 'rule'.

Perhaps the most Broomean response to the problem would be to keep r_1 and r_2 and to argue that automatic processes will induce consistency by eliminating 'surplus' intentions. That at least would have the virtue of offering a unified solution to the difficulty to achieve requirements of consistency, completeness, and conditional completeness: automatic causal processes jump in where reasoning fails.

8 Concluding assessment

The concept of rational choice is often formulated in terms of rationality axioms on preferences and beliefs, but theorists usually neglect to describe the reasoning process by which these rational attitudes come into existence. This paper has developed a formal framework, inspired by Broome's sharp distinction between the notions of reasoning and rationality, to answer the central question of whether reasoning can make us more rational.

In the introduction, we argued that a negative answer to this question would pose serious problems for rational choice theory and for those theories in behavioural social science that explain anomalous choices as resulting from reasoning error. Having broken down this general question into specific questions about different types of rationality requirement, we have arrived at one positive answer (with respect to closedness requirements) and two negative ones (with respect to completeness and consistency requirements). We must therefore conclude that rational choice theory and behavioural social science *do* face a serious problem. This conclusion however hinges on accepting Broome's account of reasoning. If, by contrast, we take the un-Broomean view that reasoning can start from, or generate, absences of attitudes, then the difficulty to become rational disappears: reasoners can then achieve consistency requirements and completeness requirements, over and above closedness requirements. An intermediate view is that there are different kinds of reasoning, and that conscious, explicit reasoning cannot handle absences whereas certain subconscious or implicit forms of reasoning can. Then we are still left with a semi-negative conclusion: reasoners can only become rational by engaging into subconscious or implicit forms of reasoning. So to say, becoming rational requires giving up conscious control over one's mind – still a remarkable conclusion.

Our negative result about completeness requirements is clearly significant in relation to the requirement of complete preferences, a standard axiom in received forms of choice theory. One possible response – a position taken by Hausman (2012: 19–20) – would be to give up preference completeness as a requirement of rationality. Clearly, that would call for major revisions to choice theory. Such revisions might allow us to understand context-dependent choice as evidence, not of error, but of preference incompleteness. It might be argued that, when an individual has to choose between two options but lacks any preference or indifference between them, it is not irrational for her to rely on automatic processes to form an intention to choose one of them, even when those processes are context-dependent (Infante et al., 2016). Accepting that conclusion would undermine many claims about 'reasoning error' made by behavioural scientists. Aside from preference completeness, some completeness-like requirements do seem to be natural requirements of rationality, as we pointed out in our discussion of Buridan's ass. It is therefore questionable whether a plausible theory of rational choice could be entirely free of completeness (or completeness-like) requirements, hence whether the difficulty could be entirely avoided.

Our negative result about consistency requirements reveals a more fundamental problem. It is uncontroversial that many consistency requirements *are* requirements of rationality. In many cases, there exist (un-Broomean) accounts of reasoning – for example, those represented in theories of belief revision – that tell you to remove attitudes where your attitudes are mutually inconsistent (sometimes leaving you some freedom as to *which* attitudes to remove, a potential problem we shall set aside here). Although our Broomean model does not allow removing an attitude, it allows adding the belief that you ought to remove a given attitude, e.g., to remove a belief, preference, or intention which conflicts with your other beliefs, preferences, or intentions. The problem is that this is second-order reasoning. It creates a second-order belief that you ought to remove a first-order attitude, rather than extinguishing that first-order attitude. As Broome points out, we lack an account of how such an ought belief can lead to actual removal of the relevant attitude through conscious reasoning. We have however suggested some possible accounts of how an individual might be able to achieve consistency requirements *with the help of* reasoning. These accounts rely not on reasoning alone, but on an interplay of reasoning and other mental processes. These other processes are fundamentally different: they are causal processes outside our conscious control, and cannot qualify as mental 'acts' in any substantive sense of agency. Calling these processes 'subconscious reasoning' (as proponents of the 'intermediate view' sketched above might do) would stretch the notion of reasoning and would risk trivialising this notion by counting *any* process of mental change as 'reasoning'.

If formal rationality can be characterised as order in the mind – and, like Broome, we think it can – then a convincing understanding of this form of order needs an account of how human minds can create order in themselves. Our tentative conclusion is that such an account requires an analysis of how reasoning interacts with other mental operations, not all of which are under conscious control.

A A general type of requirement

All types of requirements considered above are special cases of a single type, namely a generalized version of conditional-completeness requirements which no longer imposes that $U \neq \varnothing$:

Definition 18 *A* ***unified requirement*** *is a requirement R whereby possession of certain (premise) states implies possession of at least one of certain possible (conclusion) states; formally, $R = \{C : P \subseteq C \Rightarrow C \cap U \neq \varnothing\}$ for some sets of states P and U, not both empty.*

Unified requirements are unificatory in two senses. Firstly, they simultaneously generalize our three earlier requirement classes:

Remark 1 *Unified requirements $R = \{C : P \subseteq C \Rightarrow C \cap U \neq \varnothing\}$ reduce to*

- *completeness requirements if $P = \varnothing$, as then $R = \{C : C \cap U \neq \varnothing\}$,*
- *consistency requirements if $U = \varnothing$, as then $R = \{C : \text{not } P \subseteq C\}$,*
- *closedness requirements if U is a singleton $\{c\}$, as then $R = \{C : P \subseteq C \Rightarrow c \in C\}$.*[19]

[19] Furthermore, unified requirements reduce to negative-closedness requirements (defined in footnote 14), which are obtained if P is a singleton $\{c\}$ as then $R = \{C : c \in C \Rightarrow C \cap U \neq \varnothing\}$ or equivalently $R = \{C : C \cap U = \varnothing \Rightarrow c \notin C\}$.

Secondly, every theory can be axiomatized in terms of unified requirements only:

Theorem 5 *Every theory of rationality T is an intersection (conjunction) of unified requirements.*

Given the generality of unified requirements, it is natural to ask how far they are achievable through reasoning. Our earlier theorems provide the answer. If $U = \varnothing$, our negative conclusion about consistency requirements applies (Theorem 2). If U is singleton, our positive conclusion about closedness requirements applies (Theorem 1). Otherwise our essentially negative conclusion about completeness or conditional-completeness requirements applies (Theorems 3 and 4).

B Proofs

Proof of Theorem 1. Let T be a theory. Define the reasoning system S as containing all rules corresponding to closedness requirements of T. So $S = \{(P, c)$: the closedness requirement given by (P, c) is a requirement of $T\}$. Now consider any initial constitution C and any closedness requirement R of T, given by a pair (P, c).

Claim 1: $C|S$ satisfies the requirement R. This is true because if $P \subseteq C|S$, then $c \in C|S$ as $C|S$ is closed under S which contains (P, c).

Claim 2: S preserves consistency. Assume C is consistent, hence a subset of a rational constitution $C^* \in T$. We show that $C|S \subseteq C^*$ (which establishes consistency). This follows from two facts. The first is that C^* is closed under S, because it is rational and hence in particular satisfies all closedness requirements of T. The second is that $C|S$ is by definition the smallest extension of C closed under S. ■

Proof of Theorem 2. Consider a reasoning system S, and a consistency requirement R, say $R = \{C : \text{not } F \subseteq C\}$. It suffices to specify a constitution C such that $C|S$ violates R. Simply let C be any constitution such that $F \subseteq C$. Since $F \subseteq C$ and $C \subseteq C|S$, we have $F \subseteq C|S$. So the revised constitution $C|S$ violates R. ■

Proof of Theorems 3 and 4. Since Theorem 3 is the special case of Theorem 4 in which $P = \varnothing$, it suffices to prove Theorem 4. Consider any theory T.

(a) By definition, each conditional-completeness requirement of T is conditional on some set P of states, and has at least one U-state. For each such requirement, fix an arbitrary member m_U of U, and define the rule $r_U = (P, m_U)$. Let S be any reasoning system containing one such rule for each conditional-completeness requirement. Clearly, S achieves all conditional-completeness requirements of T.

(b) Consider any conditional-completeness requirement of T, defined by some P and U (where $U \neq \varnothing$). Suppose some consistency-preserving reasoning system S achieves this requirement. Since the requirement is achieved, there must be some $m' \in U$ such that $m' \in P|S$. To complete the proof, it suffices to show that m' is not falsifiable given P. To that end, we consider a consistent $C' \supseteq P$, and must show consistency of $C' \cup \{m'\}$. Because S is consistency-preserving, $C'|S$ is consistent. But $m' \in P|S$ and $P \subseteq C'$ imply $m' \in C'|S$. As $C'|S$ $(= (C'|S) \cup \{m'\})$ is consistent, so is its subset $C' \cup \{m'\}$. ■

Proof of Theorem 5. Consider a theory T. For any constitution $C \notin T$, T makes the requirement not to have that constitution, that is, the 'unique-exclusion' requirement $R_C = \{C' : C' \neq C\}$. The theory is expressible as the conjunction (intersection) of its unique-exclusion requirements: $T = \cap_{C \subseteq M : C \notin T} R_C$. Each R_C is a unified requirement, as it equivalently demands that possession of all states in C implies possession of at least one state not in C, formally $R_C = \{C' : C \subseteq C' \Rightarrow C' \cap (M \cap C) \neq \varnothing\}$. ■

C References

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

Bradley, R. (2017) Decision Theory with a Human Face, Cambridge University Press

Boghossian, P. (2016) Rationality, reasoning and rules: reflections on Broome's rationality through reasoning, Philosophical Studies 173: 3385-3397

Broome, J. (2007) Wide or narrow scope? Mind 116: 360-370

Broome, J. (2013) Rationality through reasoning, Hoboken: Wiley

Broome, J. (2015) Synchronic requirements and diachronic permissions, Canadian Journal of Philosophy 45: 630-646

Cubitt, R., Sugden, R. (2001). On money pumps. Games and Economic Behavior 37 (2001): 121–160.

Cubitt, R., Sugden, R. (2003). Common knowledge, salience and convention: a reconstruction of David Lewis's game theory. Economics and Philosophy 19: 175-210.

Cullity, G. (2016) Describing rationality, Philosophical Studies 173: 3399-3411

Dietrich, F., List, C. (2012) A model of non-informational preference change, Journal of Theoretical Politics 23(2): 145-64, 2011

Dietrich, F., List, C. (2016a) Mentalism versus behaviourism in economics: a philosophy-of-science perspective, Economics and Philosophy 32(3): 249-281

Dietrich, F., List, C. (2016b) Reason-based choice and context-dependence: an explanatory framework, Economics and Philosophy 32(3): 175-229

Dietrich, F., List, C. (2017) What matters and how it matters: a choice-theoretic representation of moral theories, Philosophical Review 126: 421-479

Dietrich, F., List, C., Bradley, R. (2016) Belief revision generalized: A joint characterization of Bayes's and Jeffrey's rules, Journal of Economic Theory 162: 352-371

Dietrich, F., Staras, A., Sugden, R. (2019) Beyond belief: Logic in multiple attitudes, working paper, available at www.franzdietrich.net/Papers/DietrichStarasSugden-BeyondBelief.pdf

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

Hansson, S. O. (2001) The structure of values and norms, Cambridge University Press

Hausman, D. (2007) The Philosophy of Economics: An Anthology, 3rd ed., Cambridge University Press

Hausman, D. (2012) Preference, Value, Choice, and Welfare, Cambridge University Press

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

Jeffrey, R. (1957) Contributions to the theory of inductive probability, PhD Thesis, Princeton University

Kahneman, D. (2011) *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux

Kolodny, N. (2005) Why be rational? Mind 114: 509-563

Kolodny, N. (2007) State or process requirements? Mind 116: 371-385

Lewis, D. (1969) Convention: A Philosophical Study. Cambridge, MA: Harvard University Press

Manzini, P., Mariotti, M. (2012) Categorize then choose: Boundedly rational choice and welfare, Journal of the European Economic Association 10: 939–1213

Okasha, S. (2016) On the interpretation of decision theory, Economics and Philosophy 32(3): 409-433

Parfit, D. (2011) On what matters, Vol. 1, Oxford University Press

Parfit, D., Broome, J. (1997) Reasons and Motivation, Proceedings of the Aristotelian Society 71: 99-146

Pettit, P. (2016) Broome on reasoning and rule-following, Philosophical Studies 173: 3373-3384

Savage, L. J. (1954) The foundations of statistics, New York: Wiley

Southwood, N. (2016) The motivation question, Philosophical Studies 173: 3413-3430

Staffel, J. (2013) Can there be reasoning with degrees of belief? Synthese 190(16): 3535-3551

Tarski, A. (1930) On Fundamental Concepts of Metamathematics

Thaler, R., Sunstein, C. (2008) Nudge: Improving Decisions About Health, Wealth, and Happiness, New Haven, CT: Yale University Press

Wason, P. C., Evans, J. B. (1974) Dual processes in reasoning?, *Cognition* 3(2): 141–54

Wilson, T. D., Lindsey, S., Schooler, T. Y. (2000) A model of dual attitudes, Psychological review 107: 101-126

Wright, C. (2014) Comment on Paul Boghossian, “What is inference?”, Philosophical Studies 169: 27-37