Beyond Belief: Logic In Multiple Attitudes

Franz Dietrich, Antonios Staras, Robert Sugden

Beyond Belief:
Logic in multiple attitudes

draft
March 2019

Franz Dietrich
Paris School of Econ. & CNRS

Antonios Staras
Univ. of East Anglia

Robert Sugden
Univ. of East Anglia

Abstract

Logical models of the mind focus on beliefs, and how one reasons with beliefs. But we also have desires, intentions, preferences, and other attitudes – and arguably we reason with them, particularly when making decisions. To enable a logical analysis of someone's psychology and decision-making, we generalize three classic logical desiderata on beliefs – consistency, completeness, and implication-closedness – towards multiple attitudes. The three resulting 'logical' desiderata on our psychology contrast with the classic notion of 'rationality requirements': requirements of having transitive preferences, non-contradictory beliefs, non-acratic intentions, intentions consistent with preferences, and so on. We prove a theorem that connects the logical desiderata to rationality requirements: each of the three logical desiderata (generalized to multiple attitudes) is equivalent to the satisfaction of a certain class of rationality requirements. This result connects logic with choice theory and psychology, and has implications for whether reasoning can make our attitudes consistent, complete, and closed.

1 Introduction

Logic is extensively used to study our beliefs, and how we form new beliefs through reasoning. But in fact we possess, and reason with, multiple attitudes: beliefs, intentions, desires, preferences, hopes, wishes, and so on. We for instance form intentions based on preferences and beliefs, or preferences based on preferences and beliefs. Such 'multi-attitude reasoning' (as we call it) differs fundamentally from reasoning with beliefs – the sort of reasoning addressed in logic. Logic finds it hard to go beyond belief due to its inherently truth-oriented nature, as explained in Section 2. By contrast, the multiplicity in attitudes has long been recognized and addressed in other disciplines like philosophy, choice theory, psychology, AI theory, and cognitive science, at least when these disciplines study the mind or decision-making.

We aim to help extend the logical analysis of the mind beyond beliefs, towards multiple attitudes. The goal is a 'logical' understanding of psychology, internal reasoning, and action. When are our attitudes consistent? When are they complete? When are they implication-closed? Such questions matter also to artificial intelligence, where the goal is more and more to design intelligent systems or robots that reason and perform actions in different environments. Such an intelligent system cannot properly select its actions if it only possesses beliefs, however perfect ('true' and 'complete') these beliefs might be. The system also needs other attitudes, perhaps preferences or intentions, in order to bridge the gap between beliefs and actions. Believing that a particular action would save a life is not enough of a basis for doing it: an intention (or preference, goal, desire etc.) to save a life is also needed.

We describe an agent not by his belief set, but by his 'constitution' which summarizes all the various attitudes; and we describe his reasoning as a process not just from old to new beliefs, but from old to new attitudes of various kinds. Within this multi-attitude framework, we propose general definitions of when a constitution is consistent, complete, or closed (under implication), and prove a theorem that relates these three logical desiderata to the choice-theoretic or philosophical notion of rationality requirements (examples of which are: transitivity of preferences, non-contradiction of beliefs, non-acrasia of intentions, and intention/preference compatibility). This result has implications about how multi-attitude reasoning can (or not) 'improve' the constitution.

The philosophical foundation of this paper is John Broome's (2013) theory of rationality and reasoning, which we aim to formalise and extend. We shall draw on the formal notions of constitutions, rationality, and multi-attitude introduced in Dietrich *et al.* (2019).

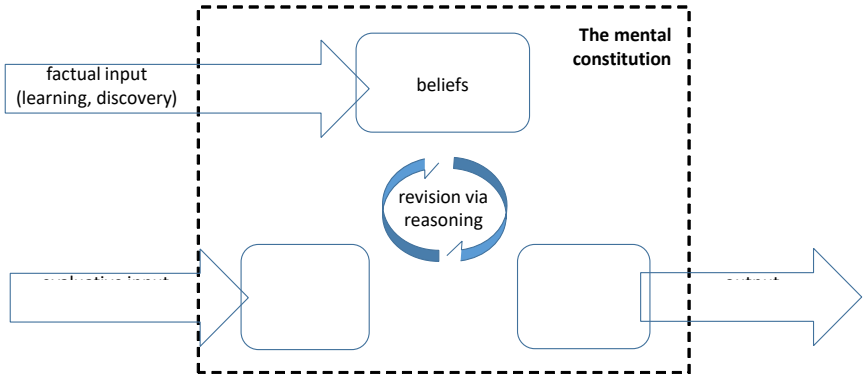

Figure 1 displays the general picture of agency underlying our and Broome's approach, and the broad structure of an intelligent system one might design. On that picture, the agent entertains multiple attitudes (Figure 1 restricts attention to beliefs, preferences, and intentions), which together form his mental constitution. Reasoning leads to revision of the constitution: new attitudes are formed based

on existing ones. Attitudes can also change through external input or learning rather than 'internal' reasoning: beliefs can change through empirical observation, and preferences through 'evaluative' input. By contrast, intentions give rise to behaviour, the 'output' of the cognitive system. This paper sets aside an analysis of external learning and how the constitution should respond to it; this would require a separate revision theory (perhaps a multi-attitude analogue of AGM-style belief revision; see Alchourrón et al. 1985).

Figure 1 is broadly in line with how many AI theorists, psychologists and cognitive scientists construe of agency (e.g., Broersen et al. 2002). [Add literature here, including on BDI-logic, e.g., Wiebe van der Hook.] Three main characteristics distinguish our Broomean model from existing multi-attitude models in AI. First, the agent's diverse attitudes are all modelled in the same unified way (as attitudes towards objects, typically propositions), and analysed as a single system, the 'mental constitution'. Second, we apply logical concepts such as consistency, completeness and closedness to that totality. Third, we construe of reasoning as a process to improve rationality of the mental constitution.

In an important and different sense, logicians *have* addressed multiple attitudes, namely through modal operators representing belief, knowledge, desire, and the like. This approach to multiple attitudes pursues different goals. Inferences in modal logic capture not how the agent forms these attitudes through reasoning, but how a third-party analyst can infer something about the agent's attitudes – still a classical form of belief formation. This is not reasoning with (formation of) multiple attitudes, but reasoning with (formation of) beliefs *about* multiple attitudes of someone else. We come back to this distinction in Section 8.

2 Why does ordinary logic not go beyond beliefs when studying the mind?

This section considers does not yet address those important modal logics in which attitudes are 'internalised' into sentences or propositions through attitude operators (like belief or desire operators). That modal-logical approach to go beyond belief attitudes is addressed in a later section. For now we ask whether logic could go beyond belief in the attitudes taken towards the sentences or propositions of the language, setting aside the modal-logical 'internalisation' of attitudes into sentences.

The difficulty of logic to go beyond belief when studying psychology is routed in the very nature of the logical approach. Logic deals with propositions or facts about the world, expressed by sentences. Logic is 'representational' of an external reality, like belief is and unlike desire. Indeed, not even our beliefs (let alone our desires or intentions) are the primary concern of logic, because logic is primarily

about the world, not the mind. But where logic *does* address the mind, it notoriously addresses beliefs: propositions stand for beliefs (not desires, ...), and logical entailments stand for belief formation, i.e., reasoning with beliefs (not desires, ...). It would be difficult to simply re-interpret logical entailment as a model of desire formation (or intention formation, ...). Could desiring p lead to desiring p *or* q just because p entails p *or* q? Could a tautology p be desired (intended, ...) just because it is logically true, i.e., entailed by anything? One may doubt this.[1] But even if logical entailment were successful in modelling reasoning with desires (or intentions, ...), then we would not have modelled *multi*-attitude reasoning, i.e., reasoning with different attitudes *at a time*. Reasoning with desires is still mono-attitude reasoning, with the unorthodox attitude of desire, not belief. Once we start mixing attitudes, logical entailment fails altogether as a model of reasoning: for instance, desiring p and believing $if\ p\ then\ q$ does not lead to intending q, since the fact that p and $if\ p\ then\ q$ logically entail q is irrelevant to multi-attitude reasoning. In sum, the attempt of reducing the process of reasoning from some attitudes to another to an entailment between the *contents* of these attitudes is shaky if the attitudes are not beliefs , and fails altogether if the attitudes are of different types. Logical entailment also fails altogether when it comes to modelling mono-attitude reasoning with a two-place attitude, held towards pairs of propositions. In particular, reasoning with preferences (Broome 2006) has nothing to do with logical entailment, already because the contents of preferences are proposition *pairs* whereas entailments hold between *single* propositions. Logical entailment can explain why our preferences of p to q and of q to r let us form a preference p to r.

We have just highlighted why logic can hardly go beyond belief in studying the dynamic phenomenon of reasoning. Can logic at least help with the static task of modelling someone's multiple attitudes at a given time? Logicians routinely model someone's static beliefs through a 'belief set' containing all currently believed propositions (or sentences). One could add a 'desire set' containing desired propositions, an 'intention set' containing intended propositions, and so on. This does not lead far however. For one, the three standard logical desiderata on belief sets – logical consistency, completeness, and deductive closedness – become less compelling when applied to desire sets, intentions set, or the like. For another, we need not just care about consistency, closedness or completeness *within* each type of attitude, but also *across* attitudes. This is why we shall generalize these three logical desiderata to desiderata on the totality of an agent's attitudes.

[1] Even reasoning with beliefs need not follow logical entailment: it need not be deductive ('truth-preserving'. This has led to the development of non-monotonic logics and other logics of inductive reasoning. These logics would however not help us much with modelling reasoning with desires or intentions.

3 Our multi-attitude framework

To go beyond an agent's beliefs, we work not with his belief set, but with his constitution. While the belief set contains all currently believed propositions, such as *it snows* or *I can go skying*, the constitution contains all currently held content-attitude pairs or 'mental states', such as $(it\ snows, belief)$ and $(I\ go\ skying, intention)$.

Our framework builds on two simple primitives (discussed in more detail in the philosophical companion paper Dietrich et al. 2019):

- a fixed non-empty set L of *objects* of attitudes. One can think of them as propositions, sentences, or (in choice-theoretic applications) as choice options, moves of other players, 'nature moves', or other constructs.
- a fixed non-empty set A of *attitudes* or more exactly *attitude types*, such as belief, intention, and preference. Each attitude comes with (i) a domain $D \subseteq L$ of possible objects of that attitude, and (ii) a number of places of that attitude $n \in \{1, 2, ...\}$. A might contain 'only' one-place attitudes of belief and intention and a two-place attitude of preference, each with some domain of possible objects. Attitudes in A might have universal domain $D = L$; or one might restrict the domain of intention to propositions under the agent's control, and the domain of belief to propositions beyond the agent's control.

For any one-place attitude a in A and proposition p in its domain, we can form the mental state (p, a), representing attitude a towards p (e.g., belief that p, or desire that p). For any two-place attitudes a in A and propositions p and q in its domain, we can form the mental state (p, q, a), representing attitude a towards (p, q) (e.g., preference of p over q). In general:

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

Terminology: We say 'attitude' not just for attitude types in A (like *desire*), but sometimes also for mental states in M (like *desire that it rains*). Mental states whose attitude type is belief (intention, preference, ...) are called *belief states* (*intention states*, *preference states*, ...), or simply *beliefs* (*intentions*, *preferences*, ...).

Definition 2 *A **(mental) constitution** is a set $C \subseteq M$ of mental states, representing the totality of an agent's current mental states.*

A notion or theory of rationality deems certain constitutions rational, and the other irrational. We thus identify a theory of rationality with the set of constitutions it deemed rational:

Definition 3 *A **notion or theory of rationality** is a set T of constitutions. Constitutions inside (outside) T are called **(ir)rational** according to the theory, or T-**(ir)rational**.*

4 Three applications

We now give two 'logical' applications, and one 'choice-theoretic' application.

Application A: syntactic or intensional model of propositions. Let propositions be defined as sentences: L is the set of all sentences of some suitable formal language, such as the language of classical propositional logic, or a richer language with modal operators and non-material conditionals, or a language of a predicate logic. The syntactic model of propositions is appealing if one thinks that propositions truly are sentences (an implausible metaphysical view) or, more interestingly, if one thinks a proposition is the *meaning* (intension, Sinn) of a sentence and can be formally represented by this sentences itself. In the second case, one has an intensional notion of propositions.

An example of a theory of rationality is the set T of all constitutions C satisfying the following conditions:

- *Modus ponens*: for all $p, q \in L$, if $(p, bel), (if\ p\ then\ q, bel) \in C$ then $(q, bel) \in C\}$.
- *Non-contradictory desires*: for all $p \in L$, if $(p, des) \in C$ then $(not\ p, des) \notin C$.
- *Enkrasia*: for all $p \in L$, if $(obligatorily\ p, bel) \in C$ then $(p, int) \in C$.
- *Necessary means*: for all $p, q \in L$, if $(p, int), (p\ only\ if\ I\ intend\ that\ q, bel) \in C$ then $(q, int) \in C$.

In these conditions the sentences are stated informally; for instance, *not* p stands for $\neg p$, and *obligatorily* p for $O(p)$, where $\neg$ and O are negation and obligation operators of the language, respectively. I have implicitly assumed two things. First, the formal language is sufficiently expressive – has 'enough' operators – for making formal sense of all the sentences considered.[2] Second, the set of attitudes A contains at least the attitudes used above, i.e., the one-place attitudes of belief *bel*, desire *des* and intention *int*, which we take to have universal domain for simplicity. Of course, A might contain other attitudes.

Application B: semantic or extensional model of propositions. Let propositions be defined as sets of possible worlds: L consists of all subsets of a given

[2] To express non-contradictory desires the language must contain a negation operator; for modus ponens it must contain an if-then operator (a material or non-material one, depending on the rendition of the principle); for encrasia it must contain an ought/obligation operator; and for encrasia it must contain an intention operator, and an if-then (or equivalently only-if) operator (which may again be material or not).

set of possible worlds Ω.[3] This model cannot distinguish between logically equivalent propositions: *it neither snows nor rains* and *It is not the case that it snows or rains* are represented by the same set of worlds, hence the same proposition. This can be problematic because mental states often ignore equivalence: we often believe or intend something without believing or intending something equivalent, say out of unawareness of the equivalence. Modelling propositions as sets of worlds reflects an extensional notion of proposition: propositions are taken to be the reference (extension, denoted thing, Bedeutung) of sentences, not the meaning (intension, Sinn) of sentences. Working with sets of worlds rather than sentences is often called a 'semantic' approach. This terminology assumes that semantics is about reference (extension). We hasten to add that one could instead take semantics to be concerned with meaning (intension), on grounds of etymology and natural use. Then Application A rather than B would qualify as semantic.

One can again define a theory of rationality by imposing the four conditions in Application A, i.e., modus ponens, non-contradictory desires, encrasia and necessary means. Formally, this theory T consists of all constitutions C satisfying the four conditions in Application A. Given the extensional notion of propositions, the propositions in these conditions now stand for particular sets of worlds, not for sentences as in Application A. For instance, *not* p stands for $\Omega \backslash p$, and *if* p *then* q (when interpreted materially) for $(\Omega \backslash p) \cup q$. Some propositions in the conditions, such as *obligatorily* p, involve non-truthfunctional operators which cannot be defined through standard set-theoretic operations like complement or union. We therefore need to add appropriate modal operators as part of the semantic model, such as an obligation operator.[4]

Application C: choice under certainty. Let us model choice theory in its simplest version. Consider a fixed non-empty set X of potential choice options, e.g., food options. Our model contains no 'choices': they are not mental states. Instead it contains intentions, the mental counterparts of choices. The agent faces a *feasible set*, the non-empty set $Y \subseteq X$ of currently feasible options; it enters our (mentalistic) model through what the agent *believes* to be the feasible set. It sum,

[3] More generally, L could be some algebra of subsets of Ω. By letting L contain only certain subsets, we can limit attention to propositions that are expressible or accessible to the agent's cognitive system.

[4] A one-place semantic operator F (e.g., an 'obligation' or 'belief' or 'intention' operator) is a function mapping propositions $p \subseteq \Omega$ to propositions $F(p) \subseteq \Omega$. One should think of *obligatorily* p as $F(p)$ for an 'obligation operator' F; $F(p)$ contains the worlds in which it is obligatory that p. A two-place semantic operator (e.g., a non-material if-then operator, or a preference operator) maps pairs of propositions to propositions. Some semantic operator are truth-functional: their output is a set-theoretic ('Boolean') combination of their input. Examples are the 'not' operator given by $F(p) = \Omega \backslash p$, the 'and' operator given by $F(p, q) = p \cap q$, and the material if-then operator given by $F(p, q) = (\Omega \backslash p) \cup q$. But many relevant semantic operators, like the *obligation* operator, are not truth-function; they are modal.

we use mental states of three types:

- (x, int), representing intention to choose option x,
- $(x, y, \succsim)$, representing weak preference of option x to option y,
- (Y, bel), representing belief that the feasible set is Y $(\in 2^X \backslash \{\varnothing\})$.

In the philosophical companion paper we work out an alternative choice-theoretic model which uses not a weak-preference attitude $\succsim$, but instead attitudes $\succ$ and $\sim$ of strict preference and indifference. Working with weak preference is choice-theoretically more common, but philosophically less natural because weak preference is arguably not a 'basic' attitude, but a 'composite' attitude which reduces to having *either* a strict preference *or* an indifference.

Since the objects of attitudes are either options or feasible sets, let L contain all options and all feasible sets: $L = X \cup (2^X \backslash \{\varnothing\})$. Those who think of attitudes as having propositional content can re-interpret any option x as the proposition that x is chosen, and any feasible set Y as the proposition that Y is the feasible set.

The set of attitudes is $A = \{int, bel, \succsim\}$, where

- int is a one-place attitude of intention, with domain X,
- $\succsim$ is a two-place attitude of weak preference, with domain X,
- bel is a one-place attitude of (feasibility) belief, with domain $2^X \backslash \{\varnothing\}$.

The received choice-theoretic view is that a full rationality agent holds transitive and complete preferences and chooses what he most prefers among what is feasible. Translated into our framework, and modulo the difference between 'choice' and 'intention to choose', and between 'feasible' and 'believed-feasible', this orthodox view deems a constitution C to be rational if and only if it satisfies five conditions:

- *Preference transitivity*: of all $x, y, z \in X$, if $(x, y, \succsim), (y, z, \succsim) \in C$ then $(x, z, \succsim) \in C$.
- *Preference completeness*: for all $x, y \in X$, $(x, y, \succsim) \in C$ or $(y, x, \succsim) \in C$.
- *Economic enkrasia* or *preference maximization*: if there is a most preferred believed-feasible option, then one such option is intended. Formally: for all feasible sets $Y \in 2^X \backslash \{\varnothing\}$, if $(Y, bel) \in C$ and there is an $x \in Y$ such that $(x, y, \succsim) \in C$ for all $y \in Y$, then $(x, int) \in C$ for some such $x \in Y$.
- *No conflicting intentions*: There is at most one option $x \in X$ such that $(x, int) \in C$.
- *Determinate feasibility beliefs*: There is exactly one feasible set $Y \in 2^X \backslash \{\varnothing\}$ such that $(Y, bel) \in C$.

So, within our framework we can define the 'classic' theory of rationality as the theory T consisting of those constitutions which satisfy the above conditions. Several other theories of rationality could be advanced, and *have* been advanced within less classical choice theory; they drop or replace some of the conditions, for instance preference completeness.

Our label 'economic encrasia' emphasizes the analogy to encrasia as standardly construed in philosophy. While standard encrasia requires a normative belief to imply an intention, economic encrasia requires particular preferences to imply an intention.

5 Partial forms of rationality: consistency, completeness, closedness

Having a rational constitution is an ideal that we rarely meet. We now introduce three weaker desiderata. They are inspired by three 'logical' desiderata on someone's *beliefs*:

(a) *Consistency.* This says: do not believe mutually inconsistent propositions, i.e., propositions which cannot be simultaneously true. This is a 'global' version of consistency. 'Local' consistency merely says: do not believe a proposition and also its negation.

(b) *Completeness.* In its 'local' version, this says: believe any proposition or its negation. In its 'global' version, it says: believe a member of each set of propositions which are mutually exhaustive, i.e., cannot be simultaneously false. So, believe not only a member of each 'trivial' exhaustive set of type $\{p, not\ p\}$ (as in the local version), but also a member of each 'non-trivial' exhaustive set, including sets of type $\{p, q, \textit{not-}p\ \textit{or}\ \textit{not-}q\}$ and many other sets.

(c) *Closedness.* This says: believe all consequences of your beliefs, i.e., all beliefs that must be true if your existing beliefs are true.

Generalizing these three concepts will allow us to analyse an agent's constitution – his full psychology – from a logical angle. But when should we count his constitution as consistent? As complete? As closed? The three (informal) definitions (a)-(c) cannot be directly translated from beliefs to multiple attitudes. Ultimately this is because we cannot appeal to the notion of 'truth' and 'possible world' in the realm of desires or other non-representational attitudes. Semantic definitions of consistency, (global) completeness, and closedness are unavailable, simply because there is no 'multi-attitude semantics'. Some might try to introduce such semantics; it is unclear whether this can be meaningful, but certainly such semantics would have to look very different. We will instead use the notion of rationality to define the three logical desiderata. So to say, rationality is our substitute for semantics.

To take inspiration from beliefs, let us first see how the definitions (a)-(c) can be re-expressed in terms of rationality of beliefs (we will do this informally; for details see Appendix B). Think of someone's belief set as 'rational' if it is consistent, complete and closed ('closed' actually follows from 'consistent' and 'complete'). This defines rationality of beliefs in terms of the three desiderata, the opposite

of what we wish to do. But the backwards strategy – going from rationality towards the three desiderata – also works for beliefs: one can characterize the three desiderata on beliefs in terms of rationality of beliefs. As shown in Appendix B, a belief set is

- consistent if and only if it becomes rational by suitably adding (zero or more) beliefs;
- complete in the global sense if and only it becomes rational by suitably removing (zero or more) beliefs;
- closed if and only if it contains each belief b which it entails, where 'entails' means equivalently that b belongs to each rational belief set containing at least all current beliefs.

These characterizations of the three desiderata in terms of rationality provide a recipe for extending the desiderata to multiple attitudes, including a recipe for extending 'entailment':

Definition 4 *Given a theory of rationality, a constitution $C \subseteq M$* ***entails*** *a mental state $m \in M$ if all rational extensions $C' \supseteq C$ contain m.*

Definition 5 *Given a theory of rationality, a constitution C is*

- ***consistent*** *if there is a rational constitution $C' \supseteq C$,*
- ***complete*** *if there is a rational constitution $C' \subseteq C$,*
- ***closed*** *if C contains each mental state which it entails.*

These definitions make intuitive sense, since they treat a constitution C as

- *consistent* if the C-states do not rule one another out, i.e., if one can rationally hold all C-states (among other states),
- *complete* if the C-states are sufficient, i.e., if one can rationally hold no other states than C-states,
- *closed* if no other state rationally follows from the C-states.

But do these three concepts truly generalize their classic belief-theoretic counterparts? They do, because they reduce to the classic concepts in the belief-only case where A contains only the belief attitude *bel*, i.e., where constitutions correspond to belief sets. This is established by the following result, re-stated formally in Appendix B:

Proposition 1 *(informal statement) In the belief-only case $A = \{bel\}$, a constitution is*

- *consistent in our sense if and only if the corresponding belief set is classically consistent,*
- *complete in our sense if and only if the corresponding belief set is classically complete (understood globally),*
- *closed in our sense if and only if the corresponding belief set is classically closed.*

6 How do consistency, completeness, and closedness relate to rationality requirements?

While logicians typically focus on 'abstract' desiderata like consistency, completeness or closedness, choice theorists and philosophers instead focus on 'concrete' rationality requirements like modus ponens, preference completeness, necessary means, and the other requirements listed in Section 4. This difference in the approach to rationality is striking. But the two worlds can be linked. Of the various 'concrete' requirements, some have a consistency flavour (e.g., non-contradictory desires); others have a completeness flavour (e.g., preference completeness); and yet others have a closedness flavour (e.g., modus ponens, preference transitivity, and necessary means).

This link holds not just intuitively, but can be turned into a general theorem, to be presented in this section. But first we must define the generic notion of a 'requirement' of rationality, and distinguish between three types of requirements, following Dietrich et al. (2019). In most general terms, a 'requirement' is something which a constitution may satisfy or violate. Therefore we simply identify a requirement with the set of constitutions satisfying it:

Definition 6 *A **requirement** is (just like a theory of rationality) a set R of constitutions; constitutions in R **satisfy** the requirement, others **violate** it.*

Each of the conditions in Application A–C defines a schema requirement. For instance, encrasia defines the requirement $R = \{C : (obligatorily\ p, bel) \in C \Rightarrow (p, int) \in C\}$ for each $p \in L$, and preference completeness defines the requirement $R = \{C : (x, y, \succsim) \in C \text{ or } (y, x, \succsim) \in C\}$ for each $x, y \in X$.

A theory of rationality implies ('makes') a bunch of requirements:

Definition 7 *The **requirements of a given theory of rationality** T (or **rationality requirements**) are those requirements R which follow from T, i.e., for which $T \subseteq R$.*

For instance, consider the 'classical' theory of rationality T in Application C, which deems those constitutions as rational which satisfy all conditions listed there: preference transitivity, preference completeness, and so on. What requirements does this theory make? For one, all the mentioned conditions define (schemas of) requirements of the theory. These requirements function as the 'axioms' used to define the theory. But the theory makes many more requirements: all logical consequences of the 'axioms'.

We distinguish between three salient types of requirements. Most requirements one encounters are of one of these types.

Definition 8 *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 states, the 'forbidden set'.*

Non-contradictory desires is a schema of consistency requirements, with forbidden sets $F = \{(p, des), (not\ p, des)\}$ $(p \in L)$. No conflicting intentions is another consistency requirement, with forbidden set $F = \{(x, int) : x \in X\}$.

Definition 9 *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 of states, the 'unavoidable set'.*

Preference completeness is a schema of completeness requirements, with unavoidable sets $\{(x, y, \succ), (y, x, \succ), (x, y, \sim)\}$ $(x, y, z \in X)$. In Application C one might consider the completeness requirement which demands believing in some feasible set; the unavoidable set is $U = \{(Y, bel) : Y \subseteq 2^X \backslash \{\varnothing\}\}$.

Definition 10 *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.*

There are many schemas of closedness requirements, such as: modus ponens (take $P = \{(p, bel), (if\ p\ then\ q, bel)\}$ and $c = (q, bel)$), preference transitivity (take $P = \{(x, y, \succsim), (y, z, \succsim)\}$ and $c = \{(x, z, \succsim)\}$), and necessary means (take $P = \{(p, int), (p\ only\ if\ I\ intend\ that\ q, bel)\}$ and $c = (q, int)$).

In sum, we now have a way to classify requirements into three types, and to analyse a theory of rationality in terms of consistency, completeness, and closedness requirements it makes.

The following result establishes a tight link between the choice-theoretic notion of rationality requirements and our general notions of rationality, consistency, completeness, and closedness.

Theorem 1 *Given any theory of rationality $T \neq \varnothing$, a constitution C is*

(a) consistent if and only if it satisfies all consistency requirements of T,
(b) complete if and only if it satisfies all completeness requirements of T,
(c) closed if and only if it satisfies all closedness requirements of T,
(d) fully rational if and only if it satisfies all requirements of T.

This result connects the choice-theoretic or philosophical approach focused on rationality requirements (like transitivity and enkrasia) with a 'logical' or 'holistic' approach focused on general structural desiderata like consistency or closedness.

7 Reasoning with multiple attitudes

The constitution of normal people is usually neither rational, nor even consistent, complete, or closed. But we often *reason* in the idea to 'improve' our constitution. This is reasoning with multiple attitudes, not reasoning with beliefs, the ordinary focus of logic. But can such reasoning make us consistent in our attitudes? Or complete? Or closed? Or even fully rational? This enquiry is a cousin of Broome's central question. Broome asks whether reasoning helps us achieve concrete rationality requirements, such as transitivity or encrasia. Broome's question is formally addressed in our companion paper Dietrich et al. (2019). Our present focus is not on achieving concrete rationality requirements, but on achieving the abstract desiderata of consistency, completeness, closedness, or even full rationality. Theorem 1 has related the abstract desiderata with rationality requirements. We should therefore expect parallels between reasoning towards rationality requirements (analysed in the companion paper) and reasoning towards consistency, completeness, closedness, or even full rationality (analysed in this section). Such parallels will indeed emerge.

7.1 Reasoning rules and the revision of constitution

This subsection formalises Broome's notion of multi-attitude reasoning, as also done in the companion paper. 'Reasoning' means forming new attitudes based on existing ones, e.g., forming the intention to study based on the belief that I ought to study (a single premise), or forming this intention based on the intention to make a change and the belief that making a change requires studying (two premises). Formally:

Definition 11 *A **reasoning rule** is a pair (P, c) of a set of ('premise') states $P \subseteq M$ and a ('conclusion') state $c \in M$ (representing the rule of forming state c based on the states P). 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, i.e.*

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

The rules underlying the two informal examples have the conclusion state $c = (I \textit{ study}, \textit{int})$ and the set of premise states given by either $P = \{(I \textit{ ought to study}, \textit{bel})\}$ or $P = \{(I \textit{ make a change}, \textit{int}), (\textit{making a change requires studying}, \textit{bel})\}$. Reasoning from a (weak) preference of x to y and of y to z towards one of x to z uses the rule with set of premise states $P = \{(x, y, \succsim), (y, z, \succsim)\}$ and conclusion state $(x, z, \succsim)$.

An agent's way to reason – his 'attitude formation policy' – is described by the totality of reasoning rules he uses. We call it his 'reasoning system'. Formally:

Definition 12 *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 premises implies possession of the conclusion, i.e., $P \subseteq C \Rightarrow c \in C$, or equivalently, $C|r = C$.*

If one starts with a constitution C, and reasons using the rules in S, one develops new attitudes, until one's constitution is closed under S, i.e., until no rule in S has any effect. We call the so-reached constitution the 'revision of C through S':

Definition 13 *The **revision (or closure) of a constitution C through a reasoning system** S is the constitution $C|S$ obtained from C by adding mental states until the constitution is closed under S. Formally, $C|S$ is the minimal extension of C closed under S.*[5]

The revised constitution $C|S$ can be gradually constructed as follows: first the agent revises C through any rule r from S that is effective, i.e., for which $C|r \neq C$; then he revises $C|r$ through any any other rule s in S that is effective, i.e., for which $C|r|s \neq C|r$; and so forth until no further rule in S is effective. Formally:

Definition 14 *A rule $r = (P, c)$ is **effective on a constitution** C if $C|r \neq C$, i.e., if $P \subseteq C$ and $c \notin C$.*

Remark 1 *The revision $C|S$ of a constitution C through a finite reasoning system S equals the consecutively revised constitution $C|r_1|r_2 \cdots |r_n$ for any maximal sequence $(r_1, ..., r_n)$ of S-rules in which each rule r_i is effective on the previous constitution $C|r_1| \cdots |r_{i-1}$.*[6]

For instance, the revision of a constitution C through a three-rule reasoning system $S = \{r, r', r''\}$ is $C|S = C|r|r'$ if r is effective on C, r' is effective on $C|r$, and r'' is not effective on $C|r|r'$, because (r, r') is then maximal. The order of revision may matter. Indeed, if r' is not yet effective on C, the opposite order of revision yields the different result $C|r'|r = C|r \neq C|S$.

[5] This minimal extension exists and is unique, and it equals the intersecion of all exended constitutions $C' \supseteq C$ that are closed under S.

[6] The sequence $(r_1, ..., r_n)$ is not unique; so one may reasoon in different ways towards $C|S$. But any two such maximal sequences $(r_1, ..., r_n)$ and $(r'_1, ..., r'_m)$ involve the same number of rules (i.e., $m = n$) and the same set of conclusions (i.e., each conclusion of some of $r_1, ..., r_n$ is the conclusion of some of $r'_1, ..., r'_m$, and vice versa).

7.2 Can we become consistent, complete and closed through reasoning?

What would it mean to achieve one of the logical desiderata through reasoning?

Definition 15 *Given a theory of rationality, a reasoning system S* ***achieves consistency, completeness, closedness****, or* ***(full) rationality*** *if for each constitution C the revision $C|S$ is, respectively, consistent, complete, closed, or rational.*

There would be little point in achieving completeness or closedness if one thereby sacrifices consistency, the arguably most basic and 'least sacrificeable' of the three logical desiderata. We shall therefore always want the reasoning system to preserve consistency, in the following sense:

Definition 16 *Given a theory of rationality, a reasoning system S* ***preserves consistency*** *if for each consistent constitution C its revision $C|S$ is still consistent.*

Theorem 1 reduces the achievement of consistency, completeness, or closedness to the achievement of certain types of requirements. Whether these types of requirements are achievable is in turn addressed in the companion paper Dietrich et al. (2017). As it turns out, Theorem 1 combined with results in the companion paper implies that

- reasoning can achieve closedness while preserving consistency,
- reasoning cannot achieve consistency,
- reasoning can achieve completeness, but only while sacrificing consistency.

Here is the formal statement:

Corollary 1 *Given any theory of rationality,*

(a) some reasoning system achieves closedness while preserving consistency,

(b) no reasoning system achieves consistency, unless the theory deems the maximal constitution $C = M$ rational (so is 'degenerate'),

(c) no reasoning system achieves completeness while preserving consistency, unless the theory deems each set of falsifiable states avoidable,

(d) no reasoning system achieves full rationality, unless the theory deems the maximal constitution $C = M$ rational (so is 'degenerate').

The definitions of 'avoidable' and 'falsifiable' in part (c) will be given shortly. The word 'unless' in parts (b)–(d) can be read not only in the sense of 'if it is not the case that', but also in the stronger sense of 'if *and only if* it is not the case that'. So Corollary 1 in fact provides necessary and sufficient conditions for when the theory of rationality permits 'successful reasoning', in the four senses of achieving consistency, completeness, closedness, or full rationality, respectively. In part (c),

the stronger reading of 'unless' however requires that we assume *compactness*: each inconsistent set of states $C \subseteq M$ has a finite inconsistent subset. Compactness holds trivially if M is finite. Compactness is the multi-attitude counterpart of ordinary logical compactness.

We now discuss each part in turn.

Part (a): the achievability of closedness. By part (a), we can always become closed in our attitudes through reasoning, without losing consistency. Why is this so? By Theorem 1, closedness is reached once all closedness *requirements* of the theory are met. A closedness requirement says: if you hold a certain set of states P, then you hold a certain state c. This immediately suggests the reasoning rule $r = (P, c)$. If the reasoning system contains this rule, it achieves the corresponding closedness requirement. The reasoning system consisting of all rules corresponding to closedness requirements of the theory does the job, because it not only achieves closedness, but also, as one can prove, preserves consistency. In practice, the same job can be done by a much slimmer (and cognitively more plausible) reasoning system, containing only the rules corresponding to *certain* closedness requirements of the theory. Suppose for instance rationality requires that believing p and $if\ p\ then\ q$ implies believing q, and also that believing q implies intending r. Then rationality also requires that believing p and $if\ p\ then\ q$ implies intending r. These are three closedness requirements. If your reasoning system contains the rules corresponding to the first two requirements,, i.e., the rules

$$r = (\{(p, bel), (if\ p\ then\ q, bel)\}, (q, bel)) \text{ and } r' = (\{(q, bel)\}, (r, int)),$$

then you do not need the rule corresponding to the third requirement, i.e., the rule

$$r'' = (\{(p, bel), (if\ p\ then\ q, bel)\}, (r, int)),$$

because what r'' achieves (i.e., the third requirement) is also achieved through applying first r and then r'. Presumably, real people have simple and 'natural' rules in their reasoning systems, as would robots programmed to reason.

Part (b): the inachievability of consistency. Part (b) is mathematically trivial, but philosophically disturbing. It is trivial (without even needing Theorem 1) because reasoning adds and never removes mental states, hence never renders any inconsistent constitution consistent. The result is disturbing because consistency is a more basic and 'minimal' normative desideratum than completeness and closedness. One would have hoped that reasoning is *at least* able to repair inconsistencies. Instead reasoning can only create inconsistencies. The only exception is a 'degenerate' theory that deems the maximal constitution $C = M$ rational: here there are no inconsistent constitutions, so that reasoning can do no harm.

Part (c): the inachievability of completeness. Why does part (c) hold? By Theorem 1, completeness is reached once all completeness *requirements* of the

theory are reached. Each completeness requirement demands having at least one state from a given 'unavoidable' set U. Formally:

Definition 17 *Given a theory of rationality, a set of mental states is* ***avoidable*** *if some rational constitution contains none of its states, and* ***unavoidable*** *otherwise.*

Clearly, a constitution satisfies all completeness requirements $R = \{C : C \cap U \neq \varnothing\}$ of the theory just in case it intersects with all unavoidable sets U. For instance, the constitution might have to intersect with $\{(p, bel), (not\ p, bel)\}$, and with $\{(p, int), (q, int), (r, int)\}$, and with many other unavoidable sets. There is a trivial way to achieve this. For each unavoidable set U of the theory, pick an (arbitrary) state m from U, and let the reasoning system contain the 'trivial' rule which always generates state m (formally, that rule is $r = (\varnothing, m)$, with empty set of premises). Through these rules, one acquires members of all unavoidable sets, hence becomes complete.

This solution is unconvincing. It seems ad hoc, if not stubborn and blind, to always adopt a fixed belief (or intention etc.), regardless of one's existing web of other beliefs (and intentions etc.). The problem can be put formally: applying such rules can make the constitution inconsistent, because the newly formed belief (or intention etc.) can clash with preexisting beliefs (or intentions etc.). If a certain set of intentions is unavoidable, but each intention is inconsistent with certain beliefs, then any reasoning rule that generates one of the intentions makes the constitution inconsistent whenever that constitution happens to contain the beliefs inconsistent with that intention.

We have just highlighted the problem that a reasoning rule (P, m) create an inconsistency. This problem however only arises if one can find other states with which m is inconsistent, i.e., if m is 'falsifiable':

Definition 18 *Given a theory of rationality, a mental state m is* ***falsifiable*** *if some consistent constitution becomes inconsistent through adding m.*

Usually most mental states are falsifiable. For instance, as long as rationality requires non-contradictory desires, any desire (p, des) is falsifiable as $\{(p, des), (not\ p, des)\}$ is inconsistent.

Part (c) rules out theories for which all sets of falsifiable states are avoidable, because in such cases all unavoidable sets contain a non-falsifiable state, so that the problem illustrated above cannot arise.

Part (d): the inachievability of full rationality. Since consistency is unachievable by part (b), so is full rationality (which implies consistency). This of course assumes the maximal constitution $C = M$ is irrational. If that constitution is rational, rationality is trivially achievable through the 'maximal' reasoning system which contains all rules and thus generates the maximal constitution.

8 The different way in which modal logic goes beyond belief

Modal logic addresses multiple attitudes through modal operators such as belief operators, desire operators, two-place preference operators, and the like. The kind of attitudes (modal operators) depends on the modal logic in question; often there are only few, but in principle there could be many (see classic expositions of Fagin et al. 1995, ...). Modal logic is important, but pursues different goals than us and Broome. We emphasize three differences:

- **Multiple attitudes 'in' rather than 'towards' propositions.** Modal logic introduces attitudes at a different level, namely in the language for expressing propositions. Modal logic has propositions about multiple attitudes, not multiple attitudes towards propositions. Modal logics can thus express propositions such as *I intend that p* or *I desire that I intend that p*. By contrast, we go beyond belief in the attitude towards propositions.
- **Reasoning 'about' rather than 'with' attitudes.** Modal logics are no different from other logics in that logical entailments capture ordinary reasoning with beliefs, not multi-attitude reasoning (see Section 2). In modal logic, one happens to reasons with beliefs *about someone's attitudes*. This is third-party reasoning about someone's attitudes, not first-person reasoning with one's own attitudes. The difference is real. Multi-attitude reasoning changes the agent in question (who acquires the new attitude), while reasoning about that agent's attitudes changes the analyst (who acquires a belief about the agent's attitudes). Multi-attitude reasoning creates these attitudes, while reasoning about attitudes teaches something about them.
- **Reduction impossible.** There is no formal correspondence, translation, or reduction between reasoning with one's own attitudes and third-party reasoning about these attitudes. Why? Consider first an instance of multi-attitude reasoning: through intending q and believing that q *only if he intends* p, the agent comes to intend p. One is tempted to translate this into third-person reasoning about him through 'internalising' the attitudes into the propositions. This leads to the following inference about his attitudes: *he intends* q; *he believes* q *only if he intends* p; therefore, *he intends* p. The trouble is that this inference is not valid: the premises can be true without the conclusion, as he may be irrational, as was indeed the case before he engaged in his reasoning. One might reply by assuming he must be rational. This renders the third-person inference valid, but assumes away the whole point of multi-attitude reasoning. Already rational agents need no reasoning to improve their rationality. This brings us to another disanalogy between the two sorts of reasoning: reasoning about someone's attitudes is (in most modal-logical systems) reasoning about *rational* attitudes, whereas

multi-attitude reasoning makes sense for irrational agents who strive for rationality. Returning to the attempt of a reduction, let us see why also the converse reduction fails. Consider this instance of reasoning about someone's attitudes: *he intends p*; *he does not both intend p and desire q*; therefore, *he does not desire q*. Here there is not even an apparent translation into first-person multi-attitude reasoning, because in the inference neither the premises nor the conclusion has a structure that would allow 'externalising' an attitude. The do not have the structure *he has attitude a towards p*, for which one could 'externalize' the attitude a and turn the proposition into the mental state (p, a).[7]

A Proof of Theorem 1

Throughout this appendix, we fix a theory of rationality T and a constitution C. Let $T \neq \varnothing$, an assumption needed only for parts (a) and (b). We now prove each part.

Part (a). We prove both directions of implication. We may assume $C \neq \varnothing$, since otherwise C is trivially consistent (as $T \neq \varnothing$) and satisfies all consistency requirements.

- First let C satisfy all consistency requirements of T. We show that C is consistent. Consider the consistency requirement R^* of not holding all states in C: formally, $R^* = \{C' : C \not\subseteq C'\}$. Since C violates R^* while satisfying all consistency requirements of T, R^* cannot be a consistency requirement of T. So some rational constitution $C' \in T$ violates R^*, i.e., $C \subseteq C'$. So C is consistent.
- Conversely, assume C is consistent. Consider any consistency requirement R of T; we must prove that C satisfies it. R takes the form $R = \{C' : F \not\subseteq C'\}$ for some 'forbidden set' F. Being consistent, C has a rational extension C^+. As C^+ is rational, it satisfies all requirements of T. So C^+ satisfies R, i.e., $F \not\subseteq C^+$. As $C \subseteq C^+$, it follows that $F \not\subseteq C$. So C satisfies R.

Part (b). The proof is the 'dual' of that for part (a). We may suppose $C \neq M$,

[7] Even for those special third-person inferences whose premises and conclusion have the special structure – i.e., state the presence of a mental state – the translation makes little sense, because it ignores the fundamental difference between 'forming' and 'inferring presence of' mental states, and also because, according to Broome's account, first-person reasoning is an operation on the contents of mental states, not on (meta) propositions about having such-and-such attitudes towards such-and-such states. Besides, someone who forms (say) an intention based on (say) certain beliefs need not even be aware of having these beliefs and later this intention: he may lack introspection about his mental states. Presumably we often form intentions without meta-level awareness that we so do.

because otherwise C is trivially conmplete (as $T \neq \varnothing$) and satisfies all completeness requirements.

- First let C satisfy all completeness requirements of T. We show that C is complete. Note that C violates the (completeness) requirement of containing a state outside C, i.e., the requirement $R^* = \{C' : (M \backslash C) \cap C' \neq \varnothing\}$. As C satisfies all completeness requirements of T, R^* is not itself a completeness requirement of T. So some rational constitution $C' \in T$ violates R^*; hence $(M \backslash C) \cap C' = \varnothing$, i.e., $C' \subseteq C$. So C is complete.
- Conversely, let C be complete. Let R be any completeness requirement of T; we show that C satisfies it. R requires to hold at least one states of a certain (unavoidable) set U: $R = \{C' : C' \cap U \neq \varnothing\}$. As C is complete, it has a rational subset C^-. Being rational, C^- satisfies all requirements of T. In particular, C^- satisfies R, i.e., $C^- \cap U \neq \varnothing$. Hence, as $C^- \subseteq C$, we have $C \cap U \neq \varnothing$. So C satisfies R.

Part (c). We prove both directions of implication.

- First, assume C satisfies all closedness requirements of T. To show that C is closed, consider a state m entailed by C; we must show that $m \in C$. Consider the closedness requirement R^* given by the set of premise states C and the conclusion state m; formally, $R^* = \{C' : C \subseteq C' \Rightarrow m \in C'\}$. As C entails m, the theory makes the closedness requirement R^*. As C satisfies all closedness requirements of the theory, C must satisfy R^*. Hence, as $C \subseteq C$, we have $m \in C$.
- Conversely, assume C is closed. Consider any closedness requirement R of the theory, say $R = \{C' : P \subseteq C' \Rightarrow c \in C'\}$ for some (premise) set $P \subseteq M$ and some (conclusion) state $c \in M$. To show that C satisfies R, assume $P \subseteq C$; we must prove that $c \in C$. Since R is a requirement of the theory, all rational constitutions which include P contain c, which in turn means that P entails c (by definition of entailment). So the larger set $C \supseteq P$ also entails c (again by definition of entailment). So $c \in C$, as C is closed.

Part (d). Trivially, rationality is equivalent to satisfaction of the theory's strongest requirement $R = T$, which is equivalent to satisfaction of all of the theory's requirements $R \supseteq T$. ■

B Proposition 1 formally re-stated and proved

Our claim to have generalized consistency, completeness and closedness from beliefs towards multiple attitudes rests on Proposition 1, which we now re-state formally and prove. This appendix section assumes the belief-only case $A = \{bel\}$ and the semantic or syntactic model of propositions (see Application A and B). So L

consists of sets of worlds or logical sentences.[8]

A *belief set* is any set of propositions $B \subseteq L$ (the 'believed' propositions). As belief is the only attitude – i.e., $A = \{bel\}$ – constitutions are notational variants of belief sets: to each constitution C corresponds a unique belief set $B = \{p : (p, bel) \in C\}$, and to each belief set B corresponds a unique constitution $C = \{(p, bel) : p \in B\}$.

Following the classical definitions, a belief set $B \subseteq L$ is

- *consistent* if it is consistent, in the sense that $\cap_{b \in B} b \neq \varnothing$ given the semantic model or in the logical sense given the syntactic model, respectively,
- *(deductively) closed* if it contains all $p \in L$ which it entails, in the logical sense given the syntactic model or in the sense that $\cap_{b \in B} b \subseteq p$ given the semantic model, respectively,
- *locally complete* or simply *complete* if it contains a member of each proposition-negation pair, i.e., each pair $\{p, \neg p\} \subseteq L$ given the syntactic model or each pair $\{p, \Omega \backslash p\} \subseteq L$ given the semantic model, respectively,
- *globally complete* if it contains a member of each exhaustive set $Y \subseteq L$. A set $Y \subseteq L$ is *exhaustive* if necessarily at least one member is true. i.e., if its disjunction is tautological or equivalently the set of negations of propositions in Y is inconsistent – in the semantic or syntactic sense, respectively.[9] The simplest exhaustive sets are the proposition-negation pairs. Global completeness implies local completeness, by quantifying over *all* exhaustive sets, not just over proposition-negation pairs. (Another equivalent definition of 'globally complete' is given in Lemma 1(b).)

The four conditions on belief sets are far from independent: any consistent and complete belief set is automatically deductively closed and globally complete. The gold standard of rational beliefs in logic is to satisfy all these conditions. Translated this logical gold standard into our framework (with the belief-only case $A = \{bel\}$), a constitution is rational just in case the corresponding belief set is consistent and complete (and hence closed and globally complete). We call this 'classical' rationality in the belief-only case. Formally:

Definition 19 *In the belief-only case $A = \{bel\}$ (with the semantic or syntactic*

[8] In the syntactic case we assume that the logic is a standard propositional logic, or more generally any well-behaved logic such as a standard propositional, predicate, modal, or conditional logic. Formally, the logic must obey a few classic conditions (namely L1–L4 in Dietrich 2007) which guarantee 'regular' notions of logical consistency and logical entailment. The notable condition is monotonicity, whereby entailments are preserved under adding premises, and so consistency of a set is preserved under removing elements.

[9] In the semantic case, $\cup_{p \in Y} p = \Omega$ or equivalently $\cap_{p \in Y} (\Omega \backslash p) = \varnothing$. In the syntactic case, $\vee_{p \in Y} p$ is tautological or equivalently $\{\neg p : p \in Y\}$ is inconsistent (if the disjunction $\vee_{p \in Y} p$ is undefined in the logic in question, e.g., because Y is infinite, then only the definition in terms of inconsistency of $\{\neg p : p \in Y\}$ can be used).

*model of L), the **classical** theory or notion of rationality is*

$$T = \{C : \textit{the belief set } \{p : (p, bel) \in C\} \textit{ is consistent \& complete}\}.$$

We are ready to re-state Proposition 1 formally:

Proposition 1 *Under the belief-only case $A = \{bel\}$ (with the semantic or syntactic model of L) and the classical theory of rationality, a constitution is*

- *consistent if and only if the corresponding belief set is consistent,*
- *complete if and only if the corresponding belief set is globally complete,*
- *closed if and only if the corresponding belief set is closed.*[10]

Since complete constitutions correspond not to complete, but to strongly complete belief sets, one might ask what type of constitutions correspond to locally complete belief sets. The answer is obvious: those constitutions C such that each proposition-negation pair in L has a member q such that $(q, bel) \in C$.

To prove the result, we first show that the notions of consistency, strong completeness and closedness for belief sets can be re-described in a way that corresponds precisely to our definitions of consistency, completeness and closedness for constitutions. The result should be partly familiar to logicians:

Lemma 1 *Given the semantic or syntactic model of L, a belief set $B \subseteq L$ is*

(a) consistent if and only if $B \subseteq B'$ for some complete and consistent belief set $B' \subseteq L$,

(b) strongly complete if and only if $B \supseteq B'$ for some complete and consistent belief set B',

(c) closed if and only if B contains each proposition contained in all complete and consistent extensions $B' \supseteq B$ (equivalently, B is the intersection of these extensions).[11]

Proof. Suppose the lemma's assumptions. Let $B \subseteq L$ be a belief set, and $\mathbf{B}$ the set of complete and consistent belief sets.

(a) We distinguish between the semantic and syntactic model of L. In the semantic case the equivalence holds trivially (if B is consistent, we can pick a $w \in \cap_{p \in B} p$ and define B' as $\{p \in L : w \in p\}$). In the syntactic case the equivalence follows from a basic property in logic, often referred to as 'Lindenbaum's lemma', which states that any consistent set of sentences in a logic is extendable to a complete and still consistent set. This property holds in well-behaved logics of the sort assumed here (see footnote 8).

(b) First let B have a subset $B' \in \mathbf{B}$. To show that B is strongly complete, consider any exhaustive set $Y \subseteq L$. We must prove that $B \cap Y \neq \varnothing$. As $B' \subseteq B$

[10] In the syntactic case we assume the logic is well-behaved as defined in footnote 8.

[11] In case of the syntactic model we assume the logic is well-behaved as defined in footnote 8.

it suffices to show that $Y \cap B' \neq \varnothing$, which holds by the following argument to be spell out for the syntactic and the semantic case:

- *In the syntactic case*, note that the (inconsistent) set $\{\neg p : p \in Y\}$ cannot be a subset of the (consistent) set B'. So there is a $p \in Y$ such that $\neg p \notin B'$, and thus $p \in B'$ as B' is complete. So $Y \cap B' \neq \varnothing$.
- *In the semantic case*, since $\{\Omega \backslash p : p \in Y\}$ has empty intersection (as Y has union Ω) while B' has non-empty intersection (as B' is consistent), the set $\{\Omega \backslash p : p \in Y\}$ cannot be a subset of B'. So there is a $p \in Y$ such that $\Omega \backslash p \notin B'$, and hence $p \in B'$ as B' is complete. So $Y \cap B' \neq \varnothing$.

Conversely, assume that B does *not* include any $B' \in \mathbf{B}$, and let us show that B is not strongly complete. By assumption, for each $B' \in \mathbf{B}$ we may pick a $p_{B'} \in B' \backslash B$. Let $Y := \{p_{B'} : B' \in \mathbf{B}\}$. This set Y is exhaustive, both in the semantic case (here each world $\omega \in \Omega$ belongs to some member of Y, namely to $p_{B'}$ with $B' := \{p \in L : \omega \in p\}$) and also in the syntactic case (here $\{\neg p : p \in Y\}$ is not included in any $B' \in \mathbf{B}$ and so is inconsistent by (a)). Yet $Y \cap B = \varnothing$ by construction of Y. So B is not strongly complete.

(c) We must show that B is closed if and only if $B = \cap_{B' \in \mathbf{B}: B' \supseteq B} B'$. In the syntactic case, this is a familiar fact, valid in in well-behaved logics of the sort considered here (see footnote 8). Now consider the semantic case. Note that $\cap_{B' \in \mathbf{B}: B' \supseteq B} B'$ is closed (in fact, not just in the semantic case). So if $B = \cap_{B' \in \mathbf{B}: B' \supseteq B} B'$ then B is automatically closed. Conversely, if B is closed, then $B = \{p \in L : p \supseteq \cap_{q \in B} q\}$, from which it easily follows that $B = \cap_{B' \in \mathbf{B}: B' \supseteq B} B'$. ■

Proof of Proposition 1. Suppose the proposition's assumptions. Let C be a constitution. We denote the content of a (belief) state m by $\widehat{m}$ and the belief set corresponding to a constitution $C \subseteq M$ by $\widehat{C} = \{\widehat{m} : m \in C\}$.

First,

$$\begin{aligned} C \text{ is consistent} &\Leftrightarrow C \subseteq C' \text{ for some } C' \in T \\ &\Leftrightarrow \widehat{C} \subseteq \widehat{C'} \text{ for some } C' \in T \\ &\Leftrightarrow \widehat{C} \subseteq B \text{ for some consistent and complete } B \subseteq L \\ &\Leftrightarrow \widehat{C} \text{ is consistent, by Lemma 1(a).} \end{aligned}$$

Second,

$$\begin{aligned} C \text{ is complete} &\Leftrightarrow C \supseteq C' \text{ for some } C' \in T \\ &\Leftrightarrow \widehat{C} \supseteq \widehat{C'} \text{ for some } C' \in T \\ &\Leftrightarrow \widehat{C} \supseteq B \text{ for some consistent and complete } B \subseteq L \\ &\Leftrightarrow \widehat{C} \text{ is strongly complete, by Lemma 1(b).} \end{aligned}$$

Third, writing $\widehat{T} := \{\widehat{C} : C \in T\} = \{B \subseteq L : B \text{ is complete and consistent}\}$,

$$\begin{aligned} C \text{ is closed} &\Leftrightarrow C \ni m \text{ for all } m \text{ entailed by } C, \text{ i.e., all } m \in \cap_{C' \in T : C' \supseteq C} C' \\ &\Leftrightarrow \widehat{C} \ni \widehat{m} \text{ for all } m \text{ entailed by } C, \text{ i.e., all } m \in \cap_{C' \in T : C' \supseteq C} C' \\ &\Leftrightarrow \widehat{C} \ni b \text{ for all } b \text{ entailed by } \widehat{C}, \text{ i.e., all } b \in \cap_{B \in \widehat{T} : B \supseteq \widehat{C}} B \\ &\Leftrightarrow \widehat{C} \text{ is closed, by Lemma 1(c).} \blacksquare \end{aligned}$$

C Proof of Corollary 1

Throughout the proof, let T be any theory of rationality. Generalizing Definition 15, we say that a reasoning system S **achieves a requirement** R if $C|S$ satisfies R for all constitutions C. Note that for each of parts (b), (c) and (d) we have to prove two directions of implication, as we read 'unless' as 'if *and only if* it is not the case that'.

Given the contradictory theory $T = \varnothing$, all four parts hold trivially. Part (a) holds because the maximal reasoning system S, which contains all rules, does the job: it achieves closedness by transforming each constitution into M (the only closed constitution), and it vacuously preserves consistency by the absence of consistent constitutions. Parts (b), (c) and (d) hold because consistency, completeness and rationality are all trivially inachievable by the absence of any consistent, complete or rational constitutions (regarding (c), note also the absence of avoidable sets).

Henceforth let $T \neq \varnothing$. We prove the four parts in turn.

Part (a). By Theorem 1(c), achieving closedness is equivalent to achieving all closedness requirements of T. Meanwhile, by Theorem 1 of the companion paper there exists a reasoning schema S which achieves all closedness requirements and preserves consistency. So S achieves closedness while preserving consistency.

Part (b). First, in the (degenerate) case that the maximal constitution $C = M$ is rational, all constitutions are consistent, and so consistency is trivially achieved by any reasoning system. Conversely, assume the maximal constitution $C = M$ is irrational. Let S be any reasoning system; we show that it fails to achieve consistency. As M is irrational, there is an inconsistent constitution C (e.g., $C = M$). As $C|S \supseteq C$, also $C|S$ is inconsistent.

Part (c). First, assume avoidability of each set of falsifiable states (along with the background assumption of compactness, whereby each inconsistent set of states has a finite inconsistent subset). For each unavoidable set U we can pick a non-falsifiable state $m_U \in U$. The reasoning system $S = \{(\varnothing, m_U) : U \text{ is unavoidable}\}$ achieves each completeness requirement of theory T, because for each completeness requirement of T a state from its unavoidable set is formed. So S achieves completeness simpliciter, by Theorem 1. We now show that S preserves consistency.

For a contradiction, consider a consistent constitution C such that $C|S$ is inconsistent. By compactness, $C|S$ has a finite inconsistent subset C'. By definition of S, $C|S = C \cup \{m_U : U \text{ is an unavoidable set}\}$. So we may pick finitely many unavoidable sets $U_1, ..., U_k$ such that $C' \subseteq C \cup \{m_{U_1}, m_{U_2}, ..., m_{U_k}\}$. Since C is consistent, so is $C \cup \{m_{U_1}\}$, as m_{U_1} is non-falsifiable; hence so is $C \cup \{m_{U_1}, m_{U_2}\}$, as m_{U_2} is non-falsifiable. Repeating this argument k times, it follows that $C \cup \{m_{U_1}, m_{U_2}, ..., m_{U_k}\}$ is consistent. Hence its subset C' is consistent.

Conversely, suppose some set of falsifiable states is unavoidable. Let R be the corresponding completeness requirement. It suffices to show that no reasoning system achieves R, because by Theorem 1 achieving completeness is equivalent to achieving all completeness requirements of the theory. By Theorem 3 in the companion paper, no reasoning system achieves any completeness requirement of the theory whose unavoidable set consists of falsifiable states. So no reasoning system S achieves R.

Part (d). First, for (degenerate) theories that deem $C = M$ rational, rationality is trivially achieved by the reasoning system S containing *all* rules, for which $C|S = M$ for all initial constitutions C. Conversely, if $C = M$ is irrational, the inachievability of rationality follows from that of the weaker demand of consistency (see part (b)). ■

D 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

Broersen, J., Dastani, M., Hulstijn, J., van der Torre, L. (2002) Goal Generation in the BOID Architecture, *Cognitive Science Quarterly* 2(3–4)

Broome, J. Rationality through reasoning

Broome, J. (2006) Reasoning with preferences? In: *Preferences and Well-Being*, S. Olsaretti ed., Cambridge University Press, 2006, pp. 183–208

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

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

Dietrich, F. (2007) A generalised model of judgment aggregation, *Social Choice and Welfare* 28(4): 529-565

Dietrich, F., Staras, A., Sugden, R. (2019) A Broomean model of rationality and reasoning, working paper

Fagin, R., Halpern, J. Y., Moses, Y., Vardi, M. Y. (1995) *Reasoning about knowledge*, Cambridge: MIT

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

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

Parfit, D. (20??) On what matters

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

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