A reasoning approach to introspection and unawareness[*]

Olivier Gossner[†] and Elias Tsakas[‡]

March 30, 2010

Abstract

We introduce and study a unified reasoning process which allows to represent the beliefs of both a fully rational agent and an unaware one. This reasoning process endows introspection and unawareness with natural properties. The corresponding model for the rational or boundedly rational agents is easy both to describe and to work with, and the agent's full system of beliefs has natural descriptions using a reduced number of parameters.

KEYWORDS: Belief, information, reasoning, introspection, unawareness.
JEL CLASSIFICATION: D80, D83, D89.

[*] We are indebted to Alpaslan Akay, Robert Aumann, Geir Asheim, Pierpaolo Battigalli, Giacomo Bonanno, Martin Dufwenberg, Amanda Friedenberg, Joe Halpern, Aviad Heifetz, Barton Lipman, Philippe Mongin, Andrés Perea, Dov Samet, Burkhard Schipper, Jeff Steif, Bernard Walliser, Mark Voorneveld and Jörgen Weibull for fruitful discussions and useful comments. We would also like to thank the audiences of the Game Theory Festival (Stony Book), ESEM (Barcelona), SAET (Ischia), Paris Game Theory Seminar, IHPST (Paris), Bocconi University, University of Warwick, HEC Lausanne and Göteborg University. The financial support from the Adlerbertska Forskningsstiftelsen Foundation and the Marie Curie Intra-European Fellowship are gratefully acknowledged.

[†] Paris School of Economics, and London School of Economics and Political Science; ogossner@ens.fr

[‡] Department of Economics, Maastricht University; e.tsakas@maastrichtuniversity.nl

1 Introduction

Modeling the beliefs of an economic agent, whether fully rational or boundedly rational, is, many decades after the seminal work of Simon (1955), still a fundamental question open to debate. Typical beliefs of a boundedly rational agent can exhibit delusion or unawareness, while the beliefs of a rational agent are exempt from such phenomena.

We can distinguish two roots of the agent's beliefs: direct observation, which consists of the agent's experiences in different states of the world, and reasoning, which is the process through which the agent completes his system of beliefs using logical deductions. It is apparent from everyday experience that the same agent can be either aware or unaware of the same facts, depending on the circumstances, and that in most situations, an agent is at the same time aware of some facts and unaware of others. Hence, awareness and unawareness are not the outcome of distinct reasoning abilities, but rather of distinct experiences. Since the same cognitive capacities are compatible with both awareness and unawareness, a unified reasoning process for both the aware and unaware agents has to be the backbone of a model that accounts for these phenomena.

Models previously introduced in the literature (see Modica and Rustichini, 1994, 1999; Halpern, 2001; Halpern and Rêgo, 2008; Heifetz, Meier, and Schipper, 2006, 2008; Li, 2009) rely on distinct reasoning processes depending on the agent's awareness, or unawareness. We find this problematic since, in order to describe an agent's state of mind in these models, one must first specify which propositions the agent is aware or unaware of, then define the agent's reasoning process accordingly, and finally complete the system of the agent's beliefs with the direct observation of phenomena by the agent. This is *as if* awareness preceded reasoning and direct experience, whereas it should arise as the product of their combination.

The aim of this paper is to introduce a model which encompasses both the rational and boundedly rational agent, with the following important features: The reasoning process is the same for both the aware and the unaware agents. Not only is this unified reasoning process not only is compatible with the agent's awareness or unawareness, but it also endows them with natural properties. The model is both simple and tractable, in the sense that it is easy to describe and easy to work with. Finally, it has low complexity, meaning that the agent's beliefs can be fully described using a limited number of parameters.

In our approach, the agent's beliefs can be of two types: universal or contingent. Universal beliefs are state independent, such as belief in the rules of logic. Contingent beliefs arise from direct observation, and are state dependent. We say that the agent has faith in a proposition if he universally believes this proposition. A key assumption in our model is that the agent has faith in his introspective capacity. This means that the agent has faith in the fact that when he believes a proposition, he also believes that he believes it (positive introspection), while if he disbelieves a proposition, he also believes that he disbelieves it (negative introspection). Introspection is a strong assumption, one which, for instance, is incompatible with unawareness. Faith in introspection, however, is distinct, and in fact much weaker than introspection, as it is a property of the agent's beliefs on the structure of his own beliefs, which does not rule out introspection by itself.

Studying the system of propositions the agent has faith in, we show (see Theorem 1) that faith in introspection for all propositions is equivalent to faith in introspection for primitive[1] propositions only. This result allows us to interpret faith in introspection as the assumption that the agent believes in his own familiarity with the relevant primitive phenomena describing his environment.

This first result has interesting implications. Consider the state space in which the agent has faith in introspection on primitives and is non-delusional[2] about these primitive propositions, and is also actually capable of introspection for primitive propositions. We show that this state space coincides with Ω_5, the benchmark state space for the rational agent (see, e.g., Chellas, 1980; Aumann, 1999) in which the agent is capable of introspection and is non-delusional on every proposition, including all epistemic propositions. Thus, introspection extends from the relatively small set of primitive propositions to the whole set of propositions. This shows that introspection does not need to be a mental process *per se*. It is the natural consequence of 1) the agent's familiarity with the environment 2) the agent assuming his own familiarity with the environment and 3) the agent's deductive process. This result thus provides a foundation for introspection, which is a central assumption in formal epistemology.

Now we turn to the more general case in which the agent has faith in introspection, while introspection may or may not hold. Recall the two main notions of unawareness from the literature. According to Modica and Rustichini (1994, 1999), the agent is unaware of a proposition if negative introspection fails for this proposition, i.e., if the agent does not believe in the proposition, and also does not believe that he disbelieves it. According to the stronger unawareness concept introduced by Dekel, Lipman, and Rustichini (1988), the

[1] Primitive propositions are those that do not involve the agent's belief. Their set is closed under negation, conjunction and disjunction. For instance, the proposition "Ann has blue eyes or it is raining in New York" is a primitive proposition.

[2] Meaning that whenever the agent believes a primitive proposition is true, this proposition is indeed true.

agent is unaware of a proposition if he does not believe in any sequence of "I believe in" and "I disbelieve in" followed by the proposition. This stronger concept formalizes a complete lack of recognition of the proposition, as the agent who is unaware of a proposition cannot, for instance, believe that he disbelieves that he believes in this proposition. We show that, whenever the agent has faith in introspection, the two notions are necessarily equivalent, namely, a failure of negative introspection on some proposition is necessarily accompanied by, and can only be explained by, a total failure of recognition of this proposition.

We study the state space Ω_u, superset of Ω_5, in which the agent has faith in introspection, while positive introspection (a much more widely-accepted assumption than negative introspection) holds. We show (see Proposition 2) that for states in Ω_u only two cases can arise. Either introspection holds for every proposition, or there exists a primitive proposition that the agent is unaware of. Thus, our state space encompasses both the standard rational agent, for whom introspection holds on every proposition, and the boundedly rational one, who exhibits unawareness on some primitive propositions. The model is also tight as no other possibilities may arise.

It is important, for practical and tractability reasons, to know how complex the description of the agent's beliefs is in the state space Ω_5. Halpern (1995) shows that elements of Ω_5 have a simple description, as every state in Ω_5 is entirely described by the values of primitive propositions and the agent's beliefs on primitive propositions at that state. Meanwhile, since Ω_u is larger than Ω_5, and allowing for unawareness, one expects the description of elements of Ω_u to be more complex than in Ω_5. This is indeed the case, but elements of Ω_u also have simple description. In Ω_u, a state is described by the values of primitive propositions and the beliefs of the agent on any proposition that contains at most once the belief operator. This implies in particular that Ω_u has bounded cardinality if the set of propositions is constructed from a finite set of primitive propositions. An alternative, and rather natural, description of elements of Ω_u relies on the set of propositions the agent is aware of. According to this description, a state in Ω_u is given by the values of the primitive propositions, the agent's beliefs on these primitives, and by the set of propositions the agent is aware of at this state.

Our primary model is a syntactic one, in the tradition of Chellas (1980) and Aumann (1999). Syntactic models explicitly represent the agent's reasoning and belief construction processes. Semantic models, previously studied by Hintikka (1962), Aumann (1976), Geanakoplos (1989) and Dekel, Lipman, and Rustichini (1988) among others, represent the outcome of this process in the form of a possibility correspondence that assigns to each state of the world the set of states that the agent considers as possible. We construct a semantic model which is equivalent to our syntactic model. An alternative way to look at the syntactic model is its semantic form, which retains the same properties in particular as regards unawareness and the role of primitive propositions.

The paper is organized as follows. Section 2 recalls the model of syntactic beliefs. Section 3 uses a motivating example to present the main questions. We introduce our approach of universal beliefs versus contingent beliefs in Section 4, and study the properties of unawareness in our model in Section 5. We present the model of semantic beliefs in Section 6, and conclude with a discussion in Section 7.

2 Propositions

We recall the syntactic model of belief proposed by Chellas (1980), Fagin, Halpern, Moses, and Vardi (1995) and Aumann (1999). We start with a set of primitive propositions, Φ_0, which describe facts about the world that do not involve the agent's belief. Examples of primitive propositions are “it is raining” and “It is sunny and Ann has blue eyes”.

The symbols $\neg$, $\vee$ and $\wedge$ express negation, disjunction and conjunction, i.e., $\neg\phi$ stands for "not ϕ", $(\phi_1 \vee \phi_2)$ stands for "ϕ_1 or ϕ_2" and $(\phi_1 \wedge \phi_2)$ stands for "ϕ_1 and ϕ_2". The set of primitive propositions Φ_0 is taken to be closed under these operations: whenever ϕ_1, ϕ_2 are primitive propositions, so are $\phi_1 \vee \phi_2$, $\phi_1 \wedge \phi_2$ and $\neg\phi_1$.

The symbol β expresses belief: $\beta\phi$ stands for "the agent believes ϕ".

The set of propositions Φ is the closure of Φ_0 under β, $\vee$, $\wedge$ and $\neg$. It is the smallest set of propositions that can be constructed from Φ_0 using these operations. For instance, whenever ϕ_1, ϕ_2, ϕ_3 are propositions, so is $\beta(\phi_1 \vee \phi_2) \wedge \beta\neg\phi_3$. Non-primitive propositions, such as $\phi_1 \vee \beta\phi_2$, are called epistemic.

The proposition "ϕ_1 implies ϕ_2" is denoted by $(\phi_1 \rightarrow \phi_2)$ and is an abbreviation for $(\neg\phi_1 \vee \phi_2)$; "ϕ_1 if and only if ϕ_2" is denoted by $(\phi_1 \leftrightarrow \phi_2)$ and is defined as $(\phi_1 \rightarrow \phi_2) \wedge (\phi_2 \rightarrow \phi_1)$.

For a proposition ϕ, the expression "the agent satisfies positive introspection for ϕ" denotes the proposition $(\beta\phi \rightarrow \beta\beta\phi)$, and "the agent satisfies negative introspection for ϕ" denotes the proposition $(\neg\beta\phi \rightarrow \beta\neg\beta\phi)$. As usual, we take positive introspection to mean that when the agent believes a proposition, he believes that he believes it, and negative introspection to mean that when the agent disbelieves a proposition, he believes that he disbelieves it. Finally, the expression "the agent satisfies the truth axiom for ϕ" stands for $(\beta\phi \to \phi)$.

3 Motivating example

We recall the example from Conan Doyle's short story "Silver Blaze". A dialog takes place between the famous detective Sherlock Holmes and the Scotland Yard detective Gregory:

> Gregory: *Is there any other point to which you would wish to draw my attention?*
>
> Holmes: *To the curious incident of the dog in the night-time.*
>
> Gregory: *The dog did nothing in the night-time.*
>
> Holmes: *That was the curious incident.*

From the fact that the dog did not bark in the night-time, Holmes infers that no stranger intruded. On the other hand, Gregory, using the same premises as Holmes, fails to reach this conclusion. This example has received considerable attention in the epistemology literature (see, for instance Geanakoplos, 1989), Dekel, Lipman, and Rustichini (1988), and is often used to illustrate unawareness. Our first exercise, which, surprisingly enough, has not been dealt with in the literature, is to detail the mental process through which Holmes reaches his conclusion. We then discuss the several steps where Gregory may fail to follow Holmes' mental process and the modeling questions that arise from this example.

3.1 Holmes: The rational benchmark

We break down Holmes' reasoning process into two parts. The first part leads to the conclusion that the dog did not bark, while the second leads to the conclusion that no stranger intruded.

For the first part, let B denote the proposition "the dog barked", and β_H Holmes' belief operator. The first premise is that Holmes did not know of the dog barking: $\neg\beta_H B$. The second is that negative introspection applies $\beta_H\neg\beta_H B$. Next, one needs to assume that Holmes believes that, had the dog barked, he would have known it: $\beta_H(B \to \beta_H B)$. Using propositional calculus (in particular, the contraposition), Holmes believes that, if he does not believe that the dog barked, then the dog indeed did not bark: $\beta_H(\neg\beta_H B \to \neg B)$. From this, Holmes infers that the dog did not bark: $\beta_H \neg B$.

As for the second part, let I denote the proposition "a stranger entered the stables". Holmes assumes that, had a stranger come into the stables, the dog would have barked: $\beta_H(I \to B)$. Using propositional calculus, Holmes believes that if there was no barking, there was no stranger: $\beta_H(\neg B \to \neg I)$. Finally, Holmes combines his belief in this implication with his belief in the absence of barking to reach the conclusion that no stranger entered the stables: $\beta_H \neg I$.

3.2 Gregory: The boundedly rational agent

Gregory's reasoning may fail to follow that of Holmes in several points. Let β_G denote Gregory's belief operator. An immediate possibility is that Gregory fails to use the rule of inference adequately: We could have in this case $\beta_G(\neg B \to \neg I)$, $\beta_G(\neg B)$, but at the same time $\neg\beta_G(\neg I)$. A second possibility is that Gregory is not capable of following the rules of propositional calculus, as for instance is the case if $\beta_G(I \to B)$ and $\neg\beta_G(\neg B \to \neg I)$. A third possibility is that Gregory does not have the same understanding of the world as Holmes, and does not believe that the presence of a stranger would have made the dog bark: $\neg\beta_G(I \to B)$.

Finally, the most interesting possibility, and, most likely, the one that is implicit in the story, is that Gregory is not aware of the possibility of the dog barking: $\neg\beta_G B$, while $\neg\beta_G\neg\beta_G B$.

3.3 Modeling questions

We aim at developing a model which can encompass both the rational agent (Holmes) and the boundedly rational one (Gregory), while retaining tractability. Let us admit that in the example, the fundamental distinction between Holmes and Gregory is their ability to apply negative introspection to the proposition "the dog barked". This means that we have in mind a model in which negative introspection does not necessarily hold for all propositions, or for all agents. This opens several modeling questions.

First, should introspection be interpreted as a mental process *per se*? If the answer is yes, it may be difficult to justify that the same agent (e.g., Gregory) is capable of applying this process to some propositions, and incapable of applying it to others. A good model of beliefs should, as much as possible, explain why introspection can hold for some propositions and not for others, or at least provide a clear interpretation for this fact.

Second, in the case of Holmes, it appears natural that he should be aware of a proposition such as "I believe that the dog did not bark". However, awareness of such a proposition is not used in the reasoning process we described for Holmes, and can *a priori* fail in a model where introspection does not necessarily hold for every proposition. A desirable property of a belief model is that awareness (or introspection capacities) between propositions should be related. This type of relation should be the consequence of natural assumptions of the model, and should not be part of the assumptions themselves.

As we see, having a model in which introspection can hold for some propositions without holding for others – and in which the structure of the set of propositions for which introspection holds has a clear and intuitive structure – is not a straightforward task. We present our approach in the next section.

4 Our approach: universal beliefs and contingent beliefs

We are ultimately interested in devising a state space in which each state describes the truth value of every proposition, including the agent's beliefs. In our approach, the agent's beliefs can be of two types: universal or contingent. Universal beliefs consist of beliefs in propositions that are not dependent on any particular state. For instance, belief in a proposition such as "a cat is mortal, or it is not" is universal, as belief in such a proposition is derived from rules of logic, and entails no beliefs about the nature of a cat, or the meaning of being mortal. On the other hand, belief in a proposition such as "it is raining in New York", is contingent by nature. Belief in such a proposition arises from particular observation, at particular states of nature, about the weather in New York, and can differ from one state of nature to another.

We describe the agent's universal beliefs in Section 4.1 and the state space in Section 4.2.

4.1 Faith: universal beliefs

We describe the set F of propositions that the agent has faith in. When considering a state space model, these propositions will be assumed to be believed by the agent at all states (see Section 4.2).

The set F is constructed from a basic set of propositions using inference rules. The basic set of propositions, otherwise called assumptions, is the set A consisting of the following propositions:

(A_0) All tautologies of propositional calculus

(A_I) $\beta(\phi_1 \rightarrow \phi_2) \rightarrow (\beta\phi_1 \rightarrow \beta\phi_2)$, for every $\phi_1, \phi_2 \in \Phi$ (Axiom of distribution)

(A_1) $\beta(\phi_1 \wedge \phi_2) \leftrightarrow (\beta\phi_1 \wedge \beta\phi_2)$, for every $\phi_1, \phi_2 \in \Phi$ (Conjunction of belief)

(A_2) $(\beta\phi_1 \vee \beta\phi_2) \rightarrow \beta(\phi_1 \vee \phi_2)$, for every $\phi_1, \phi_2 \in \Phi$ (Disjunction of belief)

(A_3) $\beta\phi \rightarrow \phi$, for every $\phi \in \Phi_0$ (Truth axiom on primitive propositions)

(A_4) $\beta\phi \rightarrow \beta\beta\phi$, for every $\phi \in \Phi_0$ (Positive introspection on primitive propositions)

(A_5) $\neg\beta\phi \rightarrow \beta\neg\beta\phi$, for every $\phi \in \Phi_0$ (Negative introspection on primitive propositions)

(A_6) $\beta\phi \rightarrow \neg\beta\neg\phi$, for every $\phi \in \Phi$ (Consistency of belief)

The first axiom (A_0) refers to obvious propositions, such as $\big((\phi_1 \rightarrow \phi_2) \leftrightarrow (\neg\phi_2 \rightarrow \neg\phi_1)\big)$, which are logically true. The axiom of distribution says that if it is believed that ϕ_1 implies ϕ_2 and it is also believed that ϕ_1 is true, then it is necessarily believed that the logical consequence ϕ_2 is also true. Conjunction says that "ϕ_1 and ϕ_2" is believed if and only if both propositions are believed. Disjunction states that "ϕ_1 or ϕ_2" is believed if at least one of the propositions is believed. Consistency says that the agent cannot believe a proposition and its negation simultaneously.

The truth axiom says that the agent is confident that his own beliefs are right, i.e. that if he believes a proposition then this proposition is true. Positive and negative introspection are also assumed to be part of the agent's faith for primitive propositions. This means that the agent assumes sufficient familiarity with his own environment: he is confident that he is capable of forming correct beliefs about his own beliefs about primitive propositions, which describe the relevant parameters of his environment.

The inference rules by which propositions in F are constructed from other propositions in F are the following:

(R_I) If $\phi_1 \in F$ and $(\phi_1 \rightarrow \phi_2) \in F$, then $\phi_2 \in F$ (Modus Ponens)

(R_F) If $\phi \in F$, then $\beta\phi \in F$ (rule of necessitation)

Modus Ponens requires that the agent is capable of making inferences on the set of propositions he has faith in. Together with A_0, it implies that the agent has faith in $(\phi_1 \wedge \phi_2)$ if and only if he has faith in both propositions.

The rule of necessitation states that the agent has faith in believing everything he has faith in, i.e., if the agent has faith in some proposition, then he has also faith in the fact that he believes this proposition.

Definition 1. *The set F of propositions the agent has faith in is the smallest set containing all propositions in A which is closed under R_I and R_F.*

Formally, A together with the rules R_I and R_F form a system of modal logic, as in Chellas (1980) or Fagin, Halpern, Moses, and Vardi (1995). Elements of A are called axioms, while R_I and R_F are inference rules, and the elements of F are called the theorems of the system of modal logic. The elements of A and the inference rules that we use are standard in modal logic, except for $A_3 - A_5$, which are weakenings of the standard axioms. The benchmark Modal Logic system, which is called S5 and used to represent a logically omniscient agent, is defined by our set of axioms and inference rules, where $A_3 - A_5$ are strengthened in that they are taken as axioms for every proposition, not just for primitive ones.

The following theorem shows that faith in introspection and the truth axiom for primitive propositions extends to the whole set of propositions.

Theorem 1. *The agent has faith in the truth axiom and introspection for every proposition: For every $\phi \in \Phi$*

1. $(\beta\phi \rightarrow \phi) \in F$

2. $(\beta\phi \rightarrow \beta\beta\phi) \in F$

3. $(\neg\beta\phi \rightarrow \beta\neg\beta\phi) \in F$

The proof of Theorem 1 can be found in Appendix A.1.

A consequence of the previous result is that the system of modal logic defining F is formally equivalent to the system S5 of a logically omniscient agent. Theorem 1 shows that, in S5, it is enough to assume the truth axiom and introspection on primitive propositions rather than on all propositions.

It is important to keep in mind that we do not take for granted that introspection holds for every proposition, or even for every primitive proposition. In that case, we would interpret elements of F as properties that necessarily hold at every state of the world. In other words, we would assume that the agent is logically omniscient. Rather, we interpret F as a set of propositions the agent has faith in, hence we have in mind an agent who has faith in his own logical omniscience. Whether or not he is right to have faith in all propositions in F, i.e. whether or not all propositions in F hold at a given state of the world, is a question studied in Section 4.2. This distinction will prove to be of considerable importance, in particular with respect to the question of unawareness.

4.2 States of the world: Contingent beliefs

The agent's beliefs at any given state are of two types: faith, which consists of belief in propositions of F, and contingent beliefs, which consist of beliefs about all other propositions. Faith is universal in the sense that belief in F holds at every state. On the other hand, contingent belief is state-dependent. Consider for instance the primitive proposition ϕ, which stands for "it is raining in New York". Obviously no element of F can inform the agent about the truth value of ϕ, and therefore the beliefs that the agent holds about ϕ depend on what he observes at every state, i.e., on whether he has credible information that it is raining. Hence, the truth value of $\beta\phi$ may differ across states.

Following Aumann (1999), a state ω assigns a truth value to every proposition in Φ. A state thus provides a complete description of the facts (primitive propositions) and the agent's beliefs (epistemic propositions). It is a mapping from Φ to $\{0, 1\}$, with the interpretation that ϕ is true at ω when $\omega(\phi) = 1$ and false otherwise. We identify ω with the set of propositions that are true at ω, and we write $\phi \in \omega$ when ϕ is true at ω. Thus, we write $\omega = \{\phi \in \Phi : \omega(\phi) = 1\}$. A state space is a collection of such states ω. We restrict attention to states that satisfy the basic rules of logic, so we let Ω_0 be the set of such mappings ω, such that for every $\phi, \phi' \in \Phi$:

- $\phi \in \omega$, if and only if $\neg\phi \notin \omega$
- $\phi \wedge \phi' \in \omega$, if and only if $\phi \in \omega$ and $\phi' \in \omega$
- $\phi \vee \phi' \in \omega$, if and only if $\phi \in \omega$ or $\phi' \in \omega$

We have in mind an agent who is capable of reasoning and believes in all propositions of F.

Definition 2. *Let Ω_r be the subset of Ω_0 containing all states ω such that*

(A_I) $\big(\beta(\phi_1 \to \phi_2) \to (\beta\phi_1 \to \beta\phi_2)\big) \in \omega$, *for every* $\phi_1, \phi_2 \in \Phi$

(A_1) $\big(\beta(\phi_1 \wedge \phi_2) \leftrightarrow (\beta\phi_1 \wedge \beta\phi_2)\big) \in \omega$, *for every* $\phi_1, \phi_2 \in \Phi$

(A_2) $\big((\beta\phi_1 \vee \beta\phi_2) \rightarrow \beta(\phi_1 \vee \phi_2)\big) \in \omega$, *for every* $\phi_1, \phi_2 \in \Phi$

(B_F) $\beta\phi \in \omega$, *for all* $\phi \in F$

The property B_F states that the agent believes in all propositions in F, i.e., the agent's contingent beliefs cannot contradict the agent's universal beliefs. In other words, the belief in some element of F is constant across states, i.e. the agent has faith in the fact that his beliefs are correct[3], implying that he has faith in himself satisfying A_3, and therefore he believes in the proposition "it rains in New York whenever I believe so": $\beta(\beta\phi \rightarrow \phi) \in \omega$, for all $\omega \in \Omega_r$.

The restriction B_F is quite reasonable, as the agent himself has proven – or assumed – these propositions, and therefore as long as he is confident that what he has assumed is true, he must also be confident that the conclusions he has reached are also true. However, the fact that the agent believes all propositions in F does not necessarily mean that he is always right to do so, i.e. he may wrongly believe some of the propositions in which he has faith. Recall the example from the previous paragraph: the agent believes that every time he receives credible information about rain in New York, then this information is necessarily true. However, this need not be the case, as it would rule out the possibility that even though he believes his source, the information provided to him is wrong.

The wrong beliefs that the agent may have at some state are not arbitrary. The agent does not wrongly believe that he is capable of reasoning, i.e., we restrict our focus to states satisfying the main principles of belief, as expressed by A_1, A_2 and A_I.

A model in which the agent's faith is potentially delusional is very interesting from the bounded rationality point of view. It provides a framework in which all agents – whether fully or boundedly rational – have faith in the same system of propositions F, and what distinguishes them is whether their assumptions on the world are satisfied or not. Bounded rationality is therefore contingent, and the reasoning processes of both the rational and boundedly rational agents are the same.

Situations of delusional faith are studied extensively in Section 5. We conclude this section by showing that delusional faith can only arise when the agent's assumptions $A_3 - A_5$ on primitives are not satisfied.

[3] Feinberg (2004) imposes the same assumption, i.e. the agent (possibly wrongly) believes that his beliefs satisfy the truth axiom.

Definition 3. *Faith in ϕ is well-founded at ω if $\phi \in F$ implies $\phi \in \omega$. Faith is well-founded at ω if it is well-founded for all propositions at ω.*

Theorem 2. *Faith is well-founded at $\omega \in \Omega_r$ if and only if $A_3 - A_5$ are satisfied for all primitives at ω.*

The "only if" part of the theorem is obvious, as propositions in $A_3 - A_5$ all belong to F, by definition. The "if" part shows that, if the agent's assumptions on his introspection capacities on primitives are correct, so are the logical conclusions he derives from them.

In particular, Theorem 2 shows that, whenever introspection holds on primitive propositions, it holds for every proposition. It provides a foundation for the introspection axioms, which are central in the literature (Samet, 1990), and allows to break down these axioms into introspection for primitive propositions – which can be understood as the product of the agent's familiarity with these propositions – and the agent's deductive process, leading to introspection for all other propositions.

The subset Ω_5 of states in Ω_r at which every proposition in F holds, is the canonical state space for the modal logic system S5.

5 Unawareness

The aim of this section is to study situations where the agent has faith in his own reasoning ability, i.e. he correctly believes at all states that he has faith in $A_0 - A_5$ and A_I, but he may wrongly believe some of the propositions in F which are related to the truth axiom and introspection. Furthermore, we examine how these wrong beliefs are connected with the notions of unawareness and delusion.

5.1 A state space with unawareness

Unawareness about a phenomenon corresponds to a strong form of ignorance about this phenomenon, in the sense that the agent fails to recognize his own ignorance. Following Modica and Rustichini (1994, 1999), we define unawareness of ϕ as the conjunction of the ignorance of ϕ together with the ignorance of this ignorance: we let $u\phi$ stand for $\neg\beta\phi \wedge \neg\beta\neg\beta\phi$.

The definition for unawareness is relatively weak, in that $u\phi$ is compatible for instance with the agent believing that he does not believe that he does not believe ϕ. Following Dekel, Lipman, and Rustichini (1988), a stronger definition of unawareness requires the agent to disbelieve any proposition made by a sequence of "the agent believes" or "the agent does not believe" and ending in ϕ: $\neg\beta\phi'$ for all $\phi' \in B(\phi)$, where $B(\phi)$ is the closure of $\{\beta\phi\}$ with respect to the operations $\neg$ and β. Our next result shows that, in every state where the agent has faith in every proposition in F, both definitions are equivalent:

Proposition 1. *Let $\omega \in \Omega_r$ be such that the agent believes every proposition in which he has faith: $\beta\phi \in \omega$ for every $\phi \in F$. For every proposition $\phi \in \Phi$, the agent is unaware of ϕ at ω if and only if $\neg\beta\phi' \in \omega$ for all $\phi' \in B(\phi)$.*

Proposition 1 shows that in Ω_r, failure of negative introspection on a proposition ϕ is necessarily accompanied by unawareness in the strong sense that the agent completely ignores ϕ, i.e. the agent cannot be aware of a primitive ϕ, while being unaware of his belief of ϕ: we cannot have $\neg u\phi$ and $u\beta\phi$ at the same ω.

One advantage of using the weaker definition of unawareness is that $u\phi$ is a well-defined proposition in Φ. As shown by Proposition 1, it is equivalent to the stronger unawareness notion that is defined through the conjunction of an infinite family of propositions.

Since unawareness is a violation of negative introspection, we study states where negative introspection is relaxed. We consider states in which the agent believes in F, and in particular believes in negative introspection, but negative introspection may or may not hold. We also relax the truth axiom, since keeping the truth axiom would imply automatically that everything believed by the agent – including negative introspection – holds. On the other hand, we assume positive introspection, which is not considered as a problematic axiom (Samet, 1990; Lipman, 1995).

Definition 4. *Let Ω_u be the set of states in Ω_r in which positive introspection holds for every proposition.*

In Ω_u, the agent is capable of reasoning, believes in every proposition in F and is capable of positive introspection. Unlike Ω_5, negative introspection and the truth axiom do not necessarily hold in Ω_u.

Now we relate unawareness to unawareness of primitive propositions:

Proposition 2. *If the agent is unaware of some proposition at $\omega \in \Omega_u$ then he is unaware of some primitive proposition.*

Proposition 2 shows that in Ω_u, the only possible source of unawareness is unawareness of a primitive proposition.

The next result characterizes F as the set of propositions that are universally believed in Ω_u, i.e., the propositions ϕ such that $\beta\phi \in \omega$ for every $\omega \in \Omega_u$.

Proposition 3. *F is the set of propositions that are universally believed in Ω_u.*

5.2 Complexity of the state space

As in e.g. Aumann (1999), we define the epistemic depth of a proposition ϕ as the number of nested belief operators found in this proposition. It is 0 for primitive propositions, the depth of $\neg\phi$ is the same as the depth of ϕ, the depth of $\phi_1 \vee \phi_2$ and $\phi_1 \wedge \phi_2$ is the maximum of the depths of ϕ_1 and ϕ_2, and the depth of $\beta\phi$ is equal to the depth of ϕ plus one. Let Φ_n denote the set of propositions of epistemic depth at most n. Formally, we define Φ_n as the closure of the set $\{\phi, \beta\phi \mid \phi \in \Phi_{n-1}\}$ with respect to $\neg$, $\vee$ and $\wedge$.

As shown by Halpern (1995), states in Ω_5 have an easy description. That is, two distinct states in Ω_5 must differ in the truth value of the primitive propositions, or in the primitive beliefs. This is particularly interesting as what the agent believes about any proposition depends only on what he believes about the primitives, and therefore a state is determined by the primitive propositions and the primitive beliefs.

Proposition 4 (Halpern (1995))**.** *Let $\omega, \omega' \in \Omega_5$. If $\omega(\phi) = \omega'(\phi)$ and $\omega(\beta\phi) = \omega'(\beta\phi)$ for every $\phi \in \Phi_0$, then $\omega = \omega'$.*

Remark 1. Note that the values of the primitive propositions do not place any restrictions on the relationship between epistemic propositions, i.e. beliefs are determined inductively by beliefs (of lower or equal depth), and not by the truth values of the primitive. ◁

The following example illustrates the relationship between primitive beliefs and states in Ω_5.

Example 1. Suppose all propositions are derived from one primitive ϕ, i.e., $\Phi_0 = \{\phi, \neg\phi\}$. From Proposition 4, it follows that the rational agent's state space is $\Omega_5 = \{\omega_1, ..., \omega_4\}$ where:

$$\begin{aligned} \omega_1 &= \{\phi, \beta\phi, \neg\beta\neg\phi, ...\}, \\ \omega_2 &= \{\neg\phi, \neg\beta\phi, \beta\neg\phi, ...\}, \\ \omega_3 &= \{\phi, \neg\beta\phi, \neg\beta\neg\phi, ...\}, \\ \omega_4 &= \{\neg\phi, \neg\beta\phi, \neg\beta\neg\phi, ...\}. \end{aligned}$$

Why is the value of every proposition ψ fixed at every $\omega \in \{\omega_1, ..., \omega_4\}$? We illustrate the mechanics underlying Proposition 4 for several such propositions.

First, note that the value of every such ψ is determined by the value of primitives and the agent's beliefs. Remark also that the belief in any ψ of the form $\psi = \psi_1 \wedge \psi_2$, is equivalent to belief in both ψ_1 and ψ_2.

How about belief in $\phi \vee \beta\phi$? Lemma 4 (in the appendix) together with $(\beta\phi \leftrightarrow \beta\beta\phi) \in \omega$ shows that for every $\omega \in \Omega_5$, $\big(\beta(\phi \vee \beta\phi) \leftrightarrow \beta\phi\big) \in \omega$. Hence $\beta(\phi \vee \beta\phi)$ holds in ω_1 only.

Lemma 4 also shows that $\big(\beta(\phi \vee \neg\beta\phi) \leftrightarrow (\beta\phi \vee \neg\beta\phi)\big) \in \omega$ for $\omega \in \Omega_5$, hence $\beta(\phi \vee \neg\beta\phi) \in \omega$ for all $\omega \in \Omega_5$.

Similarly $\big(\beta(\neg\phi \vee \beta\phi) \leftrightarrow (\beta\neg\phi \vee \beta\phi)\big) \in \omega$ for $\omega \in \Omega_5$. Thus $\beta(\neg\phi \vee \beta\phi)$ holds in ω_1 and ω_2, but not in ω_3 or ω_4.

More generally, it can be shown by induction that in Ω_5, beliefs on propositions in Φ_n are determined by the belief on propositions in Φ_{n-1}. $\triangleleft$

The next theorem shows that in Ω_u, beliefs are determined by the truth value of the primitives and the beliefs about every proposition of epistemic depth at most one.

Theorem 3. *Let $\omega, \omega' \in \Omega_u$. If $\omega(\phi) = \omega'(\phi)$ and $\omega(\beta\phi) = \omega'(\beta\phi)$ for every $\phi \in \Phi_1$, then $\omega = \omega'$.*

Remark 2. As in Ω_5, beliefs are determined inductively by beliefs of lower or equal depth, and not by the primitives. $\triangleleft$

Theorem 3 shows that, although allowing for a very rich environment including possibilities of unawareness, the state space Ω_u still remains tractable. The driving force is that structure is provided through the faith system: through a process of deductive reasoning, the agent is able to derive all higher order beliefs from beliefs about propositions of depth at most one.

In particular, Theorem 3 implies that Ω_u is finite if all propositions are constructed from an initial finite set of primitive propositions.

Example 2. Suppose as in Example 1 that all propositions are derived from a primitive ϕ, i.e., $\Phi_0 = \{\phi, \neg\phi\}$. In this case, Ω_u can be described as $\Omega_u = \{\omega_1^+, ..., \omega_9^+\} \cup \{\omega_1^-, ..., \omega_9^-\}$, where in ϕ holds in states ω_i^+, $\neg\phi$ holds in states ω_i^-, and all the agent's beliefs are the same in states ω_i^+ and ω_i^-. An agent's "state of mind", which is the same in ω_i^+ and ω_i^-, can be written as $\tilde{\omega}_i = \omega_i^+ \cap \omega_i^-$. It is straightforward that once $\tilde{\omega}_i$ is known, both ω_i^+ and ω_i^- are also known, so that the description of the state space can be completed by the description of the agent's "states of mind" $\tilde{\omega}_1, \ldots, \tilde{\omega}_9$.

- States of mind in which the agent is aware of ϕ and $\neg\phi$

$$\begin{aligned}
\tilde{\omega}_1 &= \{\beta\phi, \neg\beta\neg\phi, \beta\beta\phi, \beta\neg\beta\neg\phi, \beta(\phi \vee \beta\neg\phi), \beta(\neg\phi \vee \beta\phi), \ldots\} \\
\tilde{\omega}_2 &= \{\neg\beta\phi, \beta\neg\phi, \beta\neg\beta\phi, \beta\beta\neg\phi, \beta(\phi \vee \beta\neg\phi), \beta(\neg\phi \vee \beta\phi), \ldots\} \\
\tilde{\omega}_3 &= \{\neg\beta\phi, \neg\beta\neg\phi, \beta\neg\beta\phi, \beta\neg\beta\neg\phi, \neg\beta(\phi \vee \beta\neg\phi), \neg\beta(\neg\phi \vee \beta\phi), \ldots\}
\end{aligned}$$

States $\{\omega_1^+, \omega_2^-, \omega_3^+, \omega_3^-\}$ coincide with $\omega_1, \ldots, \omega_4 \in \Omega_5$. In states ω_1^- and ω_2^+, the agent exhibits delusion, since either ϕ or $\neg\phi$ is believed but does not hold.

- States of mind in which the agent is unaware of $\neg\phi$ and is aware of ϕ:

$$\tilde{\omega}_4 = \{\neg\beta\phi, \neg\beta\neg\phi, \beta\neg\beta\phi, \neg\beta\neg\beta\neg\phi, \neg\beta(\phi \vee \beta\neg\phi), \neg\beta(\neg\phi \vee \beta\phi), \ldots\}$$

The fact that $\neg\beta(\neg\phi \vee \beta\phi) \in \tilde{\omega}_4$ is straightforward: Let us suppose otherwise. Then, $\beta(\neg\beta\phi \rightarrow \neg\phi) \in \tilde{\omega}_4$ contradicts $\neg\beta\phi \in \tilde{\omega}_4$, because of A_I. The fact that $\neg\beta(\phi \vee \beta\neg\phi) \in \tilde{\omega}_4$ also follows by contradiction: Otherwise, it follows from A_4 and A_I that $\beta(\beta\neg\beta\neg\phi \rightarrow \beta\phi) \in \tilde{\omega}_4$. Then, it follows from A_0 that $\beta(\neg\beta\phi \rightarrow \neg\beta\neg\beta\neg\phi) \in \tilde{\omega}_4$, and again from A_I it follows that $(\beta\neg\beta\phi \rightarrow \beta\neg\beta\neg\beta\neg\phi) \in \tilde{\omega}_4$, which contradicts the unawareness of $\neg\phi$.

- States of mind in which the agent is unaware of ϕ and is aware of $\neg\phi$:

$$\tilde{\omega}_5 = \{\neg\beta\phi, \neg\beta\neg\phi, \neg\beta\neg\beta\phi, \beta\neg\beta\neg\phi, \neg\beta(\phi \vee \beta\neg\phi), \neg\beta(\neg\phi \vee \beta\phi), \ldots\}$$

The arguments regarding the beliefs about $\phi \vee \beta\neg\phi$ and $\neg\phi \vee \beta\phi$ at $\tilde{\omega}_5$ are the same as in $\tilde{\omega}_4$.

- States of mind in which the agent is unaware of both ϕ and $\neg\phi$:

$$\begin{aligned}
\tilde{\omega}_6 &= \{\neg\beta\phi, \neg\beta\neg\beta\phi, \neg\beta\neg\phi, \neg\beta\neg\beta\neg\phi, \beta(\phi \vee \beta\neg\phi), \beta(\neg\phi \vee \beta\phi), \ldots\} \\
\tilde{\omega}_7 &= \{\neg\beta\phi, \neg\beta\neg\beta\phi, \neg\beta\neg\phi, \neg\beta\neg\beta\neg\phi, \neg\beta(\phi \vee \beta\neg\phi), \beta(\neg\phi \vee \beta\phi), \ldots\} \\
\tilde{\omega}_8 &= \{\neg\beta\phi, \neg\beta\neg\beta\phi, \neg\beta\neg\phi, \neg\beta\neg\beta\neg\phi, \beta(\phi \vee \beta\neg\phi), \neg\beta(\neg\phi \vee \beta\phi), \ldots\} \\
\tilde{\omega}_9 &= \{\neg\beta\phi, \neg\beta\neg\beta\phi, \neg\beta\neg\phi, \neg\beta\neg\beta\neg\phi, \neg\beta(\phi \vee \beta\neg\phi), \neg\beta(\neg\phi \vee \beta\phi), \ldots\}
\end{aligned}$$

Simultaneous unawareness of both ϕ and $\neg\phi$ allows any beliefs for $\phi \vee \beta\neg\phi$ and $\neg\phi \vee \beta\phi$, e.g. $\tilde{\omega}_6$ and $\tilde{\omega}_7$ differ in the beliefs of $\phi \vee \beta\neg\phi$, implying that primitive beliefs and awareness about the primitives do not suffice to characterize a state. Instead, some more information is needed (see Theorem 4 below).

Two things become clear from this example, which provide an illustration of our previous results. First, whenever the agent's beliefs do not coincide with beliefs in Ω_5, the agent is unaware (see Proposition 2) or delusional about at least one primitive, and second, unlike in Ω_5, primitive beliefs alone do not suffice for characterizing the agent's state of mind in Ω_u (see Theorem 3).

Finally, note that imposing $(u\phi \leftrightarrow u\neg\phi) \in \omega$, like Modica and Rustichini (1994), would eliminate the states of mind $\tilde{\omega}_4$ and $\tilde{\omega}_5$, but not the other states. In particular, our model is compatible with such a restriction, even if, for the sake of generality, we do not impose it. $\triangleleft$

Now we present an alternative description of states in Ω_u in which a state is described through primitives and beliefs on primitives as well as the agent's awareness of propositions of depth at most one.

Theorem 4. *Let $\omega, \omega' \in \Omega_u$. If $\omega(\phi) = \omega'(\phi)$ and $\omega(\beta\phi) = \omega'(\beta\phi)$ for every $\phi \in \Phi_0$, and $\omega(u\phi) = \omega'(u\phi)$ for all $\phi \in \Phi_1$, then $\omega = \omega'$.*

This last theorem has a natural appeal: In order to describe the agent's beliefs, it is enough to describe the agent's beliefs on primitive propositions, as well as the set of propositions the agent is aware of. If at some states ω and ω', the agent is aware of the same propositions and has the same beliefs on primitive propositions, then the agent's beliefs on every proposition is the same at ω and at ω'. Furthermore, the theorem shows it is enough to restrict our attention to awareness of propositions of depth at most 1: all beliefs are fully described by beliefs on primitives and awareness of propositions of depth at most one. In particular, two states in which the truth value of primitives, the agent's beliefs on primitives, and the awareness of propositions of depth one are the same, coincide.

6 Semantics

The state space models introduced in Sections 4 and 5 are syntactic: each state corresponds to a truth assignment for every proposition, including the agent's belief. Semantic models offer an alternative representation of the agent's beliefs.

Formally, a semantic model is a tuple (Ω, P), where Ω is the state space and $P : \Omega \to 2^{\Omega}$ denotes the agent's possibility correspondence: at state Ω, $P(\omega)$ is the set of states that the agent considers as possible.

It is common to define a belief operator B from the possibility correspondence by the relation:

$$BE_{\phi} := \{\omega \in \Omega : P(\omega) \subseteq E_{\phi}\}$$

where $E_{\phi} := \{\omega' \in \Omega : \phi \in \omega'\}$. The relation $\omega \in BE_{\phi}$ reads "the agent (semantically) believes ϕ at ω", and holds whenever ϕ is true at all states in $P(\omega)$.

Each syntactic model has a natural semantic counterpart. On a syntactic state space Ω, we define a possibility correspondence P by

$$P(\omega) := \{\omega' \in \Omega : b(\omega) \subseteq \omega'\}$$

where $b(\omega) := \{\phi \in \Phi : \beta\phi \in \omega\}$.

Aumann (1999) shows that Ω_5 has a semantic representation, i.e. with the possibility correspondence P and the semantic belief operator B defined as above, the syntactic and the semantic beliefs coincide at all states in Ω_5. Formally, $BE_{\phi} = E_{\beta\phi}$ for every proposition ϕ. This result is very useful, as it allows to work equivalently using either the semantic or the syntactic model, and semantic models can often be manipulated more easily than syntactic ones.

We extend Aumann's result to Ω_u. To do so, we rely on the possibility correspondence P defined as above, but use a different definition of semantic beliefs. Let $AE_{\phi} := \{\omega \in \Omega_u : \neg u\phi \in \omega\}$ be the set of states at which the agent is aware of ϕ, and $BE_{\phi} := \{\omega \in \Omega : P(\omega) \subseteq E_{\phi}\}$ as above. We define the belief operator B_u by:

$$B_u E_{\phi} := AE_{\phi} \cap BE_{\phi}.$$

According to this definition, in order for the agent to (semantically) believe ϕ, it does not suffice that ϕ holds everywhere in $P(\omega)$. We also require that the agent is aware of ϕ. This requirement is in line with the idea of implicit and explicit belief, introduced by Fagin and Halpern (1988), which is further discussed in Section 7.1.

The following result formalizes the equivalence between the syntactic beliefs in Ω_u, and semantic beliefs induced by B_u.

Theorem 5. $B_u E_{\phi} = E_{\beta\phi}$.

The following result shows that our definition of B_u is indeed semantic, i.e. $B_u E$ does not depend on the particular proposition ϕ such that $E = E_\phi$.

Proposition 5. *If $E_{\phi_1} = E_{\phi_2}$, then $B_u E_{\phi_1} = B_u E_{\phi_2}$.*

Note that if the set of primitive propositions is finite, then every subset of Ω_u corresponds to some proposition in Φ, and therefore $B_u E$ is defined for every subset E of Ω_u.

The following consequence of Theorem 5 shows that if a proposition is true in every state of Ω_u, i.e. it is a tautology of Ω_u, the agent believes it in every state of Ω_u.

Corollary 1. $B_u \Omega_u = \Omega_u$.

To prove the Corollary, observe that tautologies of the form $\beta\phi$, $\phi \in F$, are believed at every state in Ω_u. Proposition 5 shows that this is actually the case for every tautology, i.e. if ϕ is a tautology of Ω_u, it is believed by the agent at every state of Ω_u.

The Corollary has the desirable, natural implication that Ω_u is a complete representation of the modal logic system F. Furthermore, since F coincides with S5 (see Theorem 1), we obtain the result that propositions in S5 are believed at every state of Ω_u.

Corollary 2. $B_u \Omega_5 = \Omega_u$.

This last corollary emphasizes the fact that, at all states in Ω_u, including those not belonging to Ω_5, the agent perceives Ω_5 as the "actual" state space. This is in line with the main idea of the paper that the agent reasons "as if" the state space was Ω_5, even in states outside Ω_5. We discuss this point further in Section 7.5.

7 Discussion

7.1 Explicit and implicit beliefs

The notion of unawareness was first formalized, in the context of Modal Logic, by Fagin and Halpern (1988). Their paper introduces separate modalities for explicit belief – which is equivalent to the standard notion of belief – and for implicit belief, which can be thought as the set of logical consequences of the explicitly believed propositions. A proposition is explicitly believed whenever the agent implicitly believes it and is aware of it. The relationship to our model becomes more transparent when one looks at the semantic representation in Section 6. In this semantic model, implicit belief corresponds to the usual semantic belief operator: ϕ is implicitly believed at ω when $P(\omega) \subseteq E_\phi$. Our notion of belief in the semantic model requires both implicit belief and awareness, it is therefore a notion of explicit belief. Since belief in the syntactic model is equivalent to implicit belief and awareness in the semantic model, it is also to be thought of as explicit belief.

7.2 Reasoning, unawareness and knowledge

Modica and Rustichini (1994) show that in a model of knowledge (i.e. assuming that every proposition believed by the agent holds), when positive introspection holds, then negative introspection is equivalent to the symmetry axiom[4]. This result implies that unawareness cannot be modeled using knowledge unless either symmetry or reasoning is relaxed. Although in their follow-up paper (Modica and Rustichini, 1999) they acknowledge the desirability of a unified reasoning process by mentioning that "it is not at all the case that a subject who is aware of fewer things than another must necessarily be less capable of logical reasoning than the latter", they defend the idea that the agent's reasoning must be relaxed in order to model unawareness.

Following Modica and Rustichini (1999), one strand of literature (Heifetz, Meier, and Schipper, 2006, 2008; Li, 2009) studies models of unawareness in which the agent can reason only about the propositions he is aware of.

Our model departs from the above mentioned literature in that we consider (confident) belief, instead of knowledge. Modeling unawareness in a belief model was already pointed out by Modica and Rustichini (1999) as in important question. To quote them:

> Knowledge excludes the possibility that the agent "knows" or, better, "believes", something which is false. This distinction opens a question: the one of defining and analysing awareness with belief, rather than knowledge. This would consist in dropping the truth axiom from our system. This is an important question, that we do not discuss here.

7.3 Canonical states

Our approach follows that of Aumann (1999) and Samet (1990) in that a state assigns a truth value to all propositions, including those concerning the agent's beliefs about his own beliefs, and so on. Several papers – including, for instance, Modica and Rustichini (1994), Halpern (2001) and Heifetz, Meier, and Schipper (2006, 2008) – restrict their study

[4] According to the symmetry axiom, the agent is aware of ϕ if and only if he is aware of $\neg\phi$.

to so-called "canonical states" by making the additional assumption that two states that coincide on primitives and on the agent's beliefs on primitives necessarily coincide. As shown by Proposition 4, this is without loss in generality when considering states in Ω_5. On the other hand, distinct states on Ω_u may agree on primitives and on the agent's beliefs on primitives, as these do not suffice to describe a state in Ω_u. This can be seen for instance in Example 2.

Theorems 4 shows, on the other hand, that beliefs of propositions of high epistemic order are unnecessary to describe a state, as two states that coincide on primitives and on beliefs of propositions of epistemic depth at most one necessarily coincide. This shows that, in the state space with unawareness, the appropriate notion of a canonical state is a description of primitives together with beliefs of propositions of epistemic depth at most one.

7.4 Awareness properties

Several properties of awareness have been previously introduced and studied, and are considered by some authors as desiderata for a concept of awareness. The model introduced by (Heifetz, Meier, and Schipper, 2006, 2008) is constructed so as to fulfill these properties. Our approach is more agnostic on the properties of unawareness, and tries to derive these properties from a natural representation of the agent's mind rather than imposing them. Note however that, as shown by Example 2, it is always possible to restrict attention to subsets of Ω_u where some extra structure on awareness (such as symmetry, for instance) is imposed.

7.5 States and perceived states

In our model there exist states (from the modeler's point of view) in which the agent is unaware of some primitive propositions, but the agent does not consider such states as being possible. The distinction between the states deemed as possible by the agent and those which may actually arise is not new in the literature. For instance, Geanakoplos (1989) shows that apparent failures of rational information processing can be explained by the existence of states of the world that the agent ignores. More recently, in their fundamental work showing some of the main difficulties arising when modeling unawareness, Dekel, Lipman, and Rustichini (1988) show that this distinction is necessary in order to capture any meaningful notion of unawareness.

The distinction between the states of the world that can actually arise and those considered as possible by the agent can best be seen in the semantic model of Section 6. States belong to two categories. The states in Ω_5 are the states in which the agent is aware of every proposition. As in Bacharach (1985) and Samet (1990), the agent's possibility correspondence defines a partition of the states in Ω_5. The agent does not consider states outside of Ω_5 as being possible. At such states (outside of Ω_5), the only states the agent considers as possible belong to Ω_5, hence the agent exhibits delusion.

7.6 Extensions

Our model explicitly distinguishes beliefs arising from two different sources: universal beliefs, which follow from faith, and contingent beliefs, which follow from observation at every instance. We show that this distinction provides a natural rationale for unawareness and delusion, while maintaining the agent's reasoning process.

Alternative models can be obtained under variations of the faith system, or under different assumptions on contingent beliefs. The exploration of these variations could provide a fruitful direction for future research.

7.7 Final remarks

Unawareness has been a recognized phenomenon for several decades, and constitutes a long-standing modeling puzzle in formal epistemology. Its importance in economics is exemplified by recent work (see, e.g. Feinberg (2004) and Heifetz, Meier, and Schipper (2009)) showing that unawareness allows to capture the agent's behavior in a way that differs significantly from the classical framework of incomplete information à la Harsanyi (1967/68). This paper introduces a model of unawareness which is both simple and has intuitive appeal. We hope that, by throwing light on the formation of agent's beliefs and by offering a tractable model, it will contribute to a better understanding of the role of unawareness in economic contexts.

References

Aumann, R. (1999): "Interactive epistemology I: Knowledge," *International Journal of Game Theory*, 28, 263–300.

Aumann, R. J. (1976): "Agreeing to disagree," *The Annals of Statistics*, 4, 1236–1239.

Bacharach, M. (1985): "Some extensions of a claim of Aumann in an axiomatic model of knowledge," *Journal of Economic Theory*, 37, 167–190.

Chellas, B. (1980): *Modal logic: an introduction.* Cambridge University Press, Cambridge, Uk.

Dekel, E., B. Lipman, and A. Rustichini (1988): "Standard State-Space Models Preclude Unawareness," *Econometrica*, 66(1), 159–173.

Fagin, R., and J. Halpern (1988): "Belief, Awareness, and Limited Reasoning," *Artificial Intelligence*, 34, 39–76.

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

Feinberg, Y. (2004): "Subjective Reasoning – Games with Unawareness," Research Paper Series 1875, Stanford University.

Geanakoplos, J. (1989): "Game Theory without Partitions, and Applications to Speculation and Consensus," Cowles Foundation Discussion Paper 914, Yale University.

Halpern, J. (1995): "The effect of bounding the number of primitive propositions and the depth of nesting on the complexity of modal logic," *Artificial Intelligence*, 75, 361–372.

Halpern, J. (2001): "Alternative semantics for unawareness," *Games and Economic Behavior*, 37, 321–339.

Halpern, J., and L. Rêgo (2008): "Interactive unawareness revisited," *Games and Economic Behavior*, 62, 232–262.

Harsanyi, J. (1967/68): "Games with incomplete information played by 'Bayesian' players, Parts I, II, and III," *Management Science*, 14, 159–182, 320–334, and 486–502.

Heifetz, A., M. Meier, and B. Schipper (2006): "Interactive unawareness," *Journal of Economic Theory*, 130, 78–94.

——— (2008): "A canonical model for interactive unawareness," *Games and Economic Behavior*, 62, 304–324.

——— (2009): "Dynamic unawareness and rationalizable behavior," Discussion Paper ????, UC Davis.

Hintikka, J. (1962): *Knowledge and Belief.* Cornell University Press, Ithaca, NY.

Li, J. (2009): "Information structures with unawareness," *Journal of Economic Theory*, 144, 977–993.

Lipman, B. (1995): "Information Processing and Bounded Rationality: A Survey," *Canadian Journal of Economics*, 28, 42–67.

Modica, S., and A. Rustichini (1994): "Awareness and Partitional Information Structures," *Theory and Decision*, 37, 107–124.

——— (1999): "Unawareness and partitional information structures," *Games and Economic Behavior*, 27(2), 265–298.

Samet, D. (1990): "Ignoring Ignorance and Agreeing to Disagree," *Journal of Economic Theory*, 52, 190–207.

Simon, H. (1955): "A behavioral model of rational choice," *The Quarterly Journal of Economics*, 69(1), 99–118.

A Appendix

A.1 Proofs of Section 4

Definition 5. *Let* $(\phi_1 \xrightarrow{F} \phi_2)$ *be a shorthand for the following statement:*

$$\textit{if } \phi_1 \in F \textit{ then } \phi_2 \in F.$$

Lemma 1. *For* $\phi_1, \phi_2, \phi_3, \phi_4 \in \Phi$*:*

1. $(\phi_1 \rightarrow \phi_2) \xrightarrow{F} (\neg\phi_2 \rightarrow \neg\phi_1)$,

2. $\big((\phi_1 \rightarrow \phi_2) \wedge (\phi_2 \rightarrow \phi_3)\big) \xrightarrow{F} (\phi_1 \rightarrow \phi_3)$,

3. *If* $\phi_1 \xrightarrow{F} \phi_3$ *and* $\phi_2 \xrightarrow{F} \phi_4$*, then* $(\phi_1 \wedge \phi_2) \xrightarrow{F} (\phi_3 \wedge \phi_4)$.

Proof. 1. It follows directly from the definition of the implication.

2. Consider the following sequence of tautologies:

$$\begin{aligned}\big((\phi_1 \to \phi_2) \wedge (\phi_2 \to \phi_3)\big) &\overset{\text{F}}{\to} \big((\neg\phi_1 \wedge \neg\phi_2) \vee (\neg\phi_1 \wedge \phi_3) \vee (\phi_2 \wedge \neg\phi_2) \vee (\phi_2 \wedge \phi_3)\big)\\ &\overset{\text{F}}{\to} \big((\neg\phi_1 \wedge \neg\phi_2) \vee (\neg\phi_1 \wedge \phi_3) \vee (\phi_2 \wedge \phi_3)\big)\\ &\overset{\text{F}}{\to} (\neg\phi_1 \vee \phi_3)\\ &\overset{\text{F}}{\to} (\phi_1 \to \phi_3).\end{aligned}$$

3. The following relationships hold:

$$\begin{aligned}(\phi_1 \wedge \phi_2) \overset{\text{F}}{\to} \phi_1 \overset{\text{F}}{\to} \phi_3,\\ (\phi_1 \wedge \phi_2) \overset{\text{F}}{\to} \phi_2 \overset{\text{F}}{\to} \phi_4.\end{aligned}$$

That is, if $(\phi_1 \wedge \phi_2) \in F$ then $(\phi_3 \wedge \phi_4) \in F$. □

Lemma 2. *For some $\phi \in \Phi$, let the agent have faith in the truth axiom and introspection for ϕ and $\neg\phi$. Then, the agent has faith in the truth axiom and introspection for $\beta\phi$ and $\neg\beta\phi$.*

Proof. Truth axiom: It follows by hypothesis that $(\beta\phi \to \phi) \in F$. Thus,

$$(\beta\phi \to \phi) \overset{\substack{\text{(by } A_I)\\ \text{F}}}{\to} (\beta\beta\phi \to \beta\phi).$$

It follows from A_3 that $(\beta\phi \to \beta\beta\phi) \in F$. Thus,

$$\begin{aligned}(\beta\phi \to \beta\beta\phi) &\overset{\substack{\text{(by faith in } A_6 \text{ and Lemma 1)}\\ \text{F}}}{\to} (\beta\phi \to \neg\beta\neg\beta\phi)\\ &\overset{\substack{\text{(by Lemma 1)}\\ \text{F}}}{\to} (\beta\neg\beta\phi \to \neg\beta\phi).\end{aligned}$$

Positive introspection: It follows by hypothesis that $(\beta\phi \to \beta\beta\phi) \in F$ and $(\neg\beta\phi \to \beta\neg\beta\phi) \in F$. Thus,

$$(\beta\phi \to \beta\beta\phi) \overset{\substack{\text{(by } A_I)\\ \text{F}}}{\to} (\beta\beta\phi \to \beta\beta\beta\phi),$$

and

$$(\neg\beta\phi \to \beta\neg\beta\phi) \overset{\substack{\text{(by } A_I)\\ \text{F}}}{\to} (\beta\neg\beta\phi \to \beta\beta\neg\beta\phi).$$

Negative introspection: It follows by hypothesis that $(\beta\phi \to \beta\beta\phi) \in F$. Thus,

$$\begin{aligned}
(\beta\phi \to \beta\beta\phi) \quad &\overset{\text{(by Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\beta\phi \to \neg\beta\phi)\\
&\overset{\text{(by faith in } A_5 \text{ and Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\beta\phi \to \beta\neg\beta\phi)\\
&\overset{\text{(by faith in } A_3 \text{ and } R_I\text{)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\beta\phi \to \beta\neg\beta\phi) \wedge (\beta\neg\beta\phi \to \beta\neg\beta\beta\phi)\\
&\overset{\text{(by Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\beta\phi \to \beta\neg\beta\beta\phi).
\end{aligned}$$

It follows by hypothesis that $(\neg\beta\phi \to \beta\neg\beta\phi) \in F$. Thus,

$$\begin{aligned}
(\neg\beta\phi \to \beta\neg\beta\phi) \quad &\overset{\text{(by Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\neg\beta\phi \to \beta\phi)\\
&\overset{\text{(by faith in } A_4 \text{ and Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\neg\beta\phi \to \beta\beta\phi)\\
&\overset{\text{(by faith in } A_4 \text{ and Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\neg\beta\phi \to \beta\beta\beta\phi)\\
&\overset{\text{(by faith in } A_6 \text{ and } A_I\text{)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\neg\beta\phi \to \beta\beta\beta\phi) \wedge (\beta\beta\beta\phi \to \beta\neg\beta\neg\beta\phi)\\
&\overset{\text{(by Lemma 1)}}{\xrightarrow{\mathrm{F}}} \quad (\neg\beta\neg\beta\phi \to \beta\neg\beta\neg\beta\phi),
\end{aligned}$$

which completes the proof. □

Lemma 3. *Let the agent have faith in the truth axiom and introspection for ϕ_1 and ϕ_2. Then, the agent has faith in the truth axiom and introspection for $\phi_1 \wedge \phi_2$.*

Proof. Truth axiom: It follows from A_1 that $\big(\beta(\phi_1 \wedge \phi_2) \to (\beta\phi_1 \wedge \beta\phi_2)\big) \in F$. Then,

$$\big(\beta(\phi_1 \wedge \phi_2) \to (\beta\phi_1 \wedge \beta\phi_2)\big) \quad \overset{\text{(by faith in } A_3\text{)}}{\xrightarrow{\mathrm{F}}} \quad \big(\beta(\phi_1 \wedge \phi_2) \to (\phi_1 \wedge \phi_2)\big).$$

Positive introspection: Likewise,

$$\begin{aligned}
&\big(\beta(\phi_1 \wedge \phi_2) \to (\beta\phi_1 \wedge \beta\phi_2)\big)\\
\overset{\text{(by faith in } A_4\text{)}}{\xrightarrow{\mathrm{F}}} \quad &\big(\beta(\phi_1 \wedge \phi_2) \to (\beta\phi_1 \wedge \beta\phi_2)\big) \wedge \big((\beta\phi_1 \wedge \beta\phi_2) \to (\beta\beta\phi_1 \wedge \beta\beta\phi_2)\big)\\
\overset{\text{(by Lemma 1 and faith in } A_1\text{)}}{\xrightarrow{\mathrm{F}}} \quad &\big(\beta(\phi_1 \wedge \phi_2) \to \beta\beta(\phi_1 \wedge \phi_2)\big).
\end{aligned}$$

Negative introspection: It follows from A_1 and Lemma 1 that $\big(\neg\beta(\phi_1 \wedge \phi_2) \to$

$(\neg\beta\phi_1 \vee \neg\beta\phi_2)\big) \in F$. Thus,

$$\big(\neg\beta(\phi_1 \wedge \phi_2) \to (\neg\beta\phi_1 \vee \neg\beta\phi_2)\big) \quad \overset{\text{(by faith in } A_5)}{\overset{\text{F}}{\to}} \quad \big(\neg\beta(\phi_1 \wedge \phi_2) \to \beta\neg\beta(\phi_1 \wedge \phi_2)\big),$$

which completes the proof. □

Lemma 4. *Let the agent have faith in introspection for ϕ_1 and ϕ_2. Then $\big(\beta(\beta\phi_1 \vee \phi_2) \leftrightarrow (\beta\phi_1 \vee \beta\phi_2)\big) \in F$.*

Proof. It follows from $(\beta\phi_1 \to \beta\beta\phi_1) \in F$ that $\big((\beta\phi_1 \vee \beta\phi_2) \to (\beta\beta\phi_1 \vee \beta\phi_2)\big) \in F$. Thus, it follows from A_2 and Lemma 1 that

$$\big((\beta\phi_1 \vee \beta\phi_2) \to \beta(\beta\phi_1 \vee \phi_2)\big) \in F.$$

For the converse, it follows by definition that $\big(\beta(\beta\phi_1 \vee \phi_2) \to \beta(\neg\beta\phi_1 \to \phi_2)\big) \in F$. Thus,

$$\begin{aligned}
\big(\beta(\beta\phi_1 \vee \phi_2) \to \beta(\neg\beta\phi_1 \to \phi_2)\big) \quad &\overset{\text{(by definition)}}{\overset{\text{F}}{\to}} \quad \big(\neg\beta(\beta\phi_1 \vee \phi_2) \vee \beta(\neg\beta\phi_1 \to \phi_2)\\
&\overset{\text{(by } A_I)}{\overset{\text{F}}{\to}} \quad \big(\neg\beta(\beta\phi_1 \vee \phi_2) \vee (\beta\neg\beta\phi_1 \to \beta\phi_2)\big)\\
&\overset{\text{(by definition)}}{\overset{\text{F}}{\to}} \quad \big(\beta(\beta\phi_1 \vee \phi_2) \to (\neg\beta\neg\beta\phi_1 \vee \beta\phi_2)\big)\\
&\overset{\text{(by faith in } A_4 \text{ and Lemma 1)}}{\overset{\text{F}}{\to}} \quad \big(\beta(\beta\phi_1 \vee \phi_2) \to (\beta\phi_1 \vee \beta\phi_2)\big),
\end{aligned}$$

which completes the proof. □

Lemma 5. *Let the agent have faith in the truth axiom and introspection for ϕ_1 and ϕ_2. Then the agent has faith in the truth axiom and introspection for*

1. $\beta\phi_1 \vee \phi_2$, *and*
2. $\neg\beta\phi_1 \vee \phi_2$.

Proof. 1. It follows from Lemma 4 that $\big(\beta(\beta\phi_1 \vee \phi_2) \to (\beta\phi_1 \vee \beta\phi_2)\big) \in F$.
TRUTH AXIOM: Thus,

$$\big(\beta(\beta\phi_1 \vee \phi_2) \to (\beta\phi_1 \vee \beta\phi_2)\big) \quad \overset{\text{(by faith in } A_3 \text{ and Lemma 1)}}{\overset{\text{F}}{\to}} \quad \big(\beta(\beta\phi_1 \vee \phi_2) \to (\beta\phi_1 \vee \phi_2)\big).$$

Positive introspection: Thus,

$$\begin{aligned}
&\big(\beta(\beta\phi_1 \vee \phi_2) \to (\beta\phi_1 \vee \beta\phi_2)\big)\\
\overset{\text{(by faith in } A_4 \text{ and Lemma 1)}}{\xrightarrow{\text{F}}}\quad &\big(\beta(\beta\phi_1 \vee \phi_2) \to (\beta\beta\beta\phi_1 \vee \beta\beta\phi_2)\big)\\
\overset{\text{(by faith in } A_2\text{)}}{\xrightarrow{\text{F}}}\quad &\big(\beta(\beta\phi_1 \vee \phi_2) \to \beta\beta(\beta\phi_1 \vee \phi_2)\big).
\end{aligned}$$

Negative introspection: It follows from Lemma 1 that $\big(\neg\beta(\beta\phi_1 \vee \phi_2) \to \neg(\beta\phi_1 \vee \beta\phi_2)\big) \in F$. Thus,

$$\begin{aligned}
\big(\neg\beta(\beta\phi_1 \vee \phi_2) \to \neg(\beta\phi_1 \vee \beta\phi_2)\big) \quad \overset{\text{(by faith in } A_0 \text{ and Lemma 1)}}{\xrightarrow{\text{F}}}\quad &\big(\neg\beta(\beta\phi_1 \vee \phi_2) \to (\neg\beta\phi_1 \wedge \neg\beta\phi_2)\big)\\
\overset{\text{(by faith in } A_5 \text{ and Lemma 1)}}{\xrightarrow{\text{F}}}\quad &\big(\neg\beta(\beta\phi_1 \vee \phi_2) \to (\beta\neg\beta\phi_1 \wedge \beta\neg\beta\phi_2)\big)\\
\overset{\text{(by Lemmas 1 and 3)}}{\xrightarrow{\text{F}}}\quad &\big(\neg\beta(\beta\phi_1 \vee \phi_2) \to \beta\neg(\beta\phi_1 \vee \beta\phi_2)\big)\\
\overset{\text{(by Lemma 4)}}{\xrightarrow{\text{F}}}\quad &\big(\neg\beta(\beta\phi_1 \vee \phi_2) \to \beta\neg\beta(\beta\phi_1 \vee \phi_2)\big).
\end{aligned}$$

2. It follows from faith in A_5 and Lemma 2 that $\big((\neg\beta\phi_1 \vee \phi_2) \leftrightarrow (\beta\neg\beta\phi_1 \vee \phi_2)\big) \in F$. Then, the proof is identical to Case 1, when applied for $(\beta\neg\beta\phi_1 \vee \phi_2)$. □

Proof of Theorem 1. Recall that we define Φ_n as the closure of the set $\{\phi, \beta\phi \mid \phi \in \Phi_{n-1}\}$ with respect to $\neg$, $\vee$ and $\wedge$. It is straightforward to verify that $\Phi_\infty := \bigcup_{n\geq 0} \Phi_n$ is such that $\Phi_\infty = \Phi$.

Thus, we prove the theorem by induction: we show that if the agent has faith in the truth axiom and introspection for all $\phi \in \Phi_n$, then he also has faith in the truth axiom and introspection for all $\phi' \in \Phi_{n+1}$. This follows directly from Lemmas 2, 3 and 5. □

Proof of Theorem 2. Let $A_3 - A_5$ be satisfied for all primitives at some $\omega \in \Omega_r$. By hypothesis, every proposition in A is satisfied at ω. Furthermore, the inference rules also hold locally at ω:

- $(\phi_1 \wedge \phi_2) \in \omega$, if and only if $\phi_1 \in \omega$ and $\phi_2 \in \omega$
- If $\phi_1 \in \omega$ and $(\phi_1 \to \phi_2) \in \omega$, then $\phi_2 \in \omega$
- $\beta\phi \in \omega$, for all $\phi \in F$

Thus, applying the steps of the proof of Theorem 1, locally at ω, shows that the truth axiom and introspection is satisfied for all propositions at ω. Therefore, all propositions in F hold at ω, which completes the proof. $\square$

A.2 Proofs of Section 5

Consider any sequence $\tilde{\beta} = \tau_1, \ldots, \tau_n$, $n \geq 1$, where for every $i = 1, \ldots, n$, $\tau_i = \beta$ or $\tau_i = \neg\beta$. For such a sequence $\tilde{\beta}$, we define its parity $p(\tilde{\beta}) \in \{0, 1\}$ as the parity of the number of occurrences of $\neg\beta$ in $\tilde{\beta}$. For instance, $p(\neg\beta\beta\neg\beta) = p(\beta) = 0$, whereas $p(\beta\neg\beta) = p(\neg\beta\neg\beta\neg\beta) = 1$, i.e., $p(\tilde{\beta}) = 0$ if the number of negations in $\tilde{\beta}$ is even, and $p(\tilde{\beta}) = 1$ otherwise.

Lemma 6. *Let $\omega \in \Omega_u$. For any two sequences $\tilde{\beta}$ and $\tilde{\beta}'$ such that $p(\tilde{\beta}) = p(\tilde{\beta}')$ and for any proposition $\phi \in \Phi$, we have $(\tilde{\beta}\phi \leftrightarrow \tilde{\beta}'\phi) \in F$.*

Proof. It follows inductively from Theorem 1. $\square$

Corollary 3. *Let $\omega \in \Omega_u$. For any two sequences $\tilde{\beta}$ and $\tilde{\beta}'$ such that $p(\tilde{\beta}) = p(\tilde{\beta}')$ and for any proposition $\phi \in \Phi$, we have $\beta(\tilde{\beta}\phi) \in \omega$, if and only if $\beta(\tilde{\beta}'\phi) \in \omega$.*

Proof. It follows directly from Lemma 6 and A_I. $\square$

Proof of Proposition 1. Let $(\neg\beta\phi \wedge \neg\beta\neg\beta\phi) \in \omega$, and suppose there is some $\phi' \in B(\phi)$ such that $\beta\phi' \in \omega$. By definition, $\beta\phi'$ can be rewritten as $\beta\tilde{\beta}\phi$. If $p(\tilde{\beta}) = 0$ then $(\beta\phi' \leftrightarrow \beta\phi) \in \omega$ which contradicts $\neg\beta\phi \in \omega$, whereas if $p(\tilde{\beta}) = 1$ then $(\beta\phi' \leftrightarrow \beta\neg\beta\phi) \in \omega$ which contradicts $\neg\beta\neg\beta\phi \in \omega$. Hence, $u\phi \in \omega$ implies $\neg\beta\phi' \in \omega$ for all $\phi' \in B(\phi)$. The converse is straightforward. $\square$

Proof of Proposition 2. Suppose the contrary: the agent is aware of all primitive propositions. Then the proof is identical to that of Theorem 2. Note that in order to prove that introspection is well-founded in Ω_u, we do not require the truth axiom for the primitive to be well-founded. $\square$

Proof of Proposition 3. By definition, if ϕ in F, then $\beta\phi \in \omega$ for every $\omega \in \Omega_u$. Assume that $\beta\phi \in \omega$ for every $\omega \in \Omega_u$, then $\beta\phi \in \omega$ for every $\omega \in \Omega_5$. Since the Truth axiom holds on Ω_5, this implies that $\phi \in \omega$ for every $\omega \in \Omega_5$. Therefore ϕ is a tautology of Ω_5, hence a theorem of S5, hence an element of F. $\square$

Proof of Theorem 3. First we show that for every $\phi \in \Phi$ there is another proposition $\phi_1 \in \Phi_1$ such that $(\phi \leftrightarrow \phi_1) \in F$, and therefore if ω and ω' coincide in the truth value of ϕ_1 they will also coincide in the truth value of ϕ – because of A_I – which would suffice for the proof.

It follows from Lemmas 3, 4 and 6 that for every $\phi_n \in \Phi_n$ there is some $\phi_{n-1} \in \Phi_{n-1}$ such that $(\phi_n \leftrightarrow \phi_{n-1}) \in F$. Continue inductively to obtain $(\phi_n \leftrightarrow \phi_1) \in F$, which completes the proof. □

Proof of Theorem 4. Let $\phi \in \Phi_1$. For any $\omega \in \Omega_u$, there are two possibilities: either $u\phi \in \omega$, implying $\neg\beta\phi \in \omega$, or $a\phi \in \omega$, implying $(\beta\phi \vee \beta\neg\beta\phi) \in \omega$, and therefore $(\beta\phi \vee \neg\beta\phi) \in \omega$. The truth value of $\omega(\beta\phi_0)$ for all $\phi' \in \Phi_0$ determines whether $\beta\phi \in \omega$ or $\neg\beta\phi \in \omega$. Thus, the truth value of $\omega(\beta\phi_1)$ is determined for all $\phi_1 \in \Phi_1$, and therefore the proof follows directly from Theorem 3. □

A.3 Proofs of Section 6

Proof of Theorem 5. First, we show that $E_{\beta\phi} \subseteq B_u E_\phi$. Let $\omega \in E_{\beta\phi}$, which – by definition – is equivalent to $\beta\phi \in \omega$. Then it follows from the definition of the possibility correspondence that $\phi \in \omega'$ for all $\omega' \in P(\omega)$, implying $\omega \in BE_\phi$. Furthermore, $\beta\phi \in \omega$ yields $\omega \in AE_\phi$, as required.

Now we show that $E_{\beta\phi} \supseteq B_u E_\phi$. Let $\omega \in B_u E_\phi$. It follows from $\omega \in \Omega_u$ that the agent believes all propositions in F at ω, and therefore – by the definition of P – all states $\omega' \in P(\omega)$ are well-founded. Hence, $P(\omega) \subseteq \Omega_5$. Furthermore, A_4 is satisfied at ω. Hence, $P(\omega') \subseteq P(\omega)$, for all $\omega' \in P(\omega)$ (Samet, 1990).

Now, suppose that $\neg\beta\phi \in \omega$. It follows from $\omega \in B_u E_\phi$ that the agent is aware of ϕ at ω, implying $\beta\neg\beta\phi \in \omega$ (see Proposition 1). Hence, we obtain $\neg\beta\phi \in \omega'$ for all $\omega' \in P(\omega)$. It follows from Aumann (1999) that – since $\omega' \in \Omega_5$ – there is some $\omega'' \in P(\omega')$, such that $\neg\phi \in \omega''$. Finally, from $P(\omega') \subseteq P(\omega)$, it follows that $\omega'' \in P(\omega)$, thus contradicting $\omega \in BE_\phi$, and therefore it also contradicts $\omega \in B_u E_\phi$, which completes the proof. □

Proof of Proposition 5. For arbitrary $\phi_1, \phi_2 \in \Phi$, let $E_{\phi_1} = E_{\phi_2}$, implying that $(\phi_1 \leftrightarrow \phi_2)$ is a tautology in Ω_u, and therefore also in Ω_5. Hence, $(\phi_1 \leftrightarrow \phi_2)$ is a theorem in S5, implying that $(\phi_1 \leftrightarrow \phi_2) \in F$. Finally, it follows from R_F that $\beta(\phi_1 \leftrightarrow \phi_2) \in \omega$, and therefore $(\beta\phi_1 \leftrightarrow \beta\phi_2) \in \omega$ for all $\omega \in \Omega$, which completes the proof. □

Proof of Corollary 1. It follows by definition that $\Omega_u = E_{\phi\vee\neg\phi}$. Hence,

$$\begin{aligned} B_u\Omega_u &= B_u E_{\phi\vee\neg\phi} \\ &\overset{\text{(by Theorem 5)}}{=} E_{\beta(\phi\vee\neg\phi)} \\ &\overset{\text{(by } B_F\text{)}}{=} \Omega_u \end{aligned}$$

which completes the proof. □

Proof of Corollary 2. It follows directly from the fact that every Ω_5-tautology belongs to F, and therefore the agent believes it at all states in Ω_u. □