

Formal Epistemology: Lecture Notes

Horacio Arló-Costa
Carnegie Mellon University
`hcosta@andrew.cmu.edu`

Fixed Point models for full belief

The main point of these lectures is to study the agent reasoning autoepistemically about its own beliefs, rather than agents engaged in belief attribution.

Acceptance as true and *full belief* are closely related epistemic attitudes. Commitment to full belief is mirrored by commitment to accept as true; and commitment not to accept is mirrored by commitment not to fully believe.

Focus on the set of *consistent saturated theories* σ of L (containing a modal operator B) obeying the following two constraints:

$$(A1) \quad A \in \sigma \text{ iff } B(A) \in \sigma$$

$$(A2) \quad A \notin \sigma \text{ iff } \neg B(A) \in \sigma$$

σ can be seen in this context as a *commitment set* representing the doxastic commitments of an agent at certain instant t . One can say that an agent explicitly believes a finite set of sentences M , but that he is doxastically committed to the closure of M – under the classical notion of logical consequence. Then membership in σ represents commitment to accept and lack of membership in σ represents commitment no to accept.

An *information model* is a set of theories of L closed under A1-2. One can then ask which are the modal formulae that are accepted in all saturated theories of a given model. We can call such formulae *positively valid in a model*, and *positively valid* when positively valid in all models.

Theorem 1 *The thesis of the modal system S5 are exactly the positively valid thesis.*

Proof We need to show that a formula A of L is positively valid if and only if A is a theorem of S5. From left to right the proof is not demanding. From right to left the proof is slightly harder.

We will assume that α is not a S5 theorem and we will show that α does not belong to some saturated theory T . Take $K = S5 \cup \neg B(\alpha)$. K is consistent. For suppose that K

were inconsistent. Then, by axiom (3), α is an S5 theorem.

We form now a list \mathbf{L} of all formulae of L : $g_1, g_2, \dots, g_n, \dots$. With respect to this list, we construct an infinite sequence of sets

$$I_0, I_1, \dots, I_n, \dots$$

as follows. As I_0 we take K , i.e.,

$$I_0 = K$$

Then, for each positive integer n we set:

$$I_{i+1} = \begin{cases} Cn(I_i \cup \{g_{i+1}\}) & \text{if } I_i, g_{i+1} \text{ is consistent} \\ I_i & \text{otherwise} \end{cases}$$

Form, then, a Lindenbaum-type set: $\overline{K} = C(\cup I_i)$, where $\cup I_i$ denotes the union of all the infinitely many sets I_i . \overline{K} is a complete S5-theory. Nevertheless \overline{K} is not saturated. It is interesting to see why not. Consider any formula A added at stage i . At stage

i + j one could consistently add $\neg B(A)$ - if \mathbf{L} is such that $B(A)$ has not been added yet. This is so because the formula $A \wedge \neg B(A)$, although epistemically problematic, is logically consistent. To put it in a different way, the modal system S5 is unable to capture syntactically the paradoxical nature of Moore's paradox ('this is my hand, but I do not fully believe it').*

But the formula $A \wedge \neg B(A)$ does not belong

*The philosopher G. E. Moore offered different variants of his paradox. A canonical version appears in [?].

to any consistent saturated theory. For say (by contradiction) that $A \wedge \neg B(A)$ is in a consistent saturated theory T . Then $\neg B(A) \in T$ and hence (by A2) $A \notin T$ which leads to an immediate contradiction. So, the standard Henkin method used in modal logic to prove completeness does not suffice. We propose the following *Moore saturation*[†] of K :

[†]Notice that the Moore saturation of a set K does not need to be an *extension* of \overline{K} – although it is constructed as a function of \overline{K} . For consider the case where the initial K is paradoxical (in the sense of G.E. Moore), i.e. it contains both A and $\neg B(A)$. The Lindenbaum extension of K will contain both A

$$\mathcal{M} = \{A: B(A) \in \overline{K} \}$$

\mathcal{M} is consistent. For suppose by contradiction that $\perp \in \mathcal{M}$. Then, $B(\perp) \in \overline{K}$. But since $S5 \subseteq \overline{K}$, $\neg B(\perp) \in \overline{K}$. Therefore we have a contradiction because it is easy to check that \overline{K} is consistent. \mathcal{M} is a S5 theory. It is easy to see that all S5 theorems are in \mathcal{M} . In fact, $B(A)$ is a S5-theorem whenever A is a S5-theorem and, by construction, all S5-theorems are in K . In general, assume that A and $\neg B(A)$, but the saturation of K will only contain $\neg B(A)$ - and it will *not* contain A .

entails B. Then $B(A \rightarrow B)$ is in \overline{K} . Assume that A is in \mathcal{M} . Then $B(A) \in \overline{K}$. Therefore $B(B) \in \overline{K} - \mathcal{M}$ is a complete S5-theory – and this guarantees that B is in \mathcal{M} . Assume that A, B are in \mathcal{M} . Then $B(A), B(B)$ are in \overline{K} . Now the formula $B(A) \wedge B(B) \rightarrow B(A \wedge B)$ is a S5-theorem, and therefore $A \wedge B$ is in \mathcal{M} .

We will check now that \mathcal{M} is closed under A1 and A2. First we will check A1. Assume that $B(A) \in \mathcal{M}$. Assume now by contradiction that $A \notin \mathcal{M}$. Then $B(A) \notin \overline{K}$, and since \overline{K}

is a complete S5 theory, $\neg B(A) \in \overline{K}$. Therefore, by negative introspection, $B(\neg B(A)) \in \overline{K}$. This, in turn, entails that $\neg B(A) \in \mathcal{M}$, against the consistency of \mathcal{M} . Assume now that $A \in \mathcal{M}$. Then $B(A) \in \overline{K}$. By positive introspection, $B(B(A)) \in \overline{K}$. This yields the desired result, namely that $B(A) \in \mathcal{M}$.

Secondly we should check A2. Assume that $A \notin \mathcal{M}$. Then $B(A) \notin \overline{K}$, and since \overline{K} is a complete S5 theory, $\neg B(A) \in \overline{K}$. Therefore, by negative introspection, $B(\neg B(A)) \in \overline{K}$. This, in turn, entails that $\neg B(A) \in \mathcal{M}$, as

desired. Finally assume that $\neg B(A) \in \mathcal{M}$. Assume by contradiction that $A \in \mathcal{M}$. Then $B(A) \in \overline{K}$. Positive introspection guarantees that $B(B(A)) \in \overline{K}$. But then $B(A) \in \mathcal{M}$, against the consistency of \mathcal{M} .

So, \mathcal{M} is a consistent and saturated S5 theory. But, by construction, $\neg B(\alpha) \in \mathcal{M}$. In fact, $\neg B(\alpha) \in K \subseteq \overline{K}$. Therefore, by negative introspection, $B(\neg B(A)) \in \overline{K}$, which is enough to guarantee that $\neg B(\alpha) \in \mathcal{M}$. But now, by A2, $\alpha \notin \mathcal{M}$. This completes the proof.



A better understanding of the above result can be achieved by embedding the previous model in a more complicated one, where time and the perspectives of different agents are represented.

Let $I = \{1, \dots, n\}$ be a set of rational agents. The underlying language is a Boolean standard language (L_0) augmented by the modal operators $B_{i,t}$ for each i in I and for each instant t . These various operators represent the (full) beliefs of different agents at different instants.

By the same token we can index our acceptance sets. So we will have sets $\sigma_{i,t}$ representing the sentences accepted by agent i at instant t . Notice that the interpretation of the epistemic operators, when embedded in the sets $\sigma_{i,t}$ is context dependent.

In fact, $B_{i,t}(\dots) \in \sigma_{j,t'}$ means that agent j is certain at time t' that agent i is certain that ... is the case at time t . $B_{i,t}(\dots) \in \sigma_{i,t'}$ with $t < t'$ holds whenever agent i is certain that he *was* certain that $B_{i,t}(\dots) \in \sigma_{i,t'}$ with $t > t'$ holds whenever agent i is certain that he *will be* certain that Finally $B_{i,t}(\dots) \in \sigma_{i,t}$ holds whenever agent i holds 'I am certain that ...' at time t .

We can focus on the set of all possible certainty sets of agent i at time t . The only constraint that we impose on these sets is that they must be closed under logical consequence and under the principles (A1-2).

i's background knowledge at t can be reduced to the logical consequences of a Boolean sentence, say p , or *i*'s certainties at t can contain several other items, including some certainties about other agents – for example *i* could be certain that *j* is certain that $\neg p$.

The fact that the modelling used in the previous slides did not appeal to indexes for agents and times, is only due to the fact that (A1-2) hold in the case in which the indexes in the acceptance sets and the indexes of the modal operators coincide:

$$(A1) \ A \in \sigma_{i,t} \text{ iff } B_{i,t}(A) \in \sigma_{i,t}$$

$$(A2) \ A \notin \sigma_{i,t} \text{ iff } \neg B_{i,t}(A) \in \sigma_{i,t}$$

Therefore (A1-2) can be formally studied without loss of generality by dropping indexation. Nevertheless conceptual clarity about the scope of the modelling might be loss.

Consider for example the *alethic* axiom:

$$(3) B_{i,t}(A) \rightarrow A.$$

The completeness result shows that (3) is positively valid, and this means that all substitution instances of (3) belong to every possible saturated set of i at t . This, of course, *does not* mean that if i is certain that A at t , then A is the case. What this result shows is that all rational agents at time t are certain of the truth of their certainties.

Consider the system that can be obtained from S5 by deleting the so-called alethic condition (axiom (3) in our presentation of S5) and by replacing it with the axiom $\neg B_{i,t}(\perp)$. This system is usually called KD45 and it is considered a good axiomatization of third-person belief (not certainty or full belief).

The logic of belief attributions might very well be axiomatized by KD45. The logic of belief attributions and the logic of full belief can be bridged as follows:

(T) A is positively valid for agent i at t iff
 $\vdash_{KD45_{i,t}} B_{i,t}(A)$.

In other words epistemic validity can be captured in terms of (third-person KD45) belief attributions (via T). By the same token (T) explains away paradoxical sentences (like G.E. Moore's) as *unbelievable* sentences (in accordance with the analysis offered by Hintikka in Knowledge and Belief).

Epistemic States: Other theories

A ranking function κ is a function from \mathcal{M} to the set of extended non-negative integers $\mathcal{N}^+ = \mathcal{N} \cup \{\infty\}$, such that $\kappa(w) = 0$, for some $w \in \mathcal{M}$. For each proposition $P \subseteq \mathcal{M}$ the *rank* $\kappa(P)$ of P is defined by $\kappa(P) = \min \{\kappa(w) : w \in P\}$ and $\kappa(\emptyset) = \{\infty\}$.

Spohn proposes to interpret ranks as *grades of disbelief*. $\kappa(P) = 0$ says that P is not disbelieved at all. It does not say that P is believed; this is rather expressed by $\kappa(P^c) > 0$, i.e., that non- P is disbelieved (to some degree). The set $C_\kappa = \{w: \kappa(w) = 0\}$ is called the *core* of κ and C_κ is the strongest proposition believed (to be true) in κ .

This account has already a dynamic flavor absent in the previously reviewed views. So, if A^c is believed to be true in κ , one way of representing the *contraction* of A^c from C_κ is to take the union of C_κ with the set of least disbelieved $\neg A$ points, i.e. $\{w: \kappa(w) = \kappa(\neg A)\}$. We will pay detailed attention to this process of contraction in the coming sections, devoted to belief change.