Logic of a Self-Transparent Believer

Bas C. van Fraassen

Abstract

Moore's Paradox engendered various proposals for aspects of the logic of belief, both for believers to avoid falling into its form of incoherence and for special principles to serve as axioms or rules for doxastic logic. The proposal here developed is to study the logic pertaining to believers who are self-transparent in the sense that, although they may have many false beliefs, they are right about what their beliefs are. The logic of the language of factual description of their situation is a normal modal logic **KDC4C4**, but is to be distinguished from the internal logic that governs what follows from their beliefs, on pain of incoherence. The adequacy and completeness proofs for that logic show it to be, in some respects, severely non-classical.

1 Introduction

Beliefs are often, perhaps even typically, false. Writing *B*p for the factual statement that the agent in question believes that it is the case that p, it is clear that *B*p and p can have, and often do have, opposite truth-values. But starting with G.E. Moore's seminal paper [6] it is recognized that one cannot assert that, or believe both the propositions expressed, without falling into some sense of incoherence. This has however proved quite puzzling. Agents cannot coherently state that p while denying that they believe that p, but neither can they assert the contradictory, in any general way. For the claim that always, either they believe that p or it is not the case that p, is a claim implying clairvoyance or access to a magical crystal ball. Moore's paradox appears to present a dilemma with incoherence on either horn.

The challenge for the logic of belief is then to distinguish, and formulate precisely, the relevant criteria of incoherence distinct from classical inconsistency or logical falsehood. Equivalently, the task is to make explicit the logic by which beliefs must follow from or block one another, the form of reasoning that is sound within coherent belief.

2 Introducing the self-transparent believer

To recognize Moore's point in practice takes some clarity about one's own cognitive or doxastic state, some transparency with respect to one's beliefs. If a robot were constructed to simulate rational belief and were programmed only to avoid logical falsehoods and unsatisfiable theories, it would have no qualms about entering into instances of Moore’s paradox. Only if it were programmed to have correct beliefs about what its beliefs are, could it possibly be in a position to escape such incoherence.

Accordingly I propose to study the concept of believers who, though perhaps wrong about many things, are right about what beliefs they have. With the notation introduced above, that would require the logical principles

1] If *B*p then *BB*p
2] If *BB*p then *B*p

to hold. I will call such believers *self-transparent*.[1]

These sorts of principles are not unfamiliar from efforts to formulate a logic of belief within normal modal logic, and that would seem at first blush to be the framework within which to begin. This leads us to set up a simple language for belief attribution, in which we can state both what is the case in a world and, with respect to a single believer, what is believed there. As I will elaborate briefly below, the logic that is sound and complete for this language is essentially known already, and easy to formulate.

But given that, I propose a different focus for logical exploration: to articulate what counts as coherent or incoherent belief content for such a believer, and what follows for such believers themselves when they reflect on what is implied by the beliefs they have. Specifically, of course, the relevant coherence must imply that they are not prey to instances of Moore’s paradox. But beyond that we want to have a sound and complete logic that captures their valid inference patterns.

Will classical logic remain? Yes, classical logic will remain sound, in that all the classically valid arguments remain valid. However, certain familiar meta-rules, such as reductio ad absurdum, will not. There is a complete logic of transparent belief, and despite the preserved soundness of classical logic, the two are not the same.

3 Distinguishing two consequence relations

Intuitively, statement p is a *classical consequence* of premises X exactly if p is true in all worlds in which those premises are true. The similar intuitive explanation of doxastic consequence will be: p is believed in all those worlds where those premises are believed.

Let us call the belief set in world w the set of statements **B**(w) = {p: *B*p is true in world w}, and introduce the symbol ⇒ for the doxastic consequence relation. Then we can say equivalently that p is a doxastic consequence of X if and only if p belongs to every belief set which contains X. This defines a consequence relation subject to the basic structural rules:

For all sets of sentences X and sentences p:

[1] Principle 2 has appeared in the literature, though rarely, but with a different interpretation, such as that it characterizes a ‘stable’ reasoner, in the sense that suggests a commitment not to doubt one’s own beliefs about what one believes. Taking modal statements as statements of fact, however, 2] states factually that the agent’s beliefs about what they believe are true, and that is the reading here maintained.

(Identity) if p is in X then $X \Rightarrow p$

(Weakening) if $X \Rightarrow p$ and $X \subseteq Y$ then $Y \Rightarrow p$

(Transitivity) if $X \Rightarrow p$ and $Y \Rightarrow q$ for every member q of X
then $Y \Rightarrow p$

(To shorten the text somewhat, from here on "X" will be used for a set of sentences and "p" for a sentence. But for clarity I will mostly use the phrase "is a doxastic consequence of" rather than the special symbol "⇒".)

Our purpose is now to explore the relation between the classical and doxastic consequence relations in the special case in which the relative doxastic possibility relation between worlds is the appropriate one for modeling the self-transparent believer.

4 The classical consequence relation

Our language L has the usual form of syntax of normal modal logic, with denumerably many atomic sentences and the connectives &, ∨, ⊃, *B*. Before displaying its semantics, we formulate here the logic that we mean to be sound and complete for this language:

R0. If p is a theorem of classical propositional logic then $\vdash p$

R1. If $\vdash p$ then $\vdash Bp$

R2. $p, p \supset q \vdash q$

A1. $\vdash B(p \supset q) \supset (Bp \supset Bq)$

A2. $\vdash Bp \supset BBp$

A3. $\vdash BBp \supset Bp$

A4. $\vdash Bp \supset \sim B \sim p$

Rule 2 extends the set of theorems to a deductive consequence relation. By R2 and axiom A1 it follows that if q follows from $p_1, \ldots, p_k$ then Bq follows from $Bp_1, \ldots, Bp_k$. That is, the notion of belief here is such that it includes all that is believed implicitly, in the sense of being something that follows logically from what is believed.

In the standard notation for normal modal logics, this logic is **KD4C4**.[2] But since the main concern of this essay is its interpretation as a logic for the self-transparent believer, I will use the mnemonic acronym **LSTB** for this logic. Interpretations of nearby logical systems have focused on deontic and tense logic.[3]

5 Semantic analysis

A model structure for language L is a couple $M = \langle W, R \rangle$ where W is a non-empty set, the *worlds*, and R is a binary relation on W, *relative doxastic possibility*. I will specify

[2] Here "4" stands for the S4 axiom A2, and "C4" for its converse, i.e. axiom A3, and "D" for the "deontic" principle, i.e. axiom A4 (Cf. Garson [2]).

[3] Axiom A3 is Garson's condition C4: "Density would be false if time were atomic, i.e. if there were intervals of time which could not be broken down into any smaller parts. Density corresponds to axiom (C4): $\Box \Box A \rightarrow \Box A$, [. . .] and **KDC4** [is] adequate with respect to models whose frames are serial and dense.. . . " ([3], no page number).

properties of R below.

Intuitively, if x and y are worlds, and xRy, then the beliefs the agent holds in world x do not rule out that y is the actual world, i.e. all the agent's beliefs in x are true in y.

A valuation ***v*** over M is an assignment of truth-values T, F to all sentences at each world in M, such that the truth tables for connectives &, ~ ∨, ⊃ are obeyed, and in addition:

> For all worlds x in W, ***v***(x, *B*p) = T iff ***v***(y, p) = T in all worlds y such that xRy.

We also write $\boldsymbol{v}_x$(p) for ***v***(x, p)

> $\boldsymbol{v}_x$ *satisfies* sentence p (rep. set of sentences X) iff $\boldsymbol{v}_x$(p) = T (resp. $\boldsymbol{v}_x$(q) = T for all members q of X).
>
> A sentence is *valid* in language L iff it receives T from all valuations, at all worlds, over all model structures.
>
> A set of sentences X is *satisfiable* in language L iff some valuation satisfies it.
>
> A set of sentences X *semantically entails* sentence p in language L iff p receives truth value T from all valuations, over all model structures at all worlds therein, which satisfy X.

But R needs to have special properties if the axioms are to present valid sentences. The condition on R to ensure validity for A2 is well-known from normal modal logic S4: R must be *transitive*.

Similarly, to ensure validity for A4, the requirement is again familiar: For each possible world there is a doxastically possible world as well: if x is a world there is a world y such that xRy. We can refer to the set of worlds to which x bears R as R(x). Then A4 says in effect that R(x) is never empty. A relation with this property is called *serial*.

What ensures the validity of A3, which is the converse of A2, is not as familiar. Here we may have recourse to a paper by Frederic Fitch [1] that deserves to be seminal. Fitch shows how there is a simple recipe in the traditional calculus of relations which relates modal logic principles with properties of the relative possibility relation. I won't put it in those traditional terms, now generally unfamiliar, but the required property for axiom A3 is this:

> R is *weakly reflexive*: for any worlds x, y, if xRy then there is a world z such that xRz and zRy.

We do not want R to be reflexive, though that would guarantee the validity of A3, but would also validate (*B*p ⊃ p).

It is more customary to call R *dense* if it has this property. The reasons to prefer calling this property *weak reflexivity* are two. First, if R is reflexive, so that xRx for all x, then it is weakly reflexive. Secondly, in the above formula we detect a threat of infinite regress, for of course it implies that if xRz then there must be some world u such that xRu and uRz, and so forth and so forth. But the regress can stop, and the property hold, if any of those worlds is possible relative to itself. For example if xRy and yRy then y itself can serve as the “middle” world – a dash of reflexivity will do. To sum up:

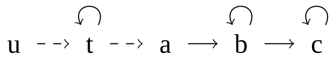
> M = $\langle$W, R$\rangle$ is a model structure for language L exactly if W is a non-empty set and R is a binary relation on W which is transitive, serial, and weakly reflexive.

With relevance to Moore’s paradox, and to doxastic consequence, it is also pertinent to point out what is not valid. Showing this also gives us the opportunity to display some specific simple models.

Thomason ([6], [7]) and Cross ([1]) included *B*(*B*p ⊃ p) as a principle of doxastic logic, which it seems was extrapolated from Hintikka’s extended discussion of Moore’s paradox with the conclusion that *B*(p & ~*B*p) as well as *B*(~p & *B*p) are indefensible ([4], 64–78, 123–125).[4] Extrapolations from Hintikka’s book to axioms systems in epistemic or doxastic logic are, at the least, difficult and contentious. In this case it is remarkable that the putative principle *B*(*B*p ⊃ p) plays a central role in the derivation of contradictions by the Gödel type arguments presented by Thomason and Cross in these articles (see further the critical discussion in van Fraassen [10], with reference to p. 17 numbered line (ii), and pp. 19–20). So it is pertinent that this putative principle cannot become a theorem in the logic of the self-transparent believer.

Theorem 1 *Neither B*(*B*p ⊃ p) *nor B*(p ⊃ *B*p) *is valid for all sentences* p.

We will define two model structures in which these formulas are falsified, and a diagram will help to guide the imagination. The arrows in the diagram represent the relative doxastic possibility relation, but for clarity, the arrows implied by the requirement of transitivity are not shown. To distinguish the two model structures, arrows that belong only to the first model structure are solid.



> Model structure M_1 = $\langle$\{a, b, c\}, $R_1\rangle$ with R_1(a) = \{b, c\}, R_1(b) = \{b, c\}, R_1(c) = \{c\}.

By inspection, R_1 is transitive and serial. We verify similarly that R_1 is weakly reflexive, noting that for example the fact that aR_1b is accompanied by the two facts that aR_1b and bR_1b, and that the fact that cR_1c is accompanied by the facts that cR_1c and cR_1c.

Valuation ***v*** is defined in part by: ***v***(a, p) = T, ***v***(b, p) = T, ***v***(c, p) = F, where p is alphabetically the first atomic sentence.

> It follows from this that ***v***(b, *B*p) = F, and hence also that ***v***(b, p ⊃ *B*p) = F.
>
> Therefore ***v***(a, *B*(p ⊃ *B*p)) = F, since aR_1b.

[4] See Thomason ([7, p. 9 numbered line 13; [6], p. 393, numbered line (ii)); in an endnote Thomason added “Other arguments to this effect can be found in that text, particularly in the applications Hintikka makes of a theory of belief meeting these conditions to matters such as Moore’s paradox of saying and disbelieving” ([6], p. 394).

Next, model structure M_2 is formed by adding to M_1 two new worlds in a way that duplicates part of the initial structure. That is,

> M2 = $\langle$\{u, t, a, b, c\}, $R_2\rangle$ with R_2(u) = \{t, a, b, c\}, R_2(t) = \{t, a, b, c\}, R_2(a) = \{b, c\}, R_2(b) = \{b, c\}, R_2(c) = \{c\}. By inspection, R_2 is transitive, serial, and weakly reflexive.

Valuation ***v'*** is defined in part by: ***v'***(a, p) = F, ***v'***(b, p) = T, ***v'***(c, p) = T.

> Therefore ***v'***(a, *B*p) = T, and hence ***v'***(a, *B*p ⊃ p) = F.
>
> Therefore ***v'***(u, *B*(*B*p ⊃ p)) = F, since uR_2a.

6 Soundness and completeness of logic LSTB

As Fitch [1] already pointed out, the soundness and completeness proofs for normal modal logics are easily adapted if some axiom is added and R is required to have the corresponding property defined by his general recipe. There is no reason to spell this out here. But we will be relying on those results later on:

> **LSTB** is sound: if X ⊢ p in **LSTB** then X semantically entails p in L.
>
> **LSTB** is complete: if X semantically entails p in L then X ⊢ p in **LSTB**.
>
> **Corollary to completeness**: if X is consistent in LSTB then X is satisfiable in L.

We note also that deducibility in **LSTB** is finitary, given the definition, and note without proof that semantic entailment in L is finitary. Hence the completeness claim and its corollary extend to infinite sets of sentences.

For soundness, let us just show that the one unfamiliar feature, axiom A3, is satisfied.

Given a model structure M = ⟨W, R⟩ and a world x in W, suppose that in world x, *B*p is false.

> So there is a world y such that xRy and p is false in y.
>
> By weak reflexivity, there is a world z such that xRz and zRy.
>
> Since zRy and p is false in y, it follows that *B*p is false in z.
>
> But then *BB*p is false in x, since xRz.

7 Approaching the doxastic consequence relation

The sentence *B*p is true in a world if it is believed there, by the agent of that world, that it is the case that p. Let us call the set of sentences that are believed in a given world a *belief* set. The doxastic consequence relation will be determined by what is characteristic of belief sets. So let's begin by asking: what is in a belief set?

Let us focus on a specific model structure M = ⟨W, R⟩ and a particular valuation ***v***. With reference to M and ***v*** understood, we will simply say that a sentence p is true in world w iff ***v***(w, p) = T. We then represent the contents of the agent's beliefs in world w, a member of W, as the agent's belief set, **B*v***(w) = {p: ***v***(w, *B*p) = T}.

In the context where M is specified I will refer to the 'access region' of world w as R(w) = {u in W: wRu}. It is clear then that **B*v***(w) = {p: for all worlds x in R(w), ***v***(x, p) = T}.

The *doxastic consequence relation* is then defined by:

> sentence p is a *doxastic consequence* of set of sentences X if and only if for all models M = ⟨W, R⟩, worlds w in W, and valuations ***v*** on M, if all members of X belong to **B*v***(w) then p belongs to **B*v***(w).

8 Characteristics of belief sets

With our focus on a single model structure M and valuation ***v*** it will be more convenient to suppress reference to them and to just write **B**(w) = {p: *B*p is true in world w}. Its members are the sentences which are true in every world doxastically possible relative to w.

> T0. If **B**(w) ⊢ p then p is a member of **B**(w).

For if p is true in all the worlds in which all members of **B**(w) are true then p is true in all the worlds in R(w), and hence p is in **B**(w).

> T1. If *B*p is in **B**(w), then p is in **B**(w).

For if *B*p is in **B**(w) then *BB*p is true in w, by the soundness of the relevant axiom A3 of **LSTB**, hence *B*p is true in w. Therefore p is true in all members of R(w), and hence belongs to **B**(w).

> T2. If p is in **B**(w), then *B*p is in **B**(w).

For if p is in **B**(w) then *B*p is true in w, hence by the soundness of the axiom A2 of **LSTB**, *BB*p is true in w, and so *B*p is in **B**(w).

Remark 2 *We see therefore that in the agent's doxastic reasoning, each of a sentence* p *and sentence B*p *can be inferred from each other.*

In classical logic the Deduction Theorem, or in natural deduction formulation the rule of Conditional Proof, would lead, from the correctness of these inferences, to the corresponding conditionals. Not so in doxastic reasoning:

> T3. Neither (*B*p ⊃ p) nor (p ⊃ *B*p) belongs to all belief sets for all sentences p.

This follows from Theorem 1, given the soundness and completeness of **LSTB**.

But there is also no way in which a rational self-transparent believer will land in Moore's Paradox:

> T4. Neither (*B*p & ~p) nor (~*B*p & p) belongs to any belief set.

For suppose that (*B*p & ~p) is true in all members of R(w). Then both *B*p and ~p are true in all members of R(w). So *BB*p and *B*(~p) are true in w. By the soundness of axiom A3 of **LSTB**, *B*p and *B*(~p) are then both true in w, which implies that both p and ~p are true in all members of R(w). By axiom A4, R(w) is not empty, so there is a world in which p and ~p are both true, which is impossible.

Equally, suppose that (~*B*p & p) belongs to belief set **B**(w). Then p is in **B**(w), hence by T2 above, *B*p is in **B**(w). So both ~*B*p and *B*p are in **B**(w), but the soundness of axiom A4 guarantees that a belief set is not inconsistent.

9 Characteristics of the doxastic consequence relation

Whatever is the logic of the doxastic consequence relation it is clear from T1–T4 that it is non-classical.

Before formulating the logic **LDOX**, which captures the doxastic consequence relation, we can discern what some of its theorems can or cannot be, on the assumption that those theorems are precisely the sentences that are true in all belief sets.

Failure of a meta-rule: *the Rule of Conditional Proof.* As noted, T1 means that p is a doxastic consequence of *B*p, while T3 shows that (*B*p ⊃ p) is not a theorem. So in this case,

> *B*p ⊢ p in **LDOX**, but it is not the case for all sentences p that ⊢ (*B*p ⊃ p).

The Deduction Theorem for classical logic entails the admissibility of the Rule of Conditional Proof, which would entitle us to infer that ⊢ (*B*p ⊃ p) from *B*p ⊢ p, but note well that Conditional Proof is not an ‘ordinary rule’ like Modus Ponens. It is a *meta-rule*. That is, it does not present simply the form of a valid argument, but rather the form of inference from validity of some arguments to the validity of others.

Another failure of a meta-rule: the rule of *Reductio ad Absurdum*

1.	(*B*p & ~p)	assumption
2.	*B*p	from 1
3.	p	from 2, because *B*p ⊢ p
4.	~p	from 1
5.	~(*B*p & ~p)	from 1–4 by Reductio ad Absurdum
6.	(*B*p ⊃ p)	from 5

We know that for arbitrary sentences p, 6 is not a theorem of **LDOX**, due to fact T3. The sequence 1–6 above is not a correct deduction (derivation) in **LDOX**. The upshot is that Reductio ad Absurdum is not an admissible rule for **LDOX**.

Nevertheless, on the positive side, we can note that the following will have to hold:

> If p is theorem of **LSTB** then p is a theorem of **LDOX**.
>
> Modus Ponens is an admissible rule for **LDOX**.

The first is clear because to be complete **LSTB** is accountable to the fact that if p is true in all worlds then it is a member of all belief sets, and so needs to be a theorem. A similar argument applies, mutatis mutandis, to inferences by Modus Ponens.

So belief sets are closed under the consequence relation of **LSTB**, but not all rules admissible for **LSTB** are admissible for **LDOX**. We will just have to be very careful not to rely on meta-rules!

10 Logical system LDOX

Motivated by the above quasi-intuitive reflections, the logical system **LDOX**, for the same syntax as **LSTB** but with its above indicated interpretation, is the following:

A1. If p is a theorem of **LSTB** then $\vdash$ p

R1. p, p $\supset$ q $\vdash$ q

R2. p $\vdash$ *B*p

R3. *B*p $\vdash$ p

Note that none of the three rules are meta-rules; no meta-rule is to be assumed to be admissible at this point. However

Theorem 3 *If* X $\vdash$ p *in* **LSTB** *then* X $\vdash$ p *in* **LDOX**.

That is so because A1 and R1 together provide us will all inferences in **LSTB**. The meta-rule that if p is a theorem then so is *B*p, in **LSTB**, will have helped to generate the set of theorems there. But it is not needed in deductions in **LDOX**, given that any theorem of **LSTB** can appear as a line in a **LDOX** derivation with no further justification.

This may all be clear enough, but because **LDOX** is not, despite its origin, a normal modal logic, and actually non-classical in some respects, it will be well to spell out some of its basic characteristics. We begin with an explicit definition of the relation $\vdash$ of deducible derivability in LDOX:

> A *derivation from* set of sentences X, in **LDOX**, is a finite sequence of sentences of L, each of which either belongs to X, or is a theorem of **LSTB**, or follows from preceding sentences by application of one of rules R1, R2, or R3.

The last member of a derivation is called its conclusion, and it is clear that every member of a derivation from X is also the conclusion of some derivation from X. All members p of the sequence are said to be (*deductively*) *derivable* from X, in symbols X $\vdash$ p.

This relation of derivability is thus, by its definition, a finitary relation, and it obeys the Structural Rules:

(Identity) if p is in X then X $\vdash$ p

(Weakening) if X $\vdash$ p and X $\subseteq$ Y then Y $\vdash$ p

(Transitivity) if X $\vdash$ p and Y $\vdash$ q for every member q of X then Y $\vdash$ p

A set of sentences X is a *theory* in **LDOX**, or LDOX-theory, exactly if all sentences deductively derivable from X are members of X. The important relation between theories and deductive derivability is this:

Theorem 4 *It is not the case that* $\mathrm{X} \vdash \mathrm{p}$ *in* **LDOX** *if and only if there is an LDOX-theory which contains* X *but does not contain* p.

The syntax, and hence the set of derivations from X, is only countably infinite. So let the derivations be enumerated as D(1), D(2), ... and define the series **S** = {X(1), X(2), ...} as follows: X(1) = X; X(n + 1) = X(n) ⋃ {q}, where q is the conclusion of D(n). Since ⊢ is a finitary relation, the union of **S** is a theory. A sentence p can belong to **S** only if it already belonged to one of the sets X(k), that is, only if it is derivable from X.

11 Initial completeness claims for LDOX

A set of sentences is *consistent* in a given logic iff there is some sentence not deducible from that set, in that logic. A set of sentences X is a *theory* in a given logic iff X is consistent in that logic and contains all that follows from it by deductive derivation in that logic. We shall index these terms to **LSTB** and **LDOX** by writing "LDOX-consistent", "LSTB-consistent". "LDOX-theory", and so forth, or equivalently, "consistent in LDOX", "theory of LSTB", and so on, to mark these distinctions. The main connections between these characteristics can be summed up as follows.

Theorem 5 *(a) If* X *is LDOX-consistent then it is LSTB-consistent.*

(b) If X *is an LDOX-theory then* X *is an LSTB-theory.*

(c) If X *is LDOX-consistent then* X *is part of an LDOX-theory.*

(d) If X *is an LDOX-theory then for any sentence* p, p *is in* X *if and only if B*p *is in* X.

(e) Any belief set is an LDOX-theory; to be precise, if M = ⟨W, R⟩ *is a model structure for* L, w *a world in* W, *and* ***v*** *a valuation on* M, *then* $\mathbf{B}_{\boldsymbol{v}}(\mathrm{w})$ *is an LDOX-theory.*

Most of this is immediate; note that characteristics T0, T1, T2 of belief sets establish (e).

The completeness property that we require **LDOX** to have is that it 'catalogues' the doxastic consequence relation. To be precise, if a set of sentences X is LDOX-consistent then we require (a) that there is a world w, in some model structure, such that X is part of **B**(w), and (b) if X does not deductively imply p in **LDOX** then we require there to be a world w, in some model structure, such that X is part of **B**(w), but p does not belong to **B**(w). Intuitively, the latter corresponds to there being a possible believer who believes all that is claimed in X but does not believe that p.

Lemma 6 *If* X *is an LDOX-theory then there is a model structure* M = ⟨W, R⟩ *for* L, *valuation* ***v*** *on* M *and world* w *in* W *such that* $\mathrm{X} \subseteq \mathbf{B}_{\boldsymbol{v}}(\mathrm{w})$.

For let X be an LDOX-theory, and define $\mathrm{X}^* = \{B\mathrm{p}\colon \mathrm{p} \in \mathrm{X}\}$. Clearly, for any sentence p, p ∈ X iff *B*p ∈ X iff *B*p ∈ X^*, hence $\mathrm{X}^* \subseteq \mathrm{X}$.

X^* is LDOX-consistent, because $\mathrm{X}^* \subseteq \mathrm{X}$; and therefore also LSTB-consistent.

Hence by the soundness of **LSTB**, there is a model structure M = ⟨W, R⟩ for L, valuation ***v*** on M and world w in W such that $\boldsymbol{v}_{\mathrm{w}}$ satisfies X^*. It follows that for all sentences p, if *B*p is in X^* then p is in $\mathbf{B}_{\mathbf{v}}$(w). Therefore X ⊆ $\mathbf{B}_{\mathbf{v}}$(w).

We have to imagine that, since in our representation belief and disbelief are constrained only by logical considerations, a given LDOX-theory X could be the whole of what a given agent believes. That would be the case precisely if for each sentence p, either *B*p is true in the world of that agent, and p is in X, or else ~*B*p is true in that world. We verify this as follows.

Theorem 7 *If* X *is an LDOX-theory then there is a model structure* M = ⟨W, R⟩ *for* L, *valuation **v** on* M *and world* w *in* W *such that* X = $\mathbf{B}_{\boldsymbol{v}}$(w).

Let X be an LDOX-theory and X^* = {*B*p: p ∊ X}. Let the family **F** of sets of sentences be defined by:

Y is in **F** if and only if:

(a) $X^* \subseteq Y$

(b) Y is LSTB-consistent

(c) if p is a member of Y then there is a sentence q such that either p = *B*q and p belongs to X^*, or p = ~*B*q

First, X^* belongs to **F**. As seen in the preceding lemma, $X^* \subseteq$ X, and is LSTB-consistent; and if p is a member of X then by definition, *B*p is in X^*. Given condition (a), X^* is therefore the smallest member of **F**.

Second, if q is a sentence not in X then it is not the case that $X^* \vdash$ *B*q in **LSTB**. For $X^* \subseteq$ X, hence if $X^* \vdash$ *B*q then X would contain *B*q; since X is an LDOX-theory it would then also contain q. Therefore $X^* \bigcup \{\sim Bq\}$ is LSTB-consistent if q is not in X.

Third, the family **F** is partially ordered by set inclusion. If Y_1, Y_2, . . . is a chain in **F** then its union is also in **F**, for that union will be consistent since ⊢ in **LSTB** is finitary. Hence all such chains have an upper bound; by Zorn's lemma, **F** has a maximal element Z.

Fourth, for every sentence p, either *B*p or ~*B*p belongs to Z. For if $Z \bigcup \{\sim Bp\}$ is LSTB-inconsistent then there are sentences $q_1, \ldots, q_k$ in Z such that $q_1, \ldots, q_k \vdash Bp$ in **LSTB**, and thus *B*p is in Z. By similar reasoning, if $Z \bigcup \{Bp\}$ is inconsistent then ~*B*p is in Z. Finally, if Z is consistent with *B*p and with ~*B*p and neither is in Z, then the addition of either sentence to Z would be a member of **F** and so Z would not be maximal.

By the completeness corollary for **LSTB** it follows that there is a model structure M = ⟨W, R⟩ for L, valuation ***v*** on M and world w in W such that $\boldsymbol{v}_{\mathrm{w}}$ satisfies Z. Hence any sentence p is in $\mathbf{B}_{\boldsymbol{v}}$(w) if and only if *B*p is in Z.

Finally, *B*p is in Z iff p is in X. For Z belongs to family **F**, hence *B*p is in Z iff *B*p is in X^*, which is the case iff p is in X.

12 The same logic, from another point of view

How logic is typically taught today tends to promote the impression that classical logic formulated with just axiom schemes and the rule of modus ponens is the same logical system as that presented in a natural deduction or Gentzen rule formulation. That is not so, as we have just seen. **LDOX** contains all of the former, but some familiar natural deduction rules (conditional proof, reductio ad absurdum) are not admissible for it. Those natural deduction or Gentzen rules are meta-rules (also sometime called epi-rules), and these can be violated without touching the theorems and 'ordinary' inferences.

This separation of 'ordinary' and 'meta' inferences is typical in supervaluation treatments that were designed to eliminate arbitrariness in assignments of truth-values (vide van Fraassen [8], pp. 491–492; [9], Chapter 5 section 3).

The plan for this section will have two parts. The first is to present a straightforward supervaluation treatment, to build a language in which the pertinent distinction is not between truth and falsity but between being believed and being disbelieved. The second is to show that the results for logic coincide precisely with what we found above for the doxastic consequence relation (Theorem 7).

To begin then, let us look at the self-transparent believer's belief sets in a different way, as determining valuations, analogous to a language that has truth-value gaps. We shall then assign Ts and Fs to represent the values '**is believed**' and '**is disbelieved**'. Many sentences will, generally, receive neither T nor F, though with that interpretation the 'gaps' are not truth-value gaps but belief-value gaps. Then we can ask: which inferences preserve the designated value T, what is the consequence relation that captures preservation (from premises to conclusion) of that value? A supervaluational language is always built on a simpler language. The simpler language comes with a well-defined class of 'classical' admissible valuations, and the valuations of the supervaluational language correspond to sets of those 'classical' valuations.

In our case the simpler language will be L that was introduced above, with the semantics spelled out in terms of model structures, worlds, and valuations.

The semantics will appear in two stages. We begin with the same definition of model structures M = ⟨W, R⟩, where W is called the set of worlds and R the relative doxastic possibility relation. R has the properties that we imposed before: it is transitive, and weakly reflexive, and serial. We now add a sort of shorthand that ignores some aspects of that semantics:

Definition 8 *An assignment* $\boldsymbol{f}$ *of a value of* T *or* F *to each sentence* p *is a classical valuation iff there is a model structure* M = ⟨W, R⟩, *a world* w *in* M, *and a valuation* $\boldsymbol{v}$ *on* M *such that for all sentences* p, $\boldsymbol{f}(\mathrm{p}) = \boldsymbol{v}_{\mathrm{w}}(\mathrm{p})$.

As above, we say that a valuation *satisfies* a set of sentences if and only if it assigns T to all the members of that set. With our new shorthand the semantic consequence relation is this:

> set X of sentences semantically entails sentence p in L if and only if all classical valuations which satisfy X also assign T to p

which from here on we will call the classical consequence relation, and define the classical consequence operator **CL**:

> **CL**(X) = {p: X semantically entails p in L}

Now we are ready to introduce a new language L*, which has the same syntax as L, but a different class of admissible valuations. First of all we specify a new relation N among sentences, the relation of *non-classical necessitation*: For any sentences p, q, it is the case that pNq if and only if either q is the sentence *B*p, or p is the sentence *B*q.

Definition 9 *A set of sentences is* saturated *if and only if it is satisfied by some classical valuation of* L *and is closed both under the classic consequence relation and under the relation* N.

We introduce the *non-classical consequence operator* as follows:

> **CNL**(X) is the least set that contains X and is closed both under the classic consequence relation and under the relation N.

Equivalently, X is saturated if and only if X is satisfiable in L and **CNL**(X) = X. Note thus that CNL(X) is itself a saturated set if X is satisfiable. Finally,

Definition 10 *An assignment **s** of a value of* T *or* F *to some sentences* p *is a supervaluation of* L *iff there is a saturated set* X *such that for all sentences* p*:*

> ***s***(p) = T *if and only if **f***(p) = T *for all classical valuations **f** that satisfy* X*;*
>
> ***s***(p) = F *if and only if **f**(*p*)* = F *for all classical valuations **f** that satisfy* X*;*
>
> ***s***(p) *is undefined otherwise.*

Definition 11 *An assignment of a value of* T *or* F *to some sentences* p *is an admissible valuation of language* L* *if and only if it is a supervaluation of* L.

Thus X semantically entails p in L* exactly if all supervaluations of L which satisfy X also satisfy p.

Theorem 12 X *semantically entails* p *in* L* *if and only if* p ∈ **CNL**(X).

Given the definitions, it is clear that if p is in **CNL**(X) then p belongs to any saturated set that contains X, and hence is satisfied by any supervaluation induced by a saturated set which contains X. Therefore if p is **CNL**(X) then X semantically entails p in L*.

For the converse, the set **CNL**(X) is either satisfied by some classical valuation or by no classical valuation. In the latter case, CNL(X) contains all sentences, so it follows trivially that X semantically entails p in L* only if p is in **CNL**(X).

In the former case, note that **CNL**(X) is itself a saturated set, and so induces a supervaluation. If that supervaluation does not satisfy p, then we have a counterexample to the claim that X semantically entails p in L*. Hence X semantically entails p in this case only if p is in **CNL**(X).

Lemma 13 *If* X *is a consistent LDOX-theory then* X *is saturated.*

An LDOX-theory is an LSTB-theory, and hence by the soundness of LSTB for language L, is satisfied by a classical valuation. By the same token, an LDOX-theory is closed under the classical consequence relation; moreover it is closed under N because of its rules R2 and R3.

Lemma 14 *If* X *is saturated then* X *is an LDOX-theory.*

If p is a classical consequences of X then X $\vdash$ p in **LSTB** since it is complete with respect to L. Being closed under N, a saturated set is closed under the **LDOX** inference rules as well.

Lemma 15 *An assignment* ***s*** *of a value of* T *or* F *to sentences is an admissible valuation of language* L^* *if and only if there is an LDOX-theory* X *such that for all sentences* p*,* $\boldsymbol{s}(p) = T$ *iff* p *is in* X *and* $\boldsymbol{s}(p) = F$ *iff* ~p *is in* X.

(Note that this assignment will not in general have all sentences in its domain.) This follows from the preceding two lemmas.

Finally then, the soundness and completeness of **LDOX** with respect to language L^*.

Theorem 16 X $\vdash$ p *in* **LDOX** *if and only if* $p \in$ **CNL**(X) *in* L^*.

Suppose p is not in **CNL**(X). Then it is not the case that X $\vdash$ p in **LDOX**, by Theorem 4, since there is an LDOX-theory, namely **CNL**(X) which contains X but not p.

Suppose conversely that it is not the case that X $\vdash$ p in **LDOX**. Then by Theorem 4, there is some LDOX-theory which contains X but not p. So by the preceding Lemma, there is a supervaluation which assigns T to all of X but not to p. Therefore by Theorem 12, p is not in **CNL**(X).

13 Conclusion: the self-transparent believer from a logical point of view

The logic of language L, which was designed for the factual description of agents whose beliefs are right about what beliefs they have, is **KD4C4**, which I gave as name also the mnemonic acronym **LSTB**. It is not a theorem of this logic that in general agents believe, for any belief they have, that they have that belief only if it is true. This is in contrast to a principle that was part of some logics of belief, and which, surprisingly, had both been argued for on the basis of Moore's paradox and been a crucial ingredient in derivation of paradoxical results.

But our focus has mainly been on characterizing the doxastic consequence relation. Intuitively, a conclusion is a doxastic consequence of some premises precisely if the beliefs of any self-transparent believer will include that conclusion if they include all the premises. The logic which catalogues the doxastic consequences, **LDOX**, has all the theorems and inferences of **LSTB** and two additional rules. These two rules capture what was right in the above mentioned principle: for transparent believers, managing the contents of their beliefs, the inferences from *B*p to p, and vice versa, are valid, but it is not the case that for just any sentence p, they believe the conditionals $(Bp \supset p)$ or $(p \supset Bp)$. Hence **LDOX** is non-classical, in that certain classical meta-theorems, such as the Deduction Theorem or the admissibility of the Reductio ad Absurdum, fail. This situation is illuminated by displaying **LDOX** as the sound and complete logic for a supervaluational language L^*, designed for doxastic reasoning.

References

[1] Cross, Charles B. (2001). A Theorem Concerning Syntactical Treatments of Nonidealized Belief. *Synthese* 129, pp. 335–341.

[2] Fitch, Frederic B. (1973). A Correlation between Modal Reduction Principles and Properties of Relations. *Journal of Philosophical Logic* 2, pp. 97–101.

[3] Garson, James (2018). Modal Logic. *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <`https://plato.stanford.edu/archives/fall2018/entries/logic-modal/`>

[4] Hintikka, Jaakko (1962). *Knowledge and Belief: an Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

[5] Moore, G. E. (1942). Russell's Theory of Descriptions. In P. Schilpp (Ed.), *The Philosophy of Bertrand Russell*, pp. 175–225. LaSalle: Open Court.

[6] Thomason, Richmond H. (1980). A Note On Syntactical Treatments Of Modality. *Synthese* 44, pp. 391–395.

[7] Thomason, Richmond H. (2011). Some Limitations to the Psychological Orientation in Semantic Theory. *Journal of Philosophical Logic* 40, pp. 1–14.

[8] Van Fraassen, Bas C. (1966). Singular Terms, Truth-Value Gaps, and Free Logic. *The Journal of Philosophy* 63, pp. 481–495.

[9] Van Fraassen, Bas C. (1971). *Formal Semantics and Logic*. N.Y.: Macmillan. URL = <`https://www.princeton.edu/˜fraassen/FS&L.htm`>

[10] Van Fraassen, Bas C. (2011). Thomason's Paradox for Belief, and Two Consequence Relations. *Journal of Philosophical Logic* 40, pp. 15–32.

Bas C. van Fraassen
fraassen@princeton.edu