Knowledge, Belief, Normality, and Introspection

Abstract

We study two logics of knowledge and belief stemming from the work of Robert Stalnaker [18], omitting positive introspection for knowledge. The two systems are equivalent with positive introspection, but not without. We show that while the logic of beliefs remains unaffected by omitting introspection for knowledge in one system, it brings significant changes to the other. The resulting logic of belief is non-normal, and its complete axiomatization uses an infinite hierarchy of coherence constraints. We conclude by returning to the philosophical interpretation underlying both models of belief, showing that neither is strong enough to support a probabilistic interpretation, nor an interpretation in terms of certainty or the "mental component" of knowledge.

The[1] starting point of this paper is the logic of knowledge and belief proposed by Robert Stalnaker in [18], and recently studied further in [1, 15]. In this system the only purely doxastic axiom is D. All other core principles for the logic of belief—K, 4, and 5— are theorems instead of axioms. They follow from a number of principles regimenting the interaction between knowledge and belief (see Table 2) together with the axioms for knowledge. Belief, furthermore, becomes definable in terms of the knowledge modality in that logic. An agent believes φ if and only if it is consistent with that agent's information that she knows φ.

In this paper we study the logic of belief that results from omitting positive introspection for knowledge in Stalnaker's system. While he explicitly rejects negative introspection, Stalnaker "provisionally" accepts that knowing implies knowing that one knows [18, p.173]. This principle, however, has been the subject of much discussion, starting with Hintikka's [10]. See [9] for an overview of the classical points of contention. In recent years Williamson's [21, chap.5] charge against the so-called "KK-principle" has attracted much attention. Williamson argues that if knowledge comes with a margin of error then assuming positive introspection leads to paradoxes. We do not take sides in this debate here. Rather, we investigate the logical question of what happens to the logic of belief in Stalnaker's system when knowledge is not introspective.

This paper is thus primarily aimed at epistemic logicians. It helps chart the landscape of combined epistemic and doxastic systems when knowledge is not positively introspective, viz. when the epistemic logic does not contain the 4 axiom. This complements Wolgang Lenzen's early work on epistemic-doxastic logics, in which knowledge is between S4 and S5 [14]. We offer a number of

[1]We would like to thank Alexandru Baltag, Johan van Benthem, Sonja Smets, Marta Bilkova, Ondrej Majer, Eric Pacuit, Vit Punochar, Igor Sedlar, two anonymous reviewers and the audience of LORI V for valuable feedback and suggestions.

K	$\vdash K(\varphi \to \psi) \to (K\varphi \to K\psi)$
NEC	If $\vdash \varphi$ then $\vdash K\varphi$
4	$\vdash K\varphi \to KK\varphi$
T	$\vdash K\varphi \to \varphi$

Table 1: S4 for knowledge

D	$\vdash B\varphi \to \langle B\rangle\varphi$
KB	$\vdash K\varphi \to B\varphi$
PI	$\vdash B\varphi \to KB\varphi$
NI	$\vdash \neg B\varphi \to K\neg B\varphi$
SB	$\vdash B\varphi \to BK\varphi$

Table 2: Stalnaker's axioms for the interaction between knowledge and belief

new completeness results together with a discussion of the interpretation of the resulting belief operators.

Section 1 presents Stalnaker's original system and some of its salient properties. It turns out that omitting positive introspection from that system gives rise to *two* rather different logics of beliefs. We present one at the end of Section 1, while the other is covered in Section 2. In Section 3 we ask whether this resulting belief operator can be supported by a probabilistic interpretation, and answer in the negative. Section 4 concludes by casting doubt on whether this belief operator can instead be read as "subjective certainty" or the "mental component" of knowledge.

1 Stalnaker's original system

In this section we review Stalnaker's original proposal. The presentation consolidates a number of known [1, 15, 18] and new results about that system. In the Appendix of this paper we also provide a cut-free display calculus for it.

1.1 Basic properties and axiomatization of the belief fragment

Throughout this paper we work with a propositional modal language augmented with one epistemic (K) and one doxastic (B) modality. Let $Prop$ be a countable set of atomic propositions. The language $\mathcal{L}$ is defined as follows.

$$\varphi := p \in Prop \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid B\varphi$$

$K\varphi$ and $B\varphi$ are read respectively as "the agent knows that φ" and "the agent believes that φ". We write $\langle K\rangle$ and $\langle B\rangle$ for the corresponding duals.

In Stalnaker's system, knowledge is taken to be normal, in the technical sense [3], factive, and positively introspective. It is an S4 modality (Table 1). Belief, on the other hand, is characterized through its relation with knowledge. This relation is encapsulated by the axioms in Table 2.

D is the only purely doxastic principle in this list. It rules out inconsistent pairs of beliefs. KB is the rather widespread assumption that knowledge implies belief. Unlike knowledge, PI and NI make belief fully transparent. The basic underlying idea here is that knowledge presupposes truth, a property to which the agent does not have direct access. Belief, on the other hand, is a mental state to which we have privileged and immediate access. So, the axiom goes, not only do agents have beliefs about their beliefs or lack thereof, they also *know* these facts about themselves. In other words, the agent is never uncertain nor mistaken in regard to what she believes.

Probably the most controversial axiom of this system is SB, standing for Strong Belief. It states that believing implies believing that one knows. One interpretation of this is that we are dealing with a form of belief that is close to absolute subjective certainty [1, 15], or, that is, the "mental component" of knowledge. This is certainly not the common-or-garden concept of belief, although here we side with Lenzen [14] in the view that might be called doxastic pluralism: that there is not one but many different concepts of belief, each to be captured by a different logical system. On this view the belief operator in Stalnaker's system is a very strong variety of belief. Logically speaking, it is in fact close to what Lenzen calls "being convinced of" (*überzeugt sein*) [14]. Hence the thought arises that one could interpret that operator, and thereby the axiom SB, by using the notions of absolute certainty or even the mental component of knowledge. The relation between knowledge and belief in this system turns out to be subtle, though, especially when knowledge is not positively introspective. So, for now, we leave the discussion of the interpretation of the belief operator under the Stalnaker axioms at that. We shall return to it in Sections 3 and 4.

Write S for the logic consisting of axioms and rule of Table 1 and 2, together with all propositional tautologies. Call the *belief fragment* $\mathcal{L}_B$ of $\mathcal{L}$ the set of all formulas in $\mathcal{L}$ that do not contain the K modality. The belief fragment of the *logic* S is defined as $\mathsf{S} \cap \mathcal{L}_B$.

Observation 1. *(cf. [18]) For all $\varphi \in \mathcal{L}_B$:*

$$\textit{If } KD45 \vdash \varphi \textit{ then } \mathsf{S} \vdash \varphi.$$

This result does not need positive introspection for knowledge. This will be seen to be important later. We sketch the proof here.

Proof of Observation 1. We show how to derive D, 4 and 5 for the belief operator, i.e. the formulas $B\varphi \to \langle B\rangle\varphi$ (D), $B\varphi \to BB\varphi$ (4) and $\neg B\varphi \to B\neg B\varphi$ (5). D is an axiom, and 4 and 5 follow directly from PI and NI, together with KB. Indeed, for 4, starting from $B\varphi$ one gets $KB\varphi$ by one application of PI, and then $BB\varphi$ follows from KB. The argument for 5 is completely analogous.

The proof of normality of B in that logic, i.e. that it admits the K axiom and the necessitation rule, is facilitated by what is probably the most crucial theorem of that logic:

$$B\varphi \leftrightarrow \langle K\rangle K\varphi \tag{EQ}$$

Observation 2. *(cf. [18]) $\mathsf{S} \vdash EQ$*

In this logic, believing is equivalent to the epistemic possibility of knowledge, i.e. one believes φ exactly when one's current knowledge is consistent with

knowing φ. The derivation of (EQ) notably does *not* involve positive introspection for knowledge. Again, this will be important later, so again we sketch the proof here.

Proof of Observation 2. Assuming that $B\varphi$ holds, we start by invoking SB to get $BK\varphi$. From there, we arrive at $\langle B\rangle K\varphi$ and finally at $\langle K\rangle K\varphi$ using D and KB, in that order. For the other direction, we start by assuming $\langle K\rangle K\varphi$. We can derive $\langle K\rangle B\varphi$ using KB and the fact that K is a normal modality. One application of NI then gives us $B\varphi$. □

With this in hand we can return to the argument for the normality of B. For NEC, assuming that φ is a theorem, NEC for K gives us that $K\varphi$ is also a theorem. From there, one application of the T axiom for K and modus ponens entails that $\langle K\rangle K\varphi$, and hence that $B\varphi$ is a theorem too. Now, it is well known that in the presence of NEC the K axiom is provably equivalent to distribution over conjunction:

$$B(\varphi \wedge \psi) \leftrightarrow B\varphi \wedge B\psi \qquad \text{(Dist-}\wedge\text{)}$$

The left-to-right implication follows straightforwardly from the fact that K is normal. The right-to-left direction is also well known and was already noted by Stalnaker. Here, however, we present a proof of it that is, to our knowledge, new and, more importantly, that does not make use of positive introspection for knowledge.

Observation 3. $\mathsf{S} \vdash B\varphi \wedge B\psi \rightarrow B(\varphi \wedge \psi)$

Proof. All the steps use normality of K. We first show that:

$$B\varphi \wedge B\psi \rightarrow B\left(B\varphi \wedge B\psi\right) \qquad (1)$$

1.	$\langle K\rangle K\varphi \wedge \langle K\rangle K\psi$	Assumption
2.	$\langle K\rangle K\langle K\rangle K\varphi \wedge \langle K\rangle K\langle K\rangle K\psi$	From 1. by 4 for B.
3.	$\langle K\rangle KK\langle K\rangle K\varphi \wedge \langle K\rangle KK\langle K\rangle K\psi$	From 2. by SB.
4.	$K\langle K\rangle K\langle K\rangle K\varphi \wedge K\langle K\rangle K\langle K\rangle K\psi$	From 3. by D for B
5.	$\langle K\rangle K(\langle K\rangle K\langle K\rangle K\varphi \wedge \langle K\rangle K\langle K\rangle K\psi)$	From 4. by Norm. of $K+T$
6.	$\langle K\rangle K(K\langle K\rangle\langle K\rangle K\varphi \wedge K\langle K\rangle\langle K\rangle K\psi)$	From 5. by D for B
7.	$\langle K\rangle K(K\langle K\rangle K\varphi \wedge K\langle K\rangle K\psi)$	From 6. by SB
8.	$\langle K\rangle K(\langle K\rangle K\varphi \wedge \langle K\rangle K\psi)$	From 7. by T

And then we show that:

$$B\left(B\varphi \wedge B\psi\right) \rightarrow B(\varphi \wedge \psi). \qquad (2)$$

1.	$\langle K\rangle K(\langle K\rangle K\varphi \wedge \langle K\rangle K\psi)$	Assumption
2.	$\langle K\rangle K(\langle K\rangle KK\varphi \wedge \langle K\rangle KK\psi)$	From 1 by SB
3.	$\langle K\rangle K(K\langle K\rangle K\varphi \wedge \langle K\rangle KK\psi)$	From 2 by D.
4.	$\langle K\rangle K\langle K\rangle(\langle K\rangle K\varphi \wedge KK\psi)$	From 3 by Normality of K
5.	$\langle K\rangle K\langle K\rangle\langle K\rangle(K\varphi \wedge K\psi)$	From 4 by Normality of K
6.	$\langle K\rangle K\langle K\rangle\langle K\rangle K(\varphi \wedge \psi)$	From 5 by Normality of K
7.	$K\langle K\rangle\langle K\rangle\langle K\rangle K(\varphi \wedge \psi)$	From 6 by D
8.	$K\langle K\rangle K(\varphi \wedge \psi)$	From 7 by SB twice
9.	$\langle K\rangle K(\varphi \wedge \psi)$	From 8 by T

□

□

This finishes the proof of Observation 1. Again, the notable feature of this proof is that it nowhere uses positive introspection for knowledge. So, even if knowledge is not introspective, which in Stalnaker's system boils down for it to be a KT modality, the logic of belief in S is still at least KD45. Again, this will be important later. For now, however, we can show more, namely that the logic of belief in Stalnaker's system is *exactly* KD45.

Observation 4. *For all $\varphi \in \mathcal{L}_B$:*

$$\mathsf{S} \vdash \varphi \text{ if and only if } KD45 \vdash \varphi$$

Proof. The right-to-left direction is Observation 1. For the left-to-right direction, assume that $KD45 \nvdash \varphi$. We have to show that $\mathsf{S} \nvdash \varphi$. Since $KD45 \nvdash \varphi$, there is a maximally KD45-consistent set $\Gamma \subseteq \mathcal{L}_B$ with $\neg\varphi \in \Gamma$. We shall show that Γ is in fact already a S-consistent set. To this end, define $K\psi$ as $B\psi \wedge \psi$. We have to show that Γ together with this newly defined K operator is S consistent, i.e. that it is S4 and satisfies the interaction axioms of Table 2. Normality of K follows from normality of B. The T axiom for K as well as Stalnaker's KB follow immediately from the definition of K. Further, note that $B\psi$ implies $BB\psi$ by KD45 for B and these two together imply $B(B\psi \wedge \psi)$, using normality once again. But this exactly means that $B\psi \to BK\psi$, which is the SB axiom. Further, by the same derivation, we find that $B\psi \wedge \psi$ implies $B(B\psi \wedge \psi) \wedge B\psi \wedge \psi$. But this means exactly that $K\psi \to KK\psi$, i.e. the 4-axiom. Further, using positive introspection, we obtain $B\psi \to BB\psi \wedge B\psi$. The consequent is by definition equivalent to $KB\psi$, thus proving PI. Finally, using 5 for B, we find that $\neg B\psi \to B\neg B\psi \wedge \neg B\psi$. Again the consequent is by definition equivalent to $K\neg B\psi$, proving NI. Thus, Γ is also a maximally S-consistent set. Since $\neg\varphi \in \mathsf{S}$, this implies that $\mathsf{S} \nvdash \varphi$. □

Furthermore, three more minor facts are worth noting regarding Stalnaker's system. First, PI is *redundant* in that system. Write S_{-PI} for Stalnaker's system minus PI.

Observation 5. *For all formulas $\varphi \in \mathcal{L}$,*

$$\mathsf{S} \vdash \varphi \text{ if and only if } \mathsf{S}_{-PI} \vdash \varphi$$

Proof. We have to show that $\mathsf{S}_{-PI} \vdash PI$. First, we note that the proof of $\mathsf{S} \vdash$ (EQ) did not rely on PI. Thus, $\mathsf{S}_{-PI} \vdash EQ$. Having this, we can show PI. From $B\varphi$, that is from $\langle K\rangle K\varphi$ by (EQ), one application of SB gives $\langle K\rangle KK\varphi$, and one application of D gives $K\langle K\rangle K\varphi$, which is just $KB\varphi$. □

Furthermore, it should be clear that Stalnaker's system is not a conservative extension of S4 for the knowledge modality. A simple illustration of that is that in the presence of (EQ), D for belief translates into the .2 axiom for knowledge,

$$\langle K\rangle K\varphi \to K\langle K\rangle\varphi$$

which is of course not a theorem of S4 alone.

On the other hand, if we augment S4 with .2 and the equivalence (EQ), then we retrieve Stalnaker's system S. More precisely: Define $S4.2 + EQ$ as the logic S4.2 for knowledge augmented with (EQ), here taken as an axiom. Observe that the latter is the only principle for belief in this logic.

Observation 6. *(cf. [15]) For all formulas φ:*

$$\mathsf{S} \vdash \varphi \text{ if and only if } S4.2 + EQ \vdash \varphi$$

Again, the proof of this result will be important later, this time because it does make use of positive introspection for knowledge.

Proof. We have already shown that $S4.2 + EQ \subseteq \mathsf{S}$, since the latter contains S4 and derives (EQ), from which we obtain .2 for K. For the converse, it suffices to show that the axioms of Table 2 are theorems of $S4.2+EQ$. Again, with (EQ), D is just .2 under another guise, and KB is a direct consequence of T for knowledge. SB is derived by one application of 4 to $\langle K\rangle K\varphi$. As shown above, PI follows from that by one application of .2. We get NI in contrapositive by one application of 4 to $\langle K\rangle\langle K\rangle K\varphi$. □

The proof of Observation 4 might seem to suggest that, in Stalnaker's system, knowledge is the same as true belief. This does not hold true in general.

Observation 7.

$$\mathsf{S} \nvdash K\varphi \leftrightarrow \varphi \wedge B\varphi$$

Proof. The left-to-right implication always holds, as it follows from KB and the T axiom. We show that the reverse direction does not hold in general by means of a counterexample. Figure 1 displays a model for a $S4.2$ knowledge relation, i.e. the relation is reflexive, transitive and satisfies the Church-Rosser property. By defining $B\varphi$ as $\langle K\rangle K\varphi$, this model becomes a S-model (cf. Observation 6), i.e. a model of Stalnaker's original axioms as shown in tables 1 and 2 in which $\varphi \wedge B\varphi \to K\varphi$ is not valid. In that model $M, w_3 \vDash Kp$. Hence $M, w_1 \vDash \langle K\rangle Kp$ and thus $M, w_1 \vDash p \wedge Bp$ holds. Yet, $M, w_1 \nvDash Kp$. □

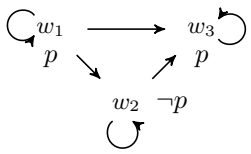


Figure 1: An $KT.2$ model in which $Kp \leftrightarrow p \wedge Bp$ fails, when B is defined through (EQ).

Let us take stock. We have now two ways to build a logic of belief on top of S4 for knowledge. The first is by Stalnaker's axioms in Table 2, and the resulting logic is S. The second is by adding .2 for knowledge and (EQ), to get S4.2+EQ. The two are provably equivalent, see Observation 6. So we are in fact dealing with one logical system, whose belief fragment we otherwise know to be exactly KD45.

1.2 Stalnaker's axioms with non-introspective knowledge

Let us now go back to the main question of this paper: to pinpoint the consequences, for the logic of belief, of omitting 4 for knowledge. Let us formulate this more precisely. Let S_{-4} be exactly as S except that the logic of knowledge is KT instead of S4. Define KT.2+EQ analogously. The question we are asking, then, is what is the logic of belief in S_{-4} and KT.2+EQ?

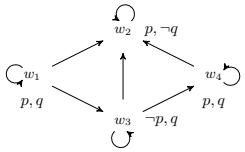
The answer for S_{-4} is already at hand. The proof of Observation 1 shows that K, NEC, 4 and 5 for B are all still derivable in S_{-4}, and so is (EQ). This means:

Observation 8. *For all $\varphi \in \mathcal{L}_B$:*

$$\mathsf{S}_{-4} \vdash \varphi \text{ if and only if } KD45 \vdash \varphi$$

Proof. We have just argued that $KD45 \vdash \varphi$ implies $\mathsf{S}_{-4} \vdash \varphi$. Since S_{-4} is a fragment of S, this implies in turn $\mathsf{S} \vdash \varphi$. But we know by Observation 4 that the latter happens if and only if $KD45 \vdash \varphi$. □

So we already know the answer to our first question, namely what happens to the logic of belief when we omit positive introspection for knowledge in Stalnaker's system? To put it bluntly, the answer is: nothing. Omitting positive introspection for knowledge in Stalnaker's leaves the logic of belief intact. Of course, this is not the case for knowledge. S_{-4} is a non-conservative extension of KT, and $K\varphi \to KK\varphi$ is not valid in that system. This can be illustrated by the following model of S_{-4} where knowledge is not introspective.



So while omitting 4 from S yields a genuinely different logic of knowledge, this does not affect the belief fragment. Is that also the case for KT.2+EQ? That 4 is used three times in the proof of Observation 6 suggests that the answer is no. This is indeed the case, as we show in the next section.

2 KT.2 and belief as epistemic possibility of knowledge

In this section we will turn to the logic of belief when it is defined as the epistemic possibility of knowledge, as in (EQ). We start with an epistemic logic where K is a KT.2 modality. The only axiom for the belief modality is (EQ). In other words, all and only the logical properties of belief are those inherited from its reduction to $\langle K \rangle K$. The question we ask can be then reformulated as follows: what is the sound and complete logic of the $\langle K \rangle K$ fragment of KT.2?

We have seen in the previous section that if knowledge is S4.2 then this logic of belief is equivalent to the one resulting from Stalnaker's axiom, and that it is completely axiomatized by KD45. This equivalence fails if the logic of knowledge is weakened to KT.2. This is what we show first, by arguing that

the resulting logic of belief is not normal. For this we will use semantic tools, which we introduce in Section 2.1. We then move to a complete axiomatization of the logic of belief, in Section 2.2.

2.1 Kripke, Neighborhoods, and MUD

Knowledge is interpreted in standard Kripke frames where the epistemic accessibility relation is reflexive and satisfies the so-called Church-Rosser property:

Definition 1. *An* **epistemic frame** $\mathcal{F}$ *is a pair* $\langle W, R\rangle$ *where* W *is a set of states and* R *is a reflexive, binary relation satisfying, for all* $w, x, y \in W$*:*

- *(Church-Rosser) If* wRx *and* wRy *then there is a* z *such that* xRz *and* yRz*.*

An **epistemic model** M *is an epistemic frame together with a valuation* V *assigning subsets of* W *to each atomic proposition in Prop.*

The truth condition for epistemic formulas thus becomes:

$$M, w \vDash K\varphi \quad \text{iff} \quad \text{If } wRv \text{ then } M, v \vDash \varphi$$

It is well known that the logic KT.2 is sound and complete with respect to the class of epistemic frames. Now define $B\varphi$ as $\langle K\rangle K\varphi$. This belief operator is not normal. The right-to-left direction of distribution under conjunction for beliefs, that is Dist-$\wedge$ on page 4, fails. Figure 2 illustrates this with a simple counter-example. This model displays the epistemic relation R. At w_1 we have both $\langle K\rangle Kp$ and $\langle K\rangle Kq$ but not $\langle K\rangle K(p \wedge q)$, thus belief is not normal.

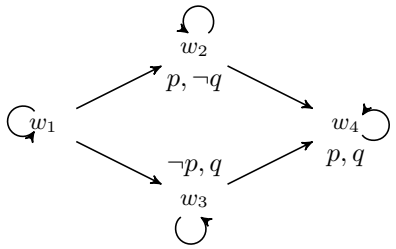


Figure 2: A KT.2 model in which the corresponding belief is not closed under conjunction

This belief operator otherwise validates necessitation and the left-to-right direction of Dist-$\wedge$. The latter can be encapsulated using the standard regularity rule [4, 16].

$$\frac{\vdash \varphi}{\vdash B\varphi}\ \text{NEC} \qquad \frac{\vdash \varphi \to \psi}{\vdash B\varphi \to B\psi}\ \text{REG}$$

Observation 9. *NEC, REG and D for* B *are sound with respect to KT.2 models where* B *is defined as* $\langle K\rangle K$*.*

NEC, REG and D are, however, not complete for the belief fragment. To show this, we move from Kripke to neighborhood frames.[2]

[2]See [16] for some background on neighborhood frames.

Definition 2. *A* MUD **neighborhood frame** $\mathcal{F}_N$*, or* MUD *frame for short, is a pair* $\langle W, n\rangle$*, where* W *is a set of possible worlds and* $n : W \to \mathcal{P}(\mathcal{P}(W))$ *is a neighborhood function, satisfying the following conditions:*

If $X \in n(w)$ *and* $X \subseteq Y$ *then* $Y \in n(w)$.	*(Monotonicity)*
$W \in n(w)$.	*(contains the Unit)*
If $X \in n(w)$ *then for all* $Y \in n(w)$, $X \cap Y \neq \emptyset$.	*(D)*

A MUD **model** M *is a neighborhood frame augmented with a valuation function* V*, as above.*

MUD models are used to interpret the belief operator only. Given a MUD model M, we write $\|\varphi\|_M$ for the truth set of φ in M, that is the set $\{w : M, w \vDash \varphi\}$. The truth condition for B is the standard one for neighborhood structures.

$$M, w \vDash B\varphi \text{ iff } \|\varphi\| \in n(w)$$

It is again well known that NEC, REG and D are together sound and complete with respect to the class of MUD frames. This logic, however, is sound but not complete for the belief fragment of KT.2. Let the formula (NBM), standing for No Belief in Moore sentence, be defined as follows:

$$\langle B\rangle(p \to Bp) \tag{NBM}$$

We will offer some interpretation of (NBM) soon. First, we show that (NBM) differentiates the two systems KT.2 and NEC, REG, and D.

Observation 10.

$$\vdash_{KT.2} \text{(NBM)}$$
$$\nvdash_{NEC,REG,D} \text{(NBM)}$$

Proof. For proving the first claim, start with the following theorem of KT: $Kp \to \langle K\rangle\langle K\rangle Kp$. This is equivalent to $\langle K\rangle\neg p \vee \langle K\rangle\langle K\rangle Kp$. K being normal, the latter is in turn equivalent to $\langle K\rangle(p \to \langle K\rangle Kp)$. One application of necessitation gives us the required formula, using (EQ): $K\langle K\rangle(p \to \langle K\rangle Kp)$. It is easy to construct a counter-model to the validity of that formula in MUD frames. □

We now return to the interpretation of (NBM). This formula is equivalent to

$$\neg B(p \wedge \neg Bp)$$

and the formula $p \wedge \neg Bp$ in the scope of the outermost belief operator is, of course, the classical "Moore sentence." The logic KT.2 thus precludes the agent from believing such a sentence about herself. This is a second-order coherence condition on belief in KT.2. Observe, however, that the stronger conditions of positive and negative introspection both fail for belief in this logic. Counter-examples are easy to construct. So, even though beliefs are not fully transparent, they are nonetheless subject to higher-order coherence constraints such as (NBM). NEC, REG and D, however, capture the full structure of first-order belief, i.e. a non-embedded belief fragment, as we will show at the end of the next section. These axioms and rules are sound *and complete* for the class of $\mathcal{L}_B$ formulas of modal depth at most one. But the full belief fragment is more demanding, as we shall see presently.

2.2 Beyond MUD: completeness

The formula (NBM) does not yet suffice to completely characterize the belief fragment of KT. 2 + EQ. We need to strengthen this condition to interact with other beliefs held by the agent. This strengthening turns out to be a set of non-trivial, infinitary constraints that lives at the frontier of expressive power for our language. So before we give a precise definition we present some intuitions guiding the construction.

The condition we need is encapsulated syntactically by the following sequence of formulas using infinite conjunction over $\mathcal{L}_B$. We will show this in Lemma 2 below.

$$B^0\psi := B\psi$$
$$B^{n+1}\psi := B^n\psi \wedge \bigwedge_{\varphi \in \mathcal{L}_B} \langle B\rangle (\varphi \to \psi \wedge B^n\varphi).$$

Note that $B^1\top = B\top \wedge \bigwedge_{\varphi \in \mathcal{L}_B} \langle B\rangle(\varphi \to B\varphi)$. The first conjunct is a theorem of KT. 2 + EQ. The second conjunct is exactly the condition expressed by taking all instances of (NBM). Since $B^1\psi \to B^1\top$, the $B^1\psi$ are thus a stronger version of the "no Moore sentence" condition.

$B^1\psi$ can best be interpreted in terms of a stability condition. While the usual $\varphi \wedge \neg B\varphi$ states that φ is true, yet the agent does not believe it, the condition $\varphi \wedge \neg (\psi \wedge B\varphi)$ states that φ is true, yet whenever ψ holds, the agent does not believe that φ. This condition thus relativizes the classical Moore sentence to those cases where ψ is true. With this in mind, $B^1\psi$ can be read as stating that the agent believes ψ and even if we deleted all $\neg\psi$ worlds, the agent would still not believe any Moore sentence about herself. In other words, the agent's belief in ψ is fully compatible with her believing a no-Moore condition.

The formulas $B^i\psi$ then are higher-order equivalents of this ψ no-Moore condition. In fact, the formulas $B^n\psi$ form an increasing hierarchy of conditions; that is, we have $B^n\psi \to B^{n-1}\psi$ for all n. As it will turn out, the $B^i\top$ are central for characterizing the belief fragment of $KT.2 + EQ$. We will see below that validity of the $B^i\top$ is necessary and sufficient for representing any MUD-models as KT.2 Kripke models (Lemma 2). We thus define:

Definition 3. *A* MUD$^\infty$ ***model*** *is an* MUD *neighborhood model in which all* $B^i\top$ *for* $i \in \omega$ *are valid.*

Fortunately, though, in order to completely axiomatize the belief fragment of KT. 2 + EQ we do not need to write down all the $B^i\top$ explicitly. We can work with finite approximations. This is what we show now.

To begin with, let us define sets of sets of $\mathcal{L}_B$ formulas X_i^ψ for $i \geq 1$ and $\psi \in L$. The construction is by induction over i. For the base case $i = 1$ and $\psi \in \mathcal{L}_B$, let

$$\overline{X}_1^\psi = \{B\psi \wedge \langle B\rangle (\varphi \to \psi \wedge B\varphi) \mid \varphi \in \mathcal{L}_B\}.$$

We then define X_1^ψ as the set of all *finite* conjunctions of formulas from $\overline{X}_1^\psi$. Similarly, for $i > 1$, we define

$$\overline{X}_i^\psi = \{\rho \wedge \langle B\rangle (\varphi \to \psi \wedge \chi) \mid \varphi \in \mathcal{L}_B, \rho \in X_{i-1}^\psi, \chi \in X_{i-1}^\varphi\}.$$

with again letting X_i^ψ be the set of finite conjunctions of $\overline{X}_i^\psi$. So the X_i^ψ are finitary approximations of $B^i\psi$ and we thus have $B^i\psi \vDash \chi$ for any $\chi \in X_i^\psi$.

A central role in our axiomatization is played by the $X_i^\top$. In fact, it is sufficient to look at these $X_i^\top$, i.e. the finitary counterparts of the formulas of the form $B^i\top$, which are in turn the classical, non-relativized no-Moore condition and their higher order variant. These $B^i\top$ are all sound with respect to the belief fragment of KT. 2 + EQ, and so are the $X_i^\top$. In fact, within KT. 2 + EQ, the entire hierarchy of X_i^ψ is situated between knowledge and belief. This is in fact true for infinitary formulas $B^i\varphi$ as well. However, showing this would require us to enter the proof theory of infinitary languages in more detail. So we leave this aside for now.

Observation 11. *For all $i \geq 1$ and $\chi \in X_i^\psi$:*

$$K\psi \vdash_{KT.2+EQ} \chi$$
$$X_{i+1}^\psi \vdash_{KT.2+EQ} \chi$$
$$X_i^\psi \vdash_{KT.2+EQ} B\psi$$

Proof. The only thing to be shown is the first implication. We prove this by induction over i. We start with $i = 1$. Assume $K\psi$. We thus have to derive any formula χ in X_1^ψ. Without loss of generality, it suffices to limit our attention to $\chi \in \overline{X}_1^\psi$. By (EQ), we thus have to show, for an arbitrary $\varphi \in \mathcal{L}_B$, that

$$B\psi \wedge K\langle K\rangle\,(\varphi \to \psi \wedge B\varphi)$$

First, we note that the first conjunct holds. Indeed, applying the T axiom, we have $K\psi \to \langle K\rangle K\psi$, i.e. $K\psi \to B\psi$, taking care of the first conjunct. It thus remains to show that $K\psi$ implies the second conjunct. We show the contrapositive. Assume that $\neg K\langle K\rangle\,(\varphi \to \psi \wedge B\varphi)$, i.e. $\langle K\rangle K\,(\varphi \wedge (\neg\psi \vee \neg B\varphi))$. Using T and the normality of K, this implies that $\langle K\rangle\,(K\varphi \wedge (\neg\psi \vee \neg B\varphi))$, which is equivalent to $\langle K\rangle(K\varphi \wedge \neg\psi) \vee \langle K\rangle(K\varphi \wedge \neg B\varphi)$. The second disjunct, $\langle K\rangle(K\varphi \wedge \neg B\varphi)$, is inconsistent with $K\varphi \to B\varphi$, which we have derived above. Thus the first disjunct $\langle K\rangle(K\varphi \wedge \neg\psi)$ is true. But this implies $\langle K\rangle\neg\psi$, and hence $\neg K\psi$. This contradicts our induction assumption of $K\psi$, thus deriving the desired contradiction. The proof of the induction step from i to $i + 1$ is similar to the above proof, with all $B\psi$ replaced by $\chi \in X_i^\psi$ and $B\varphi$ replaced by $\chi' \in X_i^\varphi$. □

Corollary 1. $\vdash_{KT.2+EQ} \chi$ *for all* $\chi \in X_i^\top$ *and* $i \in \omega$.

The axiom and rules D, NEC and REG as well as all elements of $X_i^\top$ for all i are thus sound with respect to the belief fragment of KT. 2 + EQ. In the following, we will denote this infinite set of axioms by MUD $+X$. Is MUD $+X$ also complete with respect to the belief fragment of KT. 2 + EQ? Yes.

Theorem 1. *The system* MUD $+X$ *is a sound and complete axiomatization for the belief part of* KT. 2 + EQ

In order to show completeness, we will need two auxiliary results. The first one shows that counter-models in MUD$^\infty$ frames can be constructed for each non-theorem of MUD $+X$. The second is the key representation theorem, showing that any MUD$^\infty$-model can be turned into a KT.2 model. These two together give us completeness.

Lemma 1. *Let φ such that $\nvdash_{\mathrm{MUD}+X} \varphi$. Then there is a* MUD^∞ *model in which* $M \nVdash \varphi$

Proof. We start with constructing a canonical model $M = \langle W, n, V\rangle$ of $\mathrm{MUD}+X$. As usual, the set of worlds W of M is the set of all maximally $\mathrm{MUD}+X$ consistent subsets of $\mathcal{L}_B$. First, we note that W is not empty, as $\mathrm{MUD}+X$ is sound with respect to the class of all KT.2 models, as shown in Corollary 1. Define the neighborhood function of M as $n(w) = \{\{v|\varphi \in v\}|B\varphi \in w\}$ and close all neighborhoods under supersets. As usual, the atomic valuation is given by $w \in V(p)$ iff $p \in w$. We have to show that

$$M, w \vDash \varphi \text{ iff } \varphi \in w$$

We do so by induction over the complexity of φ. The atomic and Boolean case work as usual. We only show the case where φ is of the form $B\psi$. For the right-to-left direction, assume that $\varphi = B\psi \in w$. By construction, $\{v|\psi \in v\} \in n(w)$. By induction hypothesis, we have that $\{v|\psi \in v\} = \{v|M, v \vDash \psi\}$, thus also the latter is in $n(w)$. Therefore $M, w \vDash B\psi$, which completes the proof. For the reverse direction, assume that $\varphi = B\psi \notin w$. Again, by induction hypothesis it suffices to show that $\{v|M, v \vDash \psi\} \notin n(w)$. We thus need to show for all $Y \in n(w)$ that there is some $y \in Y$ with $M, y \vDash \neg\psi$. Let an arbitrary $Y \in n(w)$ be given. By construction of the $n(w)$, there is some $B\chi \in w$ with $Y \supseteq \{v|\chi \in v\}$. First, we show that $\nvdash_{\mathrm{MUD}+X} \chi \to \psi$: If not, $\vdash_{\mathrm{MUD}+X} \chi \to \psi$ holds, which together with $B\chi \in w$ and REG implies that $B\psi \in w$, contradicting the assumption that $B\psi \notin w$. Thus $\nvdash_{\mathrm{MUD}+X} \chi \to \psi$, and hence there is some maximally $\mathrm{MUD}+X$ consistent set y with $\chi \in y$ and $\psi \notin y$. In particular $y \in Y$, but, by induction hypothesis, $M, y \nVdash \psi$, finishing the proof that $\{v|M, v \vDash \psi\} \notin n(w)$.

Now we proceed by showing that M is not just a $\mathrm{MUD}+X$, but also a MUD^∞-model, i.e. that all $B^i\top$ are valid on M. By construction, all $X_i^\top$ are valid on M. We show by induction on i that $M, w \vDash B^i\psi$ iff $M, w \vDash X_i^\psi$ for all $\psi \in \mathcal{L}_B$. The left to right direction is automatic. For the right to left direction the basic case, $i = 1$, is also straightforward. So assume now that $M, w \vDash X_i^\psi$ for some $i > 1$. To show $M, w \vDash B^i\psi$ we have to show that $i)$ $M, w \vDash B^{i-1}\psi$ and $ii)$ $M, w \vDash \bigwedge_{\chi \in \mathcal{L}_B} \langle B\rangle \left(\chi \to \psi \wedge B^{i-1}\chi\right)$. By construction of X_i^ψ we have that $M, w \vDash X_i^\psi$ implies $M, w \vDash \rho$ for all $\rho \in X_{i-1}^\psi$. By our induction assumption we get $M, w \vDash B^{i-1}\psi$, proving $i)$. For $ii)$, note that $M, w \vDash \langle B\rangle \left(\varphi \to \psi \wedge B^{i-1}\varphi\right)$ holds automatically whenever $B\varphi \notin w$. Thus, it suffices to show that this formula also holds for all $\varphi \in \mathcal{L}_B$ with $B\varphi \in w$. Let such φ be given. By assumption, we have for all $\rho \in X_{i-1}^\varphi$ that $M, w \vDash \langle B\rangle (\varphi \to \psi \wedge \rho)$.

So for every $\rho \in X_{i-1}^\varphi$, there is a maximally consistent set v with $\varphi, \psi, \rho \in v$. In particular, since X_{i-1}^φ is closed under finite conjunctions we have that for all $\rho_1, \ldots \rho_n \in X_{i-1}^\varphi$ there is a $v \in W$ with $\varphi, \psi \in v$ and $\rho_1, \ldots \rho_n \in v$. But this is just to say that every finite subset $\varphi, \psi, \rho_1, \ldots, \rho_n$ of $\{\varphi, \psi\} \cup X_{i-1}^\varphi$ is $\mathrm{MUD}+X$-consistent. Since $\vdash_{\mathrm{MUD}+X}$ is compact, this implies that $\{\varphi, \psi\} \cup X_{i-1}^\varphi$ is consistent as well. So there must be a world $v \in W$ with $M, v \vDash \varphi \wedge \psi \wedge X_{i-1}^\varphi$. By induction assumption, this implies that $M, v \vDash \varphi \wedge \psi \wedge B^{i-1}\varphi$ as desired and thus $M, w \vDash \langle B\rangle \left(\varphi \to \psi \wedge B^{i-1}\varphi\right)$, finishing the proof of $ii)$.

□

Lemma 2. *Let M be a* MUD *model.*

1. *Let $d \in \omega$: If $B^{2d}\top$ is valid in M, then for each $w \in M$, there is a KT.2 model N, v such that M, w and N, v agree on all belief formulas up to modal depth d.*

2. *If all $B^i\top$ are valid in M, then for each $w \in M$, there is a KT.2 model N, v such that M, w and N, v satisfy the same formulas in $\mathcal{L}_B$, i.e. they are modally equivalent for the belief fragment.*

Proof. Let $M = \langle W^M, n, V\rangle$ be a MUD model. We start by inductively defining a sequence of neighborhood functions $n^i : W^M \to \mathcal{PP}(W^M)$. Let $\bar{n}(w)$ denote the upward closure of $\{\|\varphi\| \mid \|\varphi\| \in n(w) \text{ for some } \varphi\}$. Note that $\bar{n}(w) \subseteq n(w)$. We define:

$$\begin{aligned} n^0(w) :=& n(w) \\ n^{i+1}(w) :=& \{X \in n^i(w) | \forall Y \in \bar{n}(w)\ \exists r \in X \cap Y : Y \in n^i(r)\} \end{aligned}$$

Since the $n(w)$ are upward closed, the $n^i(w)$ are as well. Also, since $n^i(w) \subseteq n(w)$, $X \cap Y \neq \emptyset$ for all $X, Y \in n^i(w)$ (since the same holds true for $n(w)$). Furthermore we show that the neighborhoods n^i interpret the operators B^i in the following sense: For all $v \in W^M$ and all $\varphi \in \mathcal{L}_B$ holds that

$$M, v \vDash B^i\varphi \text{ iff } \|\varphi\| \in n^i(v)$$

We show this by an induction over i. For $i = 0$, the claim is obviously true. Now assume the claim holds for $i-1$, we show that it holds for i. First, assume that $B^i\varphi$ holds at v. We have to show that $\|\varphi\| \in n^i(v)$. By definition of $\bar{n}(v)$, it suffices to show for all $\psi \in \mathcal{L}_B$ with $\|\psi\| \in n(v)$ that there is some $x \in \|\varphi\| \cap \|\psi\|$ with $\|\psi\| \in n^{i-1}(x)$. By definition, our assumption $B^i\varphi$ implies that $\bigwedge_{\psi \in \mathcal{L}_B} \langle B\rangle(\psi \to \varphi \wedge B^{i-1}\psi)$. Thus, whenever $\|\psi\| \in n(v)$, there is some $x \in \|\psi\|$ with $M, x \vDash \varphi \wedge B^{i-1}\psi$. In particular, $x \in \|\psi\| \cap \|\varphi\|$ and by induction hypothesis also $\|\psi\| \in n^{i-1}(x)$, since $M, w \vDash B^{i-1}\psi$. This finishes the proof of the first direction. For the reverse direction, assume that $\|\varphi\| \in n^i(v)$. We have to show that $M, v \vDash B^{i-1}\varphi \wedge \bigwedge_{\psi \in \mathcal{L}_B} \langle B\rangle(\psi \to \varphi \wedge B^{i-1}\psi)$. First, since the $n^i(v) \subseteq n^{i-1}(v)$, we have that $\|\varphi\| \in n^{i-1}(v)$ and thus by induction $M, w \vDash B^{i-1}\varphi$. Now let $\psi \in \mathcal{L}_B$. Since the $n(w)$ interpreting B are upward closed, it suffices to show that $X := \|\psi \wedge \neg(\varphi \wedge B^{i-1}\psi)\| \notin n(v)$. Assume for a contradiction that $X \in n(v)$. Thus $X \in \bar{n}(v)$ and, since $\|\varphi\| \in n^i(v)$, there is some $x \in \|\varphi\| \cap X$ with $X \in n^{i-1}(x)$. Since $\|\psi\| \supseteq X$, this implies that $\|\psi\| \in n^{i-1}(x)$. But by induction, this implies that $M, x \vDash \varphi \wedge B^{i-1}\psi$, contradicting the assumption that $x \in X$.

Now we can start proving 1). Assume, that $B^{2d}\top$ is valid in M (i.e. $M, v \vDash B^{2d}\top$ for all $v \in W^M$). Since $B^{2d}\top \to B^i\top$ for all $i \leq 2d$, we have that all neighborhoods $n^i(v)$ for $v \in W^M$ and $i \leq 2d$ are non-empty. We now construct an epistemic model N, v. We start by constructing the set of worlds W^N of N. The set W^N will be divided into different layers, $L_0, \ldots, L_{2d+1}$, which are constructed inductively. This construction is pictured in Figure 3.

Each world $v \in L_i$ will be indexed with a pair $\{x, X\}$ with $x \in W^M$ and $X \subseteq W^M$ such that $X \in n^{2d-i}(x)$. The inductive construction of the L_i is as follows: For the first layer, L_0, pick any $X \in n^{2d}(w)$. Layer L_0, then contains a single world $v_{w,X}$. Now, assume that L_{i-1} is already constructed for some

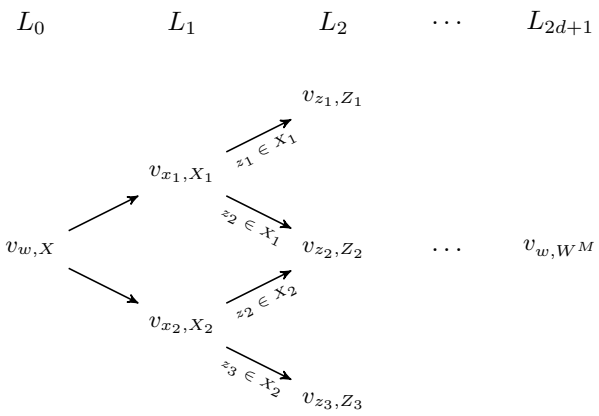


Figure 3: The construction of N as a layered model (all reflexive arrows missing)

$i < 2d$. To construct L_i, we execute the following steps for each $v_{x,X}$ in L_{i-1}. By assumption, $X \in n^{2d-(i-1)}(x)$. Thus, we have that for every $Y \in \bar{n}(x)$, there is $r \in X \cap Y$ with $Y \in n^{2d-i}(r)$. For every $Y \in \bar{n}(x)$ pick such r and add a world $v_{r,Y}$ to L_i. Further, add a world v_{y,W^M} for all $y \in X$ to L_i. This concludes the inductive construction up to L_{2d}. Finally, the last layer L_{2d+1} again contains a single world labeled v_{w,W^M}. Now, we define a valuation V^N and an accessibility relation R on the set of worlds $L_0 \cup \ldots \cup L_{2d+1}$, turning it into a KT.2 model. The valuation V^N is given by $v_{x,X} \in V^N(p)$ iff $x \in V(p)$. As for the accessibility relation, we connect every $v_{x,X}$ in L_i to itself and to every $v_{y,Y}$ in L_{i+1} with $y \in X$. We also connect every world in L_{2d} to the unique element of L_{2d+1}. Finally, the unique element of L_0 will be the designated world v.

It remains to show that N, v is indeed an epistemic model and that $M, w \vDash \varphi \Leftrightarrow N, v \vDash \varphi$ for all φ of modal depth at most d. We start by showing that N, v is an epistemic model. The relation R is reflexive by construction. For the Church-Rosser property: Let $v_{x,X} \in L_i$ and $v_{x,X}Rv_{y,Y}, v_{x,X}Rv_{z,Z}$. Wlog, $v_{y,Y}, v_{z,Z} \in L_{i+1}$. By construction, $Y, Z \in n(x)$ and thus $Y \cap Z \neq \emptyset$, since M satisfies D. Thus, there is some $r \in Y \cap Z$. By construction, $v_{r,W^M} \in L_{i+2}$ with $v_{y,Y}Rv_{r,W^M}$ and $v_{z,Z}Rv_{r,W^M}$.

Next, we show for every $v_{x,X}$ in L_i and all φ of modal depth at most $\frac{d-i}{2}$ that $N, v_{x,X} \vDash \varphi$ iff $M, x \vDash \varphi$. We show this by induction over the modal depth of φ. For atomic formulas, this holds true by the definition of the valuation V^N. We only give the proof for φ of the form $B\psi$. We start with the right to left direction. Assume, that $M, x \vDash \varphi$. Thus, by definition, $Y = \|\psi\|^M \in n(x)$. By construction of N, there is some $y \in Y$ and some $v_{y,Y}$ in L_{i+1} with $v_{x,X}Rv_{y,Y}$. Again by construction, all worlds $v_{z,Z}$ accessible from $v_{y,Y}$ satisfy $z \in Y$ and are all in L_{i+1} and L_{i+2}. Thus, by the induction hypothesis, all these $v_{z,Z}$ satisfy $N, v_{z,Z} \vDash \psi$. Thus, $N, v_{y,Y} \vDash K\psi$ and therefore $N, v_{x,X} \vDash \langle K\rangle K\psi (= B\psi)$, as desired. For the reverse direction, assume that $N, v_{x,X} \vDash B\psi$. Since $B\psi \leftrightarrow \langle K\rangle K\psi$, there is some $v_{y,Y}$ with $v_{x,X}Rv_{y,Y}$ and for all $v_{z,Z}$ with $v_{y,Y}Rv_{z,Z}$ holds $N, v_{z,Z} \vDash \psi$. By the induction hypothesis and the fact that all $v_{z,Z}$ accessible from $v_{y,Y}$ are in L_{i+1} or L_{i+2}, we again have $M, z \vDash \psi$ for all these $v_{z,Z}$. By construction of the L_i, we have $Y = \{z | \exists v_{z,Z} : v_{y,Y}Rv_{z,Z}\} \in n(x)$. Thus, using

the fact that $n(x)$ is upwards closed, we have that $\|\psi\|^M \in n(x)$, completing the proof.

The proof of 2) is similar to the proof of 1). We only sketch the argument. This time, the set of worlds of N consists of an infinite hierarchy of layers $L_0, L_1, \ldots$. The individual worlds $v \in W^N$ will be labeled with triples $v_{x,X,n}$, where $x \in W^M$, $X \in n(x)$ and $n \in \omega$. To construct layer L_0, we pick some $X \in \bigcap n^i(w)$ (which is non-empty, as $W^M \in n^i(w)$ for all i) and set $L_0 = \{v_{w,X,0}\}$. Second, to construct L_1, for every $Y \in \bar{n}(w)$ and every $j \in \omega$ pick some $r \in X$ with $Y \in n^j(r)$. Add a new world to L_1 with label $x_{v,Y,j}$. Further add $v_{r,W^M,j}$ for every $r \in X$ and $j \in \omega$ to L_1. From there on, the inductive procedure goes as usual: For $i > 0$ assume that L_i is already constructed. For every $v_{x,X,j} \in L_i$ with $j > 1$ and every $Y \in \bar{n}(x)$ pick some $y \in X \cap Y$ with $Y \in n^{j-1}(y)$ and add $v_{y,Y,j-1}$ to L_{i+1}. Also add $v_{x,W^M,l}$ for all $x \in X$ and $l \in \omega$ to L_{i+1}. The valuation V^N is defined as above. The relation R is defined as follows: Relate every $x \in W^N$ to itself and relate the unique element of L_0 to all elements of L_1. For $i > 0$, relate $v_{x,X,j} \in L_i$ with $j > 0$ to all $v_{y,Y,j-1} \in L_{i+1}$ with $y \in X$. Also relate $v_{x,X,j} \in L_i$ to all $v_{y,W^M,l}$ for $l \geq j-1$ and $y \in X$. Then the proof that N is an epistemic model proceeds as above.

We still have to show that this new pointed model is modally equivalent to M, w. We do so in two steps. First, we observe that for every $v_{x,Y,n}$ in W^N that is not the root v, and every formula φ of modal depth at most $\frac{n}{2}$, we have that $M, x \vDash \varphi$ iff $N, v_{x,Y,n} \vDash \varphi$. The proof is basically the same as in 1). Now, we can finally show that M, w is modally equivalent to N, v, i.e. that $M, w \vDash \varphi$ iff $N, v \vDash \varphi$. We do so by induction over the complexity of φ. We only show the case, where φ is of the form $B\psi$. First, assume that $M, w \vDash B\psi$. Thus, there is some $Y \in \bar{n}(w)$ with $M, r \vDash \psi$ for all $r \in Y$. Since $Y \in \bar{n}(w)$, there is some $r \in X \cap Y$ and some $v_{r,Y,j}$ in L_1, such that $j > 2 \cdot md(\psi)$, where $md(\cdot)$ denotes the modal depth. Thus, all $x \in R[v_{r,Y,j}](= \{y | v_{r,Y,j} R y\})$ are of the form $v_{s,Z,m}$ for some $s \in Y$ and $m \geq j-1$. By the previous argument, these worlds all satisfy that $N, v_{s,Z,m} \vDash \psi$. Thus $N, v \vDash \langle K \rangle K\psi = B\psi$.

Now, assume $M, w \nvDash B\psi$. Thus, $M, w \vDash \langle B \rangle \neg\psi$. We thus have to show, that both $R[v]$ and every $R[z]$ for $z \in L_1$ have a non-empty intersection with $\|\neg\psi\|$. First, we show this for the element $R[v]$. Since $M, w \nvDash B\psi$, there is some $r \in X$ with $M, r \vDash \neg\psi$. By construction, there is thus some $v_{r,W^M,j}$ in L_1 with $j > 2 \cdot md(\psi)$. By the previous argument, we then have that $N, v_{r,Y,j} \vDash \neg\psi$, showing that $N, v \nvDash K\psi$. We now show that all $R[z]$ for z in L_1 have a non-empty intersection with $\|\neg\psi\|$. Each such z is labelled by $v_{y,Y,l}$ for some $Y \in n(w)$. Again, since $M, w \vDash \langle B \rangle \neg\psi$, there is some $r \in Y$ such that $M, r \vDash \neg\psi$. By construction, there is a world $v_{r,W^M,j}$ in L_2 with $j > 2md(\psi)$ and $z R v_{r,W^M,j}$. Again by the previous argument, $N, v_{z,W^M,j} \vDash \neg\psi$. Therefore $N, z \nvDash K\psi$ thus $N, v \nvDash \langle K \rangle K\psi$ finishing the proof.

□

Before proceding to the proof of Theorem 1, we illustrate the construction in the first part of the above lemma with an example.

Example 1. *Let the* MUD$^\infty$ *neighborhood model* $M = \langle W, n, V \rangle$ *be given by the set of worlds* $W = \{a, b, c\}$*, the valuation* $V(p) = \{a\}, V(q) = \{b\}$ *and*

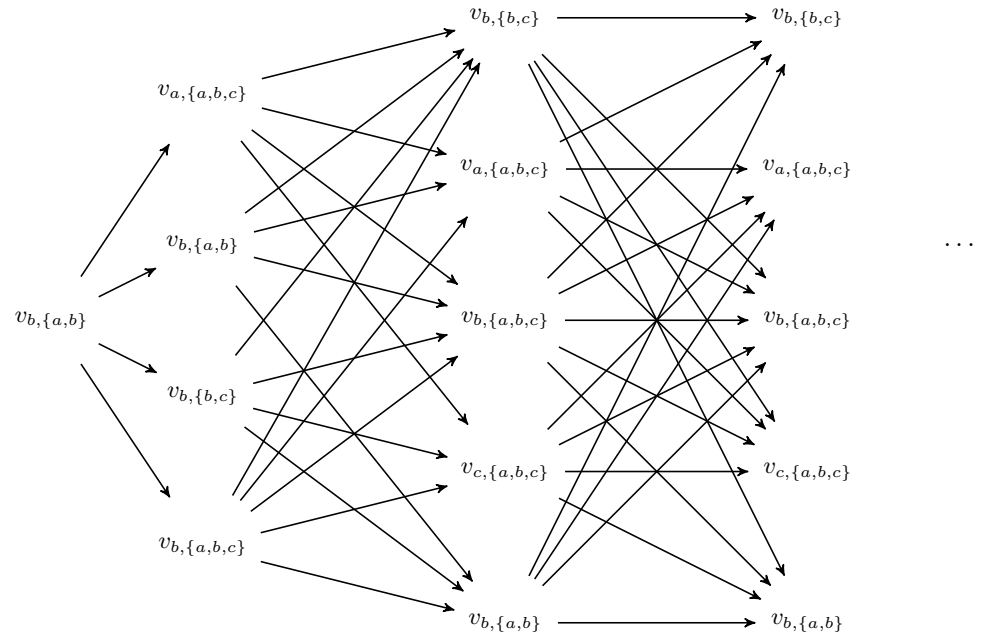


Figure 4: The first four layers $L_0, \ldots, L_3$ constructed in Example 1.

$V(r) = \{c\}$ *and the neighborhood function* n*:*

$$\begin{aligned} n(a) =& \{\{b,c\},\{a,b,c\}\} \\ n(b) =& \{\{a,b\},\{b,c\},\{a,b,c\}\} \\ n(c) =& \{\{a,b\},\{a,b,c\}\} \end{aligned}$$

First, we note that for all $x \in W$ *holds*

$$\bar{n}(x) = n(x) = n^1(x) = n^2(x) = \ldots$$

In particular, all $B^i\top$ *for* $i \in \omega$ *are valid on* M*. By the first part of Lemma 2, there is a KT.2 model* N, v *such that* M, b *and* N, v *agree on all formula up to modal depth, say,* 5*. Figure 4 shows the first four levels* $L_0, \ldots, L_3$ *of such a* N, v*, constructed as in the proof of Lemma 2, part 1). For the initial level, we start with the choice* $v_{b,\{a,b\}}$ *as the unique world in* L_0*. The valuation on* $L_0, \ldots, L_3$ *is given by* $V(p) = \{v_{a,X} | X \subseteq W\}, V(q) = \{v_{b,X} | X \subseteq W\}$ *and* $V(r) = \{v_{c,X} | X \subseteq W\}$*.*

Now, we can finally prove the soundness and completeness theorem

Proof of Theorem 1. Soundness was shown in Observation 9 and Corollary 1 above. Now, we show the completeness of our axioms. Let φ be a belief formula

that is valid on all epistemic frames. We have to show that $\vdash_{\mathrm{MUD}+X} \varphi$. Assume the contrary, i.e. $\nvdash_{\mathrm{MUD}+X} \varphi$ and therefore by Lemma 1 there is a MUD$^\infty$ model M on which φ is not valid. Thus, there is some w with $M, w \vDash \neg\varphi$. By Lemma 2, there is then an epistemic model N, v with $N, v \vDash \neg\varphi$. But this contradicts the assumption that φ is valid on the class of epistemic frames. □

As already alluded to at the end of Section 2.1, a slight variation of the previous construction shows, that the axioms NEC, REG and D are sound and complete with respect to the first-order fragment of the belief language.

Remark 1. *Let $\varphi \in \mathcal{L}_B$ of modal depth at most 1. Then $\vdash_{NEC,REG,D} \varphi$ iff $\vdash_{KT.2+EQ} \varphi$.*

Proof. Soundness, i.e. the left-to-right direction, was shown in Observation 9. For completeness, i.e. the right-to-left direction, it suffices to show that whenever there is a MUD model M, x with $M, x \vDash \varphi$, there is also an epistemic model $N, v \vDash \varphi$. Starting with a MUD model M, x, we construct N, v similar to the last proof. The set of worlds of N will again consist of layers $L_0, \dots L_3$ and each $v \in W^N$ will be labelled with a pair x, X with $x \in W^M$ and $X \subseteq W^M$. The first and last layers L_0 and L_3 consist of a singleton with label v_{w,W^M}. The second layer, L_2, consists of a world v_{v,W^M} for every $v \in W$. Finally, to construct layer L_1, add a world v_{v,W^M} for every $v \in W$. Furthermore, for every $X \in n(w)$ pick some $x \in X$ and add a world with label $v_{x,X}$. Define the relation R such that vRv for every $v \in W^N$. Additionally, the unique world in L_0 is related to every element of L_1 and every element of L_2 is related to the unique element of L_3. Finally, an element $v_{x,X}$ of L_1 is related to $v_{y,Y} \in L_2$ iff $y \in X$. Defining the valuation again as $v_{x,X} \in V^N(p) \Leftrightarrow x \in V^M(p)$, we find that N is an epistemic model and $N, v \vDash \varphi$, where v is the unique element of L_0. □

2.3 Beyond MUD: Model theory

We now take a closer look at the model theory of KT. 2 + EQ, and in particular at its relationship to epistemic models. To start with, we relate epistemic models to MUD neighborhood models.

Definition 4. *Let $M = \langle W, R, V\rangle$ be an epistemic model. The* **corresponding belief model** *$M' = \langle W', n, V'\rangle$ is defined as $W' = W, V' = V$ and $X \in n(w)$ iff there is $v \in W$ with wRv such that $X \supseteq \{z | vRz\}$.*

The construction of the belief neighborhoods is the semantic counterpart of the syntactic (EQ) definition of belief as $B\varphi \leftrightarrow \langle K\rangle K\varphi$. It immediately follows that

Observation 12. *For an epistemic model M, w and all $\varphi \in \mathcal{L}_B$:*

$$M, w \vDash \varphi \Leftrightarrow M', w \vDash \varphi$$

Thus, every epistemic model, representing the knowledge and uncertainty of an agent, can be translated into a corresponding neighborhood model, picturing the beliefs of that same agent. But what about the converse: Given an MUD neighborhood model M, w in which all $B^i\top$ are valid, can we retrieve the agent's knowledge? Or, if not, can we at least find *some* epistemic model N such that

M is (equivalent to) the derived epistemic model of N? As it turns out, the answer to both these questions is negative. In the remainder of this section we will show why.

First, we need to define when two models are equivalent. As usual, we will spell this out in terms of bisimulations. For neighborhood models, these are defined as follows (cf. [16]).

Definition 5. *Let M and N be monotonic neighborhood models and $Z \subseteq M \times N$ a relation. Then Z is a* **bisimulation** *iff, whenever wZv:*

- *(Atomic Harmony) For each $p \in At$ holds $w \in V^M(p) \Leftrightarrow v \in V^N(p)$*
- *(Zig) For all $X \in n(w)$ there is some $X' \in n(v)$ with $\forall x' \in X' \exists x \in X$ such that xZx'.*
- *(Zag) For all $X' \in n(v)$ there is some $X \in n(w)$ with $\forall x \in X \exists x' \in X'$ such that xZx'.*

We call Z **bitotal** *iff for every $x \in M$ there is some $y \in N$ with xZy and vice versa.*

We recall that bisimilarity is, in general, a stricter notion than logical equivalence. In the case of Kripke frames, bisimilar models are logically equivalent, but the converse need not hold, see [3, p.69]. The same holds true for neighborhood models, see [16]. The question we ask here is: Is every pointed MUD$^\infty$ model M, w bisimilar to a derived neighborhood model N', v of an epistemic model?

We start by giving a sufficient condition for when a pointed MUD$^\infty$ model M, w is bitotally bisimilar to the derived model N', v of an epistemic model. To do so, we extend the sequence of neighborhoods n^i defined in the proof of Lemma 2 transfinitely. Recall, that[3]

$$\begin{aligned} n^0(w) :=& n(w) \\ n^{i+1}(w) :=& \{X \in n^i(w) \mid \forall Y \in \bar{n}(w) \exists r \in X \cap Y : Y \in n^i(r)\} \end{aligned}$$

Since the n^i form a decreasing sequence (i.e. $n(w) \supseteq n^1(w) \supseteq n^2(w) \ldots$), a natural way of defining n^ω is to take the intersection of all $n^i(w)$ for $i < \omega$.

$$n^\omega(w) := \bigcap_{i<\omega} n^i(w).$$

We can continue this construction arbitrarily, thus getting $n^\alpha(w)$ for every ordinal α. Since the $n^\alpha(w)$ form a $\subseteq$-decreasing sequence, they must eventually reach a fixed point. Hence, there is some minimal ordinal α_0 such that $n^{\alpha_0}(w) = n^{\alpha_0+1}(w)$ for all $w \in W$. This n^{α_0} will help define the necessary and sufficient condition for the existence of a bitotal bisimulation between M, w and the derived model N', v of some epistemic model N, v. In the proof of Lemma 2,

[3]Note that through referring to $\bar{n}(w)$, rather than $n(w)$, the construction of the $n^i(w)$ does depend upon the valuation V. In particular, there is no straightforward lifting of the construction in Lemma 2 from models to frames. We are not aware of any viable definition of MUD$^\infty$ neighborhood *frames*, i.e. neighborhood frames that result in a MUD$^\infty$ model when equipped with *any* valuation.

we showed that the $n^i(w)$ interpret the $B^i\varphi$, i.e. $M, w \vDash B^i\varphi \Leftrightarrow \|\varphi\| \in n^i(w)$. For a MUD$^\infty$ model M, we thus have $n^i(w) \neq \emptyset$ for all $i < \omega$. The conditions of the following theorem are even stronger than this.

Theorem 2. *Let M, w be a pointed* MUD$^\infty$ *model. Then M, w is bitotally bisimilar to the derived model N', v of an epistemic model if and only if $n^{\alpha_0}(w') \neq \emptyset$ for all $w' \in M$*

Proof. We start by showing that $n^{\alpha_0}(w') \neq \emptyset$ for all $w' \in M$ is a sufficient condition for being bitotally bisimilar to some derived model N', v of an epistemic model. Assume that $n^{\alpha_0}(w') \neq \emptyset$ for all w'. Then we can create an epistemic model N, v, using a similar construction as in the proof of Lemma 2 (1). This time, there are infinitely many levels $L_0, L_1, \ldots$ and instead of picking the $v_{x,X}$ such that $X \in n^i(x)$, we pick them such that $X \in n^{\alpha_0}(x)$. Then, the relation Z defined by $xZv_{y,Y}$ iff $x = y$ is a bitotal bisimulation.

We now show that $n^{\alpha_0}(w') \neq \emptyset$ is also a necessary condition for being bitotally bisimilar to some derived model N', v of an epistemic model. We proceed in two steps. First, we show that in every KT. $2 +$ EQ model N, all $n^{\alpha_0}(w)$ are non-empty. In fact, $R[v] \in n^{\alpha_0}(w)$, where R is the relation corresponding to the K operator. As a second step, we show that if Z is a bisimulation between belief models M and N, then it is also a bisimulation between the belief models M^i and N^i, where the neighborhood functions n_M and n_N have been replaced by n^i_M and n^i_N, respectively. In particular, if xZy then $n^i_N(x)$ is empty iff $n^i_M(y)$ is empty.

We start with the first step. Let N', v be the derived model of an epistemic frame. We show by transfinite induction that

$$R[w] = \{v \mid wRv\} \in n^\alpha(w)$$

for all ordinals α (where R is the accessibility relation of the epistemic frame N). For $n^0(w) = n(w)$, this holds true since R is reflexive. Now, assume $\alpha = \beta + 1$ is a successor ordinal and $R[v] \in n^\beta(v)$ for all v. We have to show that $R[v] \in n^\alpha(v)$ for all v. Thus, we have to show that for all $Y \in \bar{n}(v)$, there is some $r \in Y \cap R[v]$ with $Y \in n^\beta(r)$. Let such Y be given. By construction of the derived model, there is some s with vRs and $R[s] \subseteq Y$. In particular, $s \in Y \cap R[v]$ and, by assumption, $R[s] \in n^\beta(s)$. Since the n^β are monotonous, $R[s] \in n^\beta(s)$ implies $Y \in n^\beta(s)$, completing the proof. Finally, if α is a limit ordinal, $n^\alpha(w) = \bigcap_{i<\alpha} n^i(w)$. By induction, all these $n^i(w)$ contain $R[w]$, thus n^α does too.

Finally, as a second step, we show that if Z is a bisimulation between belief models M and N, then it is also a bisimulation between the belief models M^α and N^α, i.e. the models M and N where the neighborhood functions n_M and n_N are replaced by n^α_M and n^α_N. We do so by induction over α. Since $n^0(w) = n(w)$, the base case is trivial. Now, assume that $\alpha = \beta + 1$ and that Z is a bitotal bisimulation between M^β and N^β. We have to show that Z is a bisimulation between M^α and N^α. The atomic harmony condition is trivial. We only show the (zig) condition. The proof of (zag) is similar. Thus let wZv and $X \in n^\alpha(w)$. To prove (zig), it is sufficient to show that $Z[X] = \{v' | xZv' \text{ for some } x \in X\}$ is in $n^\alpha(v)$. Note that by induction assumption, $Z[X] \in n^\beta(v)$. So let $Y' \in \bar{n}(v)$. We have to show that there is some $r' \in Z[X] \cap Y'$ with $Y' \in n^\beta(r')$. By definition of $\bar{n}$, there is some $\|\varphi\| \in \bar{n}(v)$ with $\|\varphi\| \subseteq Y'$. Thus we can assume

without loss of generality that Y' is of the form $\|\varphi\|$. Since wZv, there is some $Y \in n(w)$ such that for all $y \in Y$ there is $y' \in Y'$ with yZy'. Since Z is a bisimulation, $M, y \vDash \varphi$ for every $y \in Y$. Thus $Y \subseteq \|\varphi\|$. In fact, since Z is bitotal, every $y \in \|\varphi\|^M$ has some y' with yZy'. Using monotonicity, we can thus assume that $Y = \|\varphi\| \in \bar{n}(w)$. Since $X \in n^\alpha(w)$, there is some $r \in X \cap Y$ with $Y \in n^\beta(r)$. Since Z is bitotal, there is some r' with rZr'. Since Z is a bisimulation, we get $r' \in Y' \cap Z[X]$. The only thing that remains to show is that $Y' \in n^\beta(r')$. By assumption, Z is a bitotal bisimulation between M^β and N^β. Since $Y \in n^\beta(r)$, there is some $Y'' \in n^\beta(r')$ such that for all $z'' \in Y''$ there is $z \in Y$ with zZz''. Thus, since Z is a bisimulation, $Y'' \subseteq \|\varphi\|$ and thus by monotonicity $Y' = \|\varphi\| \in n^\beta(r')$. Finally, we show the claim for the case of α is a limit ordinal. Assume that $X \in n^\alpha_M(w)$. Again, it suffices to show that $Z[X] = \{v' | xZv' \text{ for some } x \in X\} \in n^\alpha_N(v)$. Assume not. Then there is some $\beta < \alpha$ with $X \notin n^\beta_N(v)$. Thus, since n^β_N is upward closed, there is no $Z \in n^\beta_N(v)$ such that for all $z \in Z$ $\exists x \in X$ with xZz. But this, together with the fact that $X \in n^\beta_M(w)$ (since $n^\beta_M(w) \supseteq n^\alpha_m(w)$), contradicts the induction assumption that Z is a bisimulation between M^β and N^β. □

As mentioned above, the condition that all $B^i\top$ are valid in a MUD model M is equivalent to stating that $n^i(w) \neq \emptyset$ for all $i \in \omega$. For infinite models, this condition is weaker than demanding that $n^{\alpha_0}(w) \neq \emptyset$. Since the former condition guarantees that M is modally equivalent to a $KT.2$ model (by Lemma 2 (2)), we conjecture:

Conjecture 1. *There is an infinite pointed* MUD^∞ *model* M, w*, that is modally equivalent to some KT.2 model, but not bitotally bisimilar to the derived model* N', v *of any epistemic model.*

Finally, we end with a quick note on the relationship between knowledge and belief. Within $\mathrm{KT.2 + EQ}$ logic, belief is defined through knowledge via the equivalence (EQ): $B\varphi \leftrightarrow \langle K \rangle K\varphi$. Thus, given an epistemic frame, we can read off the agents beliefs. But what about the converse? Assume we are given the corresponding derived belief model N', v of an epistemic model N, v. Can we retrieve the agents knowledge from N', v? As it turns out, the answer to this is again negative:

Example 2. *There are two epistemic models* M, v *and* N, x *such that the derived belief models* M', v *and* N', x *are bisimilar, while there is some formula* $K\psi$ *with* $M, v \vDash K\psi$ *and* $N, x \vDash \neg K\psi$*.*

Proof. Let the neighborhood model O have as set of worlds $W^O = \{w_1, w_2, w_3\}$. The neighborhood function is constant (i.e. $n(w_1) = n(w_2) = n(w_3)$) and given by $n = \{\{w_1, w_2\}, \{w_1, w_3\}, \{w_1, w_2, w_3\}\}$. Finally $V(p) = \{w_1, w_2\}$. It is easy to check that $\alpha_0 = 0$, i.e. $n(w) = n^1(w) = n^2(w) \ldots$ for all w. Thus, by Theorem 2, there is an epistemic model M, v such that the corresponding belief model is bisimilar to O, w_1. Furthermore, since $\{w_1, w_2\} \in n^{\alpha_0}(w)$, we can execute the construction from the proof of Theorem 2 in such a way, that the unique element of L_0 is labelled with $v_{w_1,\{w_1,w_2\}}$. Thus, every $z \in M$ with vRz is labelled with w_1 or w_2, where R is the relation corresponding to the knowledge operator. Since $V(p) = \{w_1, w_2\}$, this implies that $M, v \vDash Kp$. Similarly, since $\{w_1, w_3\} \in n^{\alpha_0}(w)$, we can also construct N, x bisimilar to O, w_1 such that the unique element of L_0 is labelled with $v_{w_1,\{w_1,w_3\}}$.In this case, every $z \in N$ with

xRz is labelled with w_1 or w_3, and the latter label occurs at least once. Thus, $N, x \not\models Kp$. Since the derived belief models of M, v and N, x are both bisimilar to O, w_1, they are bisimilar to each other. □

3 No Lockean interpretation for belief

Having a sound and complete analysis of its logical structure, we now turn our attention to finding a suitable interpretation of the belief operator. Stalnaker, in his original paper, offers an understanding of the belief operator as "subjective certainty" on the side of the believing agent. This is in line with later probabilistic and Bayesian interpretations (see [6, 7, 11]), where a KD45 theory of belief has been linked to the notion of belief with probability[4] 1. In fact, KD45 is sound and complete with respect to belief with probability 1. So what about the belief operators introduced here? Do they lend themselves to a probabilistic interpretation? In the previous sections, we have introduced two new logics for knowledge and belief. The first of these, S_{-4} logic, combines a KT.2 notion of knowledge with the original Stalnaker axioms. In this logic, the resulting belief operator is still KD45 (Observation 8), thus we can maintain the interpretation of belief as probability 1 or, in Stalnaker's words, as subjective certainty.

Observation 13. *The belief part of S_{-4} logic is sound and complete with respect to belief with probability 1.*

The second logic we studied, combines a KT.2 knowledge operator with the identity $B\varphi \leftrightarrow \langle K\rangle K\varphi$. As shown above, this belief operator is no longer a KD45 operator anymore. In fact, it is not even normal, as $B\varphi \wedge B\psi \rightarrow B(\varphi \wedge \psi)$ is not valid. Thus, the resulting belief will not be complete with respect to belief of probability 1. One tempting interpretation of the belief operator under consideration, motivated by the Lockean thesis [13], is in terms of sufficiently high enough credence. That is, rather than demanding subjective certainty, a formula should be believed if its credence is above a given threshold $t > 1/2$, i.e.,

$$B\varphi \text{ iff } p(\varphi) \geq t > 1/2$$

for some given probability measure p. It is well known that such an operator would not be closed under intersection. Moreover, such an operator would validate all axioms and rules of the first-order part of

$$\text{MUD} + X$$

logic.

Observation 14. *NEC, REG and D are sound for B interpreted as "probability at least t", for any $t > \frac{1}{2}$.*

Proof. Take a probability measure over a σ-algebra and let it be the set $\mathcal{X}$ of measurable sets that have probability $> 1/2$. It is immediately clear that $\mathcal{X}$ satisfies NEC and REG. To see that $\mathcal{X}$ satisfies D, let X and Y in $\mathcal{X}$. Since $p(X)$ and $p(Y)$ are strictly greater than 0.5, we have $X \cap Y \neq \emptyset$, showing that D holds. □

[4] See [7] for combined knowledge-belief models with a probabilistic interpretation. There, belief in φ is interpreted as the set of φ worlds having probability one, while knowledge of φ means that all worlds in the knowledge cell are φ worlds

NEC, REG and D, however, are not complete with respect to that interpretation. There are MUD$^\infty$ models that cannot be equipped with a probability measure in such a way that the belief operator respects the equivalence above. We show in fact something stronger. Rather than focusing on a threshold of .5, we show that for any threshold ϵ there is some MUD$^\infty$ model that cannot be equipped with a probability measure in such a way that the agent only believes propositions with probability at least ϵ. Even stronger still, the following example will be such that there is some proposition of low probability, at most ϵ, that the agent believes as well as some proposition of high probability, at least $(1-\epsilon)$, that the agent fails to believe.

Example 3. *Assume a company advertises a new position. As it happens, n qualified candidates apply, so the company decides to make two hires. Assume now, that our agent just learned that two people will be hired, but he has not yet learned who. We will show that it is consistent with* MUD $+X$ *to believe of each candidate simultaneously that she will be hired. But, of course, no matter how subjective credence is attributed to the different possible hires, some candidate needs to receive an extremely low subjective probability. Let us fill in some details.*

The set of atomic propositions is $p_1, \dots p_n$, where p_i stands for candidate i getting hired. We assume that there are at least 2 candidates, i.e. $n \geq 2$. The model $M = (W, n, V)$ is then constructed as follows. The possible worlds are $W = \{w_{i,j} | 1 \leq i < j \leq n\}$. For the valuation, let $w_{i,j} \in V(p_k)$ if $k = i$ or $k = j$. Thus, world $w_{i,j}$ represents a situation in which agents i and j are hired. Finally, the neighborhood function n is constant (i.e. the same for all worlds) and given by the upward closure of $\{\|p_1\|, \dots \|p_n\|\}$. In particular, we have $M \vDash Bp_i$ for all i. Our agent believes of every candidate that they will be hired. Furthermore, M is a MUD$^\infty$ *model: The neighborhoods n are obviously monotonous and contain the unit. Since $w_{i,j} \in \|p_i\| \cap \|p_j\|$, also D holds, thus M is a* MUD *model. It is easy to check that $n^i(w) = n(w)$ for all $i \in \omega$, thus also $B^i\top$ hold for all i.*

So what about the subjective probabilities? As only two candidates are hired, not every $\|p_i\|$ can be assigned a high probability. If n is large enough, we are guaranteed to find some $\|p_i\|$ that receives a low probability, no matter how we choose to assign probabilities. To be somewhat more explicit about this argument, assume we want to find a probability function that makes each $\|p_i\|$ the agent believes as probable as possible. More specifically, we look for the probability distribution that maximizes $min_i prob(\|p_i\|)$; that is, we want to make the most improbable proposition that the agent still believes as probable as possible. It is not difficult to see that the probability distribution maximizing $min_i prob(\|p_i\|)$ assigns equal weight to all worlds. Since there are $\frac{n(n-1)}{2}$ many worlds, this probability distribution will assign a weight of $\frac{2}{n(n-1)}$ to every world w_{ij}. All $\|p_i\|$ have cardinality $n-1$, thus they each receive a weight of $\frac{2}{n}$. Thus, in every possible probability distribution with n candidates, there will be some i such that $\|p_i\|$ has a subjective probability of at most $\frac{2}{n}$. In particular, if n becomes large, the agent will believe some proposition p_i that is extremely implausible, i.e. one to which she assigns probability at most $\frac{2}{n}$. At the same time, she will not believe the proposition $\neg p_i$, even though it receives a credence of at least $\frac{n-2}{n}$.

The example shows that the notion of belief defined above is not sufficiently

strong to enforce the "belief with high enough credence" interpretation. Now, a natural question to ask is: What additional constraints on beliefs would be required? We leave that question open for future work. Instead, we now inquire into a suitable interpretation of the B^i operators defined in the last section.

In section 2.2 we have introduced a sequence of belief operators $B^i\varphi$, for all natural numbers i. These satisfy $B^j\varphi \to B^i\varphi$ for all $j > i$, thus imposing increasingly strict conditions on belief as i grows. In fact, each $B^j\varphi$ is defined by some (infinite) coherence conditions on the lower level B^i. In other words, the lower level B^i could be seen as an internal scaffolding, giving further structure to B^j. Notably, we have shown that $K\varphi \to B^i\varphi \to B\varphi$, i.e. B^i is situated somewhere between belief simpliciter and knowledge. So, what is a reasonable interpretation of the B^i? Could they, perhaps, allow for a more knowledge-like interpretation? Or at least for an interpretation of belief with high probability? No. Despite being stronger operators, the logical properties of the B^i are exactly the same as those of the B operators. In the following we denote by $\mathcal{L}_{B^i}$ the logical language, in which all B are replaced by B^i.

Fact 1. *The axioms REC, NEC and D are sound and complete with respect to the fragment of $\mathcal{L}_{B^i}$ consisting of formulas of modal depth at most 1.*

Proof. Soundness follows from the fact that REC, NEC and D are sound for B, together with the definition of the B^i. For the completeness part, we start with a general observation: It is easy to check that if M, w is a neighborhood model with constant neighborhood function, i.e. $n(v) = n(w)$ for all $v, w \in M$, then $n^i(v) = n(v)$ for all i and thus $M, w \vDash B^i\varphi \Leftrightarrow M, w \vDash B\varphi$. Now, assume that φ is a first-order belief formula with $\nvdash_{KT.2+EQ} \varphi$. Since MUD frames are complete with respect to first-order belief formulas, there is a MUD model M, w with $M, w \vDash \neg\varphi$. Since the truth of φ depends only on $n(w)$, we can assume that M has a constant neighborhood function. But by the above observation, this implies that $M, w \vDash \neg\varphi'$, where φ' is φ with all B replaced by B^i. □

4 Conclusion: Inter-Definability and Higher-Order Consistency

Omitting positive introspection for knowledge in Stalnaker's system brings with it a number of surprises for the logic of belief. On the one hand, the interaction axioms in Table 2 (page 2) are strong enough to keep the logic of belief unchanged, that is keep it to KD45, even in the absence of 4 for knowledge. This is not true if, instead of these interaction axioms, we take the definition of belief as the epistemic possibility of knowledge. Normality goes—belief no longer distributes over conjunction—together with introspection, either positive or negative. Not all introspective properties are lost, though. Agents, in that logic, never believe the Moore sentence about themselves. They never believe that something is the case but that they don't believe this. We have shown in fact that the agents are subjected to an infinite hierarchy of such anti-Moorean coherence constraints, and that this hierarchy completely axiomatizes this new logic of belief.

Compared to KD45, which as we saw in Section 3 is sound and complete with respect to the "probability 1" interpretation, the non-normal modal operator

resulting from defining belief with (EQ) when knowledge is KT.2 is harder to interpret. In Section 3 we ruled out the Lockean reading. Could we then go back to viewing this belief operator as a formulation of "absolute subjective certainty" or even as the "mental component" of knowledge? The example developed on page 22 shows that the absolute subjective certainty interpretation fails in the strongest way possible, at least if "certainty" is in any substantial way related to credences. The second reading goes back at least to Lenzen [14], who showed that in a slightly stronger doxastic-epistemic system, knowledge and belief become interdefinable; belief as the epistemic possibility of knowledge, and knowledge as true belief:

$$K\varphi \leftrightarrow \varphi \wedge B\varphi$$

This equivalence fails already if knowledge is only $S4$. An inspection of the proof of Observation 4, however, reveals that in Stalnaker's system the logic of "true belief" is at least S4.2. So in that logic belief can be seen as the mental component of *some* epistemic-like attitude, although *not* of the primitive knowledge operator. Unsurprisingly, though, even this weaker result fails in KT. 2 + EQ. The logic of true belief will validate NEC, REG and, trivially, T, but not the full K axiom nor any introspective principle. So this belief is not the "mental component" of knowledge either, at least not of a common form of knowledge.

The main philosophical output of this study of the belief fragment of KT. 2 + EQ is thus to turn the emphasis onto the higher-order consistency constraints that might bear on knowledge and belief, instead of adding yet another iteration to the debate on introspection. Mirroring that debate, though, an obvious question to ask is whether principles like "no belief in Moore sentence" (NBM) are prone to paradoxical consequences in cases of vagueness. A multi-agent extension of KT. 2 + EQ or MUD $+X$ also raises interesting questions. In contrast to introspective properties, which most multi-agent epistemic logics assume to be common knowledge, it seems natural to allow uncertainty regarding (NBM). There is nothing paradoxical about believing a Moore sentence concerning someone else. The multi-agent perspective also raises interesting technical questions, as notions such as common belief become more subtle both to define and to axiomatize in the absence of full distribution under conjunction [16]. Along the same line, an obvious next step is to develop a plausible theory of revision and update for weaker beliefs. Here the main challenge, this time echoing [1, 15], would be to see whether – if not in general then when – information dynamically preserves the higher-order consistency constraints of MUD $+X$. All in all, then, omitting introspection from Stalnakers original system has turned out to be a technically rewarding enterprise, opening up interesting philosophical avenues.

References

[1] Baltag, A., N. Bezhanishvili, A. Özgün, and S. Smets (2013). The topology of belief, belief revision and defeasible knowledge. In *Logic, Rationality, and Interaction*, pp. 27–40. Springer.

[2] Belnap, N. (1982). Display logic. *Journal of Philosophical Logic 11*, 375–417.

[3] Blackburn, P., M. De Rijke, and Y. Venema (2002). *Modal logic*, Volume 53. Cambridge University Press.

[4] Chellas, B. F. (1980). *Modal Logic: An Introduction*. Cambridge University Press.

[5] Ciabattoni, A., R. Ramanayake, and H. Wansing (2014). Hypersequent and display calculi a unified perspective. *Studia Logica*, 1–50.

[6] Fagin, R., J. Geanakoplos, J. Y. Halpern, and M. Y. Vardi (1999). The hierarchical approach to modeling knowledge and common knowledge. *International Journal of Game Theory 28*(3), 331–365.

[7] Galeazzi, P. and E. Lorini (2015). Epistemic logic meets epistemic game theory: a comparison between multi-agent kripke models and type spaces. *Synthese*, 1–31.

[8] Gentzen, G. (1934-35). Untersuchungen über das logische Schließen. *Mathematische Zeitschrift vol.39*, 176–210, 405–431.

[9] Hendricks, V. F. (2003). Active agents. *Journal of Logic, Language and Information 12*(4), 469–495.

[10] Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Cornell University Press.

[11] Klein, D. and E. Pacuit (2014). Changing types: Information dynamics for qualitative type spaces. *Studia Logica 102*(2), 297–319.

[12] Kracht, M. (1996). Power and weakness of the modal display calculus. In H. Wansing (Ed.), *Proof Theory of Modal Logic*, pp. 95–120. OUP.

[13] Leitgeb, H. (2014). The stability theory of belief. *Philosophical Review 123*(2), 131–171.

[14] Lenzen, W. (1979). Epistemologische betrachtungen zu [s4, s5]. *Erkenntnis 14*(1), 33–56.

[15] Özgün, A. (2013). Topological models for belief and belief revision. Master's thesis, Universiteit van Amsterdam.

[16] Pacuit, E. (2016). *Neighborhood Semantics for Modal Logic*. Forthcoming.

[17] Poggiolesi, F. (2011). Display calculi and other modal calculi: a comparison. *Synthese 173*, 259–279.

[18] Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical studies 128*(1), 169–199.

[19] Wansing, H. (1998). *Displaying modal logic*. Kluwer Academic Publishers.

[20] Wansing, H. (2002). Sequent systems for modal logics. In *Handbook of philosophical logic*, pp. 61–145. Springer.

[21] Williamson, T. (2000). *Knowledge and its Limits*. Oxford UP.

Appendix - Proof Theory for Stalnaker's System

In this appendix we present a version of Stalnaker's S in the setting of Display logic; which we shall call DS. This is interesting because this system is a bimodal logic and proof-theoretic investigations of such logics are still sparse.

Display logic has been designed by Belnap [2] to provide for a powerful syntactic framework which is also a generalization of Gentzen's sequent calculus [8]. Display logic allows for an elegant proof of cut elimination given that several conditions hold. These conditions are usually easy to verify. For this end the system contains not only formulas and sequents (also termed *consecutions*) but also structures. Due to its richness, however, there also downsides to Display logic, see e.g.[12]. For connections to other logical frameworks cf. [5, 17].

Wansing [19, 20] enriched Belnap's orignial work with an intensional marker $\bullet$, in order to allows for a smooth formalization of normal modal logics within. For our purposes we introduce two bullets, one corresponding to knowledge, $\bullet_K$ and one corresponding to belief, $\bullet_B$.

In what follows we present a concise version of DS. We start with definitions of formulas and structures; throughout we use standard terminology as eg. in [19].

Definition 6 (Formulas of DS)**.** $\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid K\varphi \mid B\varphi$

Definition 7 (Structures of DS)**.** $X := \mathrm{I} \mid \varphi \mid *X \mid X \circ Y \mid \bullet_K \mid \bullet_B$

The particular system DS consists of an axiom, structural and (propositional) logical rules, Display equivalence rules, and eventually rules for the introduction of the modalities K and B.

A presentation of the system DS

Axiom $A \Longrightarrow A$ $\quad$ A is atomic.

Structural Rules

$$\frac{X \Longrightarrow Y}{\mathrm{I} \circ X \Longrightarrow Y}\ (\mathrm{I}+) \qquad \frac{\mathrm{I} \circ X \Longrightarrow Y}{X \Longrightarrow Y}\ (\mathrm{I}-) \qquad \frac{\mathrm{I} \Longrightarrow Y}{X \Longrightarrow Y}\ (\mathrm{Il})$$

$$\frac{X \Longrightarrow \mathrm{I}}{X \Longrightarrow Y}\ (\mathrm{Ir}) \qquad \frac{X \circ Y \Longrightarrow Z}{Y \circ X \Longrightarrow Z}\ (\mathsf{P}) \qquad \frac{X \circ X \Longrightarrow Y}{X \Longrightarrow Y}\ (\mathsf{C})$$

$$\frac{X \circ (Y \circ Z) \Longrightarrow U}{(X \circ Y) \circ Z \Longrightarrow U}\ (\mathsf{A}) \qquad \frac{X \Longrightarrow A \qquad A \Longrightarrow Y}{X \Longrightarrow Y}\ (\mathsf{Cut})$$

Logical Rules (Propositional)

$$\frac{*\varphi \Longrightarrow X}{\neg\varphi \Longrightarrow X}\ (\neg\mathsf{l}) \qquad \frac{X \Longrightarrow *\varphi}{X \Longrightarrow \neg\varphi}\ (\neg\mathsf{r}) \qquad \frac{\varphi \circ \psi \Longrightarrow X}{\varphi \wedge \psi \Longrightarrow X}\ (\wedge\mathsf{l})$$

$$\frac{X \Longrightarrow \varphi \qquad Y \Longrightarrow \psi}{X \circ Y \Longrightarrow \varphi \wedge \psi}\ (\wedge\mathsf{r}) \qquad \frac{\varphi \Longrightarrow X \qquad \psi \Longrightarrow Y}{\varphi \vee \psi \Longrightarrow X \circ Y}\ (\vee\mathsf{l}) \qquad \frac{X \Longrightarrow \varphi \circ \psi}{X \Longrightarrow \varphi \vee \psi}\ (\vee\mathsf{r})$$

$$\frac{X \Longrightarrow \varphi \qquad \psi \Longrightarrow Y}{\varphi \to \psi \Longrightarrow *X \circ Y}\ (\to\mathsf{l}) \qquad \frac{X \circ \varphi \Longrightarrow \psi}{X \Longrightarrow \varphi \to \psi}\ (\to\mathsf{r})$$

Display Equivalence Rules (DE)

$$\frac{\dfrac{X \circ Z \Longrightarrow Y}{X \Longrightarrow Y \circ *Z}}{Z \Longrightarrow *X \circ Y} \qquad \frac{\dfrac{X \Longrightarrow Y}{*Y \Longrightarrow *X}}{X \Longrightarrow * * Y} \qquad \frac{\dfrac{X \Longrightarrow Y \circ Z}{X \circ *Z \Longrightarrow Y}}{*Y \circ X \Longrightarrow Z}$$

The rules for the introduction of K and B are structurally identical. So present them using $\Box$ for referring to either K or B and, do not index the intensional marker $\bullet$ with $_K$ or $_B$.

$$\frac{\varphi \Longrightarrow Y}{\Box\varphi \Longrightarrow \bullet Y}\,(\Box\mathsf{l}) \qquad \frac{\bullet X \Longrightarrow \varphi}{X \Longrightarrow \Box\varphi}\,(\Box\mathrm{r})$$

Additionally we need the structural rule (I$\bullet$) and the display equivalence rule ($\bullet$):

$$\frac{\mathrm{I} \Longrightarrow Y}{\bullet\mathrm{I} \Longrightarrow Y}\,(\mathrm{I}\bullet) \qquad \frac{\dfrac{\bullet X \Longrightarrow Y}{X \Longrightarrow \bullet Y}}{}\,(\bullet)$$

The four rules above capture the fact that both modalities K and B are normal operators (in the technical sense). The additional logical content of Stalnaker's S is expressed by the following structural rules. Due to the fact that the logical content is expressed by use of structural rules and applications of cut are warranted only for formulas no new cases arise for the general cut elimination procedure.

$$\frac{\bullet_B X \circ \bullet_B Y \Longrightarrow *\mathrm{I}}{X \Longrightarrow *Y}\,(\mathsf{d}) \qquad \frac{X \Longrightarrow \bullet_K Y}{X \Longrightarrow Y}\,(\mathsf{t})$$

$$\frac{X \Longrightarrow \bullet_K Y}{X \Longrightarrow \bullet_K \bullet_K Y}\,(4) \qquad \frac{\bullet_B X \Longrightarrow Y}{\bullet_B \bullet_K X \Longrightarrow Y}\,(\mathsf{sb})$$

$$\frac{\bullet_K X \Longrightarrow Y}{\bullet_B X \Longrightarrow Y}\,(\mathsf{kb}) \qquad \frac{\bullet_B X \Longrightarrow Y}{\bullet_K \bullet_B X \Longrightarrow Y}\,(\mathsf{pi})$$

$$\frac{\bullet_B X \Longrightarrow Y}{\bullet_B * \bullet_K * X \Longrightarrow Y}\,(\mathsf{ni})$$

This completes the presentation of DS. Here are three important derivations that illustrate how the system works.

$$\cfrac{\cfrac{\cfrac{\varphi \Longrightarrow \varphi}{K\varphi \Longrightarrow \bullet_K \varphi}}{\bullet_K K\varphi \Longrightarrow \varphi}}{\bullet_B K\varphi \Longrightarrow \varphi}{\Big/}\,K\varphi \Longrightarrow B\varphi \qquad \cfrac{\cfrac{\cfrac{\varphi \Longrightarrow \varphi}{B\varphi \Longrightarrow \bullet_B \varphi}}{\bullet_B B\varphi \Longrightarrow \varphi}}{\bullet_B \bullet_K B\varphi \Longrightarrow \varphi}{\Big/}\,\bullet_K B\varphi \Longrightarrow B\varphi \,/\, B\varphi \Longrightarrow KB\varphi$$

$$\cfrac{\cfrac{\cfrac{\cfrac{\cfrac{\cfrac{\cfrac{\varphi \Longrightarrow \varphi}{B\varphi \Longrightarrow \bullet_B \varphi}}{\bullet_B B\varphi \Longrightarrow \varphi}}{\bullet_B * \bullet_K * B\varphi \Longrightarrow \varphi}}{* \bullet_K *B\varphi \Longrightarrow B\varphi}}{*B\varphi \Longrightarrow \bullet_K * B\varphi}}{\neg B\varphi \Longrightarrow \bullet_K * B\varphi}}{\bullet_K \neg B\varphi \Longrightarrow *B\varphi}}{\bullet_K \neg B\varphi \Longrightarrow \neg B\varphi}}{\neg B\varphi \Longrightarrow K\neg B\varphi}$$

Fact 2. *The right counterparts of the structural rules P, C, and A are derivable.*

Fact 3. $\mathsf{DS} \vdash \varphi \Longrightarrow \varphi$ *for all* φ.

At the center of any Display logic is the Display theorem which is a main ingredient for establishing the cut elimination theorem. For this end we need two further definitions. It is not difficult to see that from these the Display theorem follows.

Definition 8 (Positive and Negative Occurrence)**.** *An occurrence of a substructure in a given structure is called* positive *if it is in the scope of an even number of* $*$ *(otherwise it is coined* negative*).*

Definition 9 (Antecedent and Succedent Parts)**.** *In a sequent* $Y \Longrightarrow Z$ *an occurrence of* X *is an* antecedent part *if it occurs positively in the antecedent or negatively in the succedent. An occurrence that is not an antecedent part is a* succedent part*.*

Theorem 3 (Display theorem)**.** *Each antecedent part of* X *of a sequent* S *can be displayed as the whole antecedent of a display-equivalent sequent* $X \Longrightarrow Y$*. Likewise, each consequent part of a sequent can be displayed as the whole succedent of a display-equivalent sequent.*

From the Display theorem together with the conditions (listed below) a general cut elimination result follows in a straightforward way. The conditions (C2)-(C8) guarantee cut elimination, whereas (C1) ensures the subformula property.

Definition 10. *(Belnap's conditions (C1)-(C8))*

(C1) Preservation of formulas*: Each formula occurring in a premise of a rule instance is a subformula of some formula in the conclusion (except Cut).*

(C2) Shape-likeness of parameters*: Congruent parameters are occurrences of the same structure.*

(C3) Non-proliferation of parameters*: Each parameter is congruent to at most one constituent in the conclusion; that is, no two constituents in the conclusion are congruent to each other.*

(C4) Position-likeness of parameters*: Congruent parameters are either all antecedent or all consequent parts in their respective sequence.*

(C5) Display of principal constituents*: If a formula is principal constituent in the conclusion of an inference, then it is either the entire antecedent or the entire consequent of the conclusion.*

(C6) Closure under substitution of consequent parts*: Each inference rule is closed under simultaneous substitution of arbitrary structures in consequent parts for congruent parameters.*

(C7) Closure under substitution of antecedent parts*: Each inference rule is closed under simultaneous substitution of arbitrary structures in antecedent parts for congruent parameters.*

(C8) Cut of matching principal constituents*: If there are inferences* Inf_1 *and* Inf_2 *with respective conclusions (1)* $X \Longrightarrow \varphi$ *and (2)* $\varphi \Longrightarrow Y$*, with* φ *principal in both inferences, then either (3)* $X \Longrightarrow Y$ *is identical to one of (1) or (2), or there is a derivation of (3) from the premises of* Inf_1 *and* Inf_2 *in which (Cut) is only used on proper subformulas of* φ*.*

The following two theorems are direct consequences of the above definition.

Theorem 4 (Cut elimination for DS)**.** *Cut is eliminable for* DS*.*

Let us pause to note that condition (C1) does not play a role in proving the eliminability of cut, i.e. conditions (C2) through (C8) are sufficient to prove the general cut elimination theorem. However, if all eight conditions do hold for a system, then it follows that the system possesses the subformula property meaning that each provable sequent has a proof where every formula occurring in any step of the derivation is a subformula of a formula in the conclusion.

Theorem 5 (Subformula property of DS)**.** *The display calculus* DS *without (Cut) has the subformula property.*

We can in fact prove that if (1) $\mathsf{S} \vdash \varphi$ then $\mathsf{DS} \vdash \mathrm{I} \Longrightarrow \varphi$ and furthermore that (2) if $\mathsf{DS} \vdash X \Longrightarrow Y$, then $\mathsf{D} \vdash \tau(X) \to \tau(Y)$. Part (2) needs an explicit treatment of the translation function τ, which is tedious but not particularly difficult. We omit it here.

Fact 4. DS *and* S *are deductively equivalent.*

This fact gives rise to soundness and completeness:

Fact 5. DS *is sound and complete with respect to the semantics of* S*.*