Computationele logica

Kamans, Jim
10302905

Roosingh, Sander
11983957

Schenk, Stefan
11881798

December 2017

Exercise 1

For this exercise, please use the semantic definitions (given in the slides of Lectures 4.1 and 4.2) of knowledge $K\varphi$ in plausibility models, belief $B\varphi$, conditional belief $B^{\varphi}\psi$ and safe belief $\Box\varphi$ (defined as the Kripke modality for the plausibility relation).

(a) *Show the following equivalence:*

$$K_a\varphi \Leftrightarrow B_a^{\neg\varphi} false$$

Proof. Let $\mathbf{S} = (S, \leq_a, \sim_a, ||\cdot||)$ be an arbitrary plausibility model and $s \in S$ also be arbitrary.
To prove our equivalence, we need to prove two implications:

Proof of left-to-right direction ($K_a\varphi \Rightarrow B_a^{\neg\varphi} false$)

We assume that $s \models K_a\varphi$. By the semantic definition of K_a, this means that $s(a) \subseteq ||\varphi||$ (1). By definition of $best_a$, $best_a(||\neg\varphi|| \cap s(a)) \subseteq (||\neg\varphi|| \cap s(a))$. From (1), it follows that $(||\neg\varphi|| \cap s(a)) = \emptyset \subseteq ||false||$. By the semantic definition of conditional belief, this means that $s \models B_a^{\neg\varphi} false$.

Proof of right-to-left direction ($B_a^{\neg\varphi} false \Rightarrow K_a\varphi$)

We assume that $s \models B_a^{\neg\varphi} false$. By the semantic definition of conditional belief, this means that $best_a(||\neg\varphi|| \cap s(a)) \subseteq (||\neg\varphi|| \cap s(a)) \subseteq ||false||$. $(||\neg\varphi|| \cap s(a)) \subseteq ||false||$ iff $s(a) \subseteq ||\varphi||$. By the semantic definition of K_a, this means that $s \models K_a\varphi$. □

(b) *Prove semantically the equivalence claimed on Slide 24 of Lecture Notes 4.2:*

$$B_a\varphi \Leftrightarrow \Diamond_a\Box_a\varphi$$

where $\Diamond_a\varphi := \neg\Box_a\neg\varphi$ is the dual modality to safe belief $\Box_a$.

Proof. Let $\mathbf{S} = (S, \leq_a, \sim_a, ||\cdot||)$ be an arbitrary plausibility model and $s \in S$ also be arbitrary.
To prove our equivalence, we need to prove two implications:

Proof of left-to-right direction ($B_a\varphi \Rightarrow \Diamond_a \Box_a \varphi$)

We assume $s \models B_a\varphi$. By the semantic definition of B_a, this means that $best_a s(a) \subseteq ||\varphi||$. Now we are going to show that the $best_a s(a) \subseteq ||\Diamond_a \Box_a \varphi||$. To do so, we consider **any arbitrary element** $t \in best_a s(a)$. By definition of $best_a$, we have $u \leq_a t$ for all $u \in s(a)$. Let $w \in s(a)$ be **any arbitrary state s.t.** $w \geq_a t$. Then, by transitivity of $\geq_a$, we must have $u \leq_a w$ for all $u \in s(a)$. So we have $w \in best_a s(a)$. Since $best_a s(a) \subseteq ||\varphi||$, it follows that $w \in ||\varphi||$, i.e. $w \models \varphi$. Because w is chosen arbitrary, it follows that $t \models \Box_a \varphi$. Then, by reflexivity of $\leq_a$, we must have $t \models \Diamond_a \Box_a \varphi$. Because t is chosen arbitrary, it follows that $best_a s(a) \subseteq ||\Diamond_a \Box_a \varphi||$ and thus $s \models \Diamond_a \Box_a \varphi$.

Proof of right-to-left direction ($\Diamond_a \Box_a \varphi \Rightarrow B_a\varphi$)

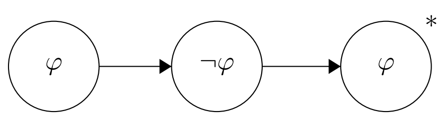
We assume $s \models \Diamond_a \Box_a \varphi$. By semantics of $\Diamond_a$ and $\Box_a$, this means that there is a t for which it holds that $t \geq_a s$ s.t. for all $u \geq_a t$: $u \models \varphi$. Then, by transitivity of $\geq_a$, it follows that $u \geq_a s$. By semantics of B_a, it follows that $s \models B_a\varphi$. $\Box$

(c) *Prove (via a counterexample) that* ***safe belief does NOT imply strong belief****; i.e. that*

$$\Box_a \varphi \not\Rightarrow Sb_a \varphi$$

Safe belief in some world s means that $t \models \varphi$ for all t such that $s \leq_a t$. Strong belief implies that for all worlds reachable from that world, the φ-worlds are strictly more plausible than the non-φ-worlds.

In the model below, the real world is denoted by *. In this world $\Box_a \varphi$ holds (there are no a-arrows pointing towards worlds in which φ is not true), but $Sb_a \varphi$ doesn't hold, because not all non-φ-worlds reachable from the real world are strictly more plausible than the φ-worlds reachable from the real world (the middle world is more plausible than the left world).



HINT: The positive statements (a) and (b) need general proofs: you need to show that, for *every* plausibility model and *every sentence* ϕ, the desired formula is true at *all* the worlds of that model. But the negative statement (c) must be shown by giving a counterexample: construct *some* plausibility model and find *some* sentence φ for which the implication fails to be true at *some* world of that model (which we can think of as the "real world").

Exercise 2

A virtual agent in a video game doesn't know his current position in the virtual space, but all he cares is (a) whether or not he's in a "Dangerous" zone (say, close to a dangerous monster), and (b) whether or not he's close to this Target (say, a treasure). Let's use the letter d to denote the sentence *the agent is in a dangerous zone*, and the letter t to denote the sentence *the agent is close to the target*. These possibilities are independent of each other, and the agent **doesn't know** which is the case, so he **cannot exclude any** of the four possible cases $d \wedge t$, $d \wedge \neg t$, $\neg d \wedge t$ and $\neg d \wedge \neg t$. However, our agent **believes both** that he's close to the target AND that he's NOT in a dangerous zone. **If** he would learn that this belief is WRONG (i.e. that at least one of his two beliefs is false), then he'd still believe (conditional on this new information) that he is close to the target. But **if** he would learn instead that he's far from (=NOT close to) the target, then (conditional on this information) he'd keep his initial belief that he's NOT in a dangerous zone.

1. *Write down a logical formula in the language of beliefs, knowledge and conditional beliefs to encode all the above assumptions.*

   $\neg K(t) \wedge \neg K(\neg t) \wedge \neg K(d) \wedge \neg K(\neg d) \wedge B(t \wedge \neg d) \wedge B^{\neg(t\wedge\neg d)}(t) \wedge B^{\neg t}(\neg d)$

2. *Represent the agent's beliefs (and conditional beliefs), using a* ***plausibility model*** *with four possible worlds. Specify the* ***valuation*** *(which atomic sentences of the two atomic sentences d and t are true at which worlds). Represent the agent's* ***plausibility relation*** *on these worlds, by drawing arrows going from the less plausible worlds to the more plausible ones.*

   $$\boxed{\neg t,\ d} \longrightarrow \boxed{\neg t,\ \neg d} \longrightarrow \boxed{t,\ d} \longrightarrow \boxed{t,\ \neg d}$$

3. *Suppose somebody who never lies tells our agent "You are close to the target if and only if you believe that you are in a dangerous zone."* ***Write down formally*** *this sentence as a formula φ in doxastic logic (using the atomic sentences).*

   $\varphi := t \leftrightarrow B(d)$

4. *Interpreting the above truthful announcement $!\varphi$ as an* ***update*** *with the sentence φ in the previous part, represent the* ***updated model***.

   $$\boxed{\neg t,\ d} \longrightarrow \boxed{\neg t,\ \neg d}$$

5. *After the previous announcement, another truthful announcement is made: "You are in a dangerous zone if and only if you don't believe that you are*

in a dangerous zone." ***Write down formally*** *this sentence as a formula ψ in doxastic logic (using the atomic sentences).*

$\psi := d \leftrightarrow \neg B(d)$

6. ***What is the real world****? (In other words, answer the question: is the agent in a dangerous zone or not, and is he close to the target or not?)* ***Justify your answer****, by interpreting the announcement in the previous part as a new update $!\psi$ and* ***representing the updated model****.*

   Let S be the model of question 2.4. In S, ψ holds only in world $(\neg t, d)$. In this world, d is true and $\neg B(d)$ is true (because $B(d)$ is false, the agent believes $B(\neg d)$). In the other world, $(\neg t, \neg d)$, d is false but $\neg B(d)$ is true, so ψ doesn't hold here. Because ψ is truthful, the real world is $(\neg t, d)$.

   $\neg t,\ d$