# Homework 3

This homework is worth 100 points. The answers must be submitted via Blackboard. You can answer them either individually, or in pairs or small groups of (at most) 3-4 people. Only one homework per group should be submitted, but the names of everybody in the group should be written on top of the paper.

**Question 1**. (*52 points*)

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:

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

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

$$B_a\phi \;\Leftrightarrow\; \Diamond_a\Box_a\phi$$

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

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

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

**HINT**: The positive statements (a) and (b) need general proofs: you need to show that, for *every* plausibility model and *every* sentence $\phi$, 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 $\varphi$ for which the implication fails to be true at *some* world of that model (which we can think of as the "real world").

**Question 2** (*48 points*) 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 his 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.

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 relatio**n on these worlds, by drawing arrows going from the less plausible worlds to the more plausible ones.

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 $\varphi$ in doxastic logic (using the atomic sentences).

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

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 $\psi$ in doxastic logic (using the atomic sentences).

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**.