Homework 4

This homework is worth 100 points in total. The answers **have to be typed, and submitted via Blackboard, BUT if you have difficulties in drawing models in LaTEX you man attach scans of drawings done by hand**. 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. The names of everybody in the group should be written on top of the paper**.

Question 1. (*48 points*)

A robot is in a war zone. It doesn't know its location, but all it cares is whether or not there is a **mine** in front (m), and whether or not there is an **enemy** approaching (e). The robot has two sensors, one detecting pieces of metal in the ground in front of the robot (to assess if there is a mine ahead), and another one detecting tiny vibrations in the ground (to assess if there is any enemy approaching). The robot *trusts* its sensors to be reliable, but is aware that they might not be completely reliable: it is in principle possible that the sensors might malfunction (indicating mines/enemies when there are none, or vice-versa). But, in case the sensors malfunction, the robot considers as *equally plausible* that the metal detector malfunctions and that the vibration detector malfunctions; however, the robot considers that it is *more plausible* that *only one* (either) of the sensors malfunctions than that *both* sensors malfunction.

At the moment, the sensors don't indicate any mines or enemies. So the robot *believes* there are none, although it doesn't know this for sure (since the sensors might malfunction).

1. *Represent the robots's belief-revision structure using a single-agent plausibility model*, with four possible states, using the atomic sentences m for "there is a mine in front of the robot", and e for "an enemy is approaching". Draw arrows going from the less plausible worlds to more plausible worlds (according to the robot's plausibility relation). In your drawing, you may skip the loops (since plausibility relations

are assumed to be reflexive) and the arrows that can be obtained by composing other arrows (since plausibility relations are assumed to be transitive), but be aware that they are there. Also, be sure to use all the information given in the text above.

2. Suppose now that the robot's sensors indicate some vibrations. Assume that the robot *strongly trusts* its sensors (though it doesn't know for sure they are reliable), and so interpret this event as a radical upgrade $\Uparrow e$ of the robot's belief structure with the sentence e. *Represent the robots's new belief structure (as a plausibility model) after this upgrade.*

3. *Immediately after* the event in the previous part, the robot's metal detector indicates the presence of a mine. Again, assume the robot strongly trusts the reliability of its detector, and so interpret this a new radical upgrade $\Uparrow m$ of the robot's belief structure with the sentence m. *Represent the robots's new belief structure (as a plausibility model) after this new upgrade.*

4. *Immediately after the previous two events*, the robot receives a message from its controller, saying that: "(*Right now, your vibration detector malfunctions, so*) **Whatever you currently believe about the enemy (approaching or not) is false**."

   Let us denote by φ the sentence announced by the controller. Write this sentence formally, and determine in which possible states it is true.

5. Assume that the controller is known to be *infallible*, and so that the robot performs an *update* with the announced sentence. *Represent the robots's new belief structure (as a plausibility model) after this update.* What does the robot believe about e and m after the update?

6. What would have been the final belief structure (plausibility model) of the robot if the above three events happened in *a different order*, namely: first the robot received the above message from the (infallible) controller, then its (strongly trusted) sensor indicated vibrations (hence enemies), and then its (strongly trusted) metal detector indicated a mine?

Exercise 2. (*52 points*)
A covered coin is on the table, lying either Heads up (H) or Tails up (T). There are two agents, Alice and Bob, and *it is common knowledge that neither of them can see the coin, but that for some reason they both believe that the coin lies Tails up.*

8, 10, 10, 12, 12

1. *Draw a multi-agent plausibility model* $\mathbf{M}_0$ (with two agents, a for Alice and b for Bob, and atomic sentence H and T) to accurately represent all the information above.

2. Some external referee publicly announces: "*The coin lies Heads up*". It is *common knowledge that: Bob strongly trusts the referee, but that Alice is neutral towards the referee* (neither trusts nor distrusts him).

   Represent this action as an **event plausibility model $\mathbf{\Sigma}$**.

3. *Represent (draw) a model* $\mathbf{M}_1$ for the situation *after* the action described in the previous part, by computing the Action-Priority update $\mathbf{M}_1 = \mathbf{M}_0 \otimes \mathbf{\Sigma}$.

4. We now change the scenario as follows. It is *still common knowledge that Bob strongly trusts the referee*, but Alice's attitude is NOT common knowledge. Instead, it is common knowledge that: *either Alice strongly trusts the referee, or she is neutral (neither trusting nor distrusting), or she strongly distrusts him.* Moreover, it is also common knowledge that: *Bob believes that Alice trusts the referee*; but that, *IF given that this belief is wrong, he'd believe that she's neutral* (rather than that she is strongly distrusting the referee).

   Of course, we also assume *each agent knows his/her own attitude.*

   As before, the referee publicly announces: "*The coin lies Heads up*".

   Represent this action as an **event plausibility model $\mathbf{\Sigma}'$**.

   HINT: This is a plausibility event model with 6 actions.

5. Starting from the original situation (in part 1), suppose the action that we described in the previous part happens.

   Represent (draw) a model $\mathbf{M}'_1$ for the situation *after* this action, by computing the Action-Priority update $\mathbf{M}'_1 = \mathbf{M}_0 \otimes \mathbf{\Sigma}'$.