COSC410 Lecture 7
Epistemic Logic

Willem Labuschagne
University of Otago

2016

Introduction

Suppose we want to generalise belief change theory to a group of agents. Typically, an agent in a multi-agent system wants to reason about both the system of interest and the other agents. Players in a game of bridge or poker want to know not only what cards are in each hand (objective knowledge) but also what each of the other players is thinking (subjective[1] knowledge). Subjective knowledge may be represented by a new connective in a way first explained in the book *Knowledge and Belief* written by the Finnish logician Jaakko Hintikka in 1962.

Take any propositional language L_A and, for each agent in the multi-agent system, add to the symbols of L_A a new connective called a *modal box operator*, in other words add $[i]$ for each agent $i \in I$, where I is the set of agents in the multi-agent system. Let L_A^I be the language that results. (If there is only one agent, the box operator is simply $\Box$, without an index inside, and the language is $L_A^{\Box}$.) The syntax of the new language is simple: in concert with the other connectives, apply $[i]$ as a prefix to form new sentences. (So box operators are applied like negation, and have as scope the shortest sentence following the box.)

The propositional language with the new box operators is called an *epistemic* language. Suppose we design an epistemic language in which to represent knowledge about the 3 Card System. Since there are three players, we could take $I = \{1, 2, 3\}$. The set of atoms could stay the same as before: $A = \{r_1, r_2, r_3, b_1, b_2, b_3, g_1, g_2, g_3\}$. Now the sentences of L_A^I include atomic sentences like r_1 and g_3, as before, but also new compound sentences like $[2]r_1$, and $\neg[2]r_1$, and $[2]\neg r_1$, and $[3]([1]r_1 \vee [2]g_3)$ and so forth.

The idea will be that we should read a sentence of the form $[i]\varphi$ as "Agent i knows that φ", but before we can justify this reading we have to describe the semantics of our new language. Something to keep in mind is that epistemic logic ignores heuristic information about what is typical or probable and focuses solely on high-confidence definite information such as might be gained by observation or by report from a trusted source, i.e. epistemic logic is classical.

[1] Subjective information is information about agents. We do not intend "subjective" to mean that the information is merely a matter of opinion, i.e. uncertain.

Adding accessibility relations to the semantics

The semantics of L_A^I must represent both the objective system of interest and the agents who are observing and interacting with the system.

To specify the semantics we first give the set S of states of the objective system and a labelling function $V : S \longrightarrow W_A$ that determines which atomic sentences are true in which states. This should sound familiar.

We must also describe, for each agent $i \in I$, an *accessibility* relation $\approx_i \subseteq S \times S$. Here the idea is that an agent will usually have limited information about the system of interest (i.e. about the environment). For example, an agent observing something happening some distance away may be able to see that two other agents are talking but not hear what is being said. Or, an agent listening to your phone calls may hear what you are saying, but perhaps cannot see what you are doing. The sensors (or senses) of an agent can probe the environment only in certain places and to a certain depth. Each accessibility relation in the semantics of L_A^I reflects the constraints that limit the relevant agent's capacity to obtain information about the objective system.

Note that in the simple case of a single agent, the semantics we're describing resembles that of nonmonotonic logic except that we have an accessibility relation $\approx$ instead of a total preorder $\preccurlyeq$.

What kind of thing is the accessibility relation of an agent? Suppose an agent has a limited amount of information about what the state of the system is. Then there will be several candidates for being the actual state, and the agent will not be able to pick out the one that is the real state of the system. In other words, some of the states will be equivalent in the sense that the agent cannot distinguish between them on the basis of the available information. Accordingly we shall take the accessibility relation to be an *equivalence* relation.

Definition 1 *An equivalence relation $\approx$ on S is a subset of $S \times S$ that is reflexive, transitive and symmetric, i.e.*

- *$x \approx x$ for every $x \in S$ (reflexivity)*
- *if $x \approx y$ and $y \approx z$ then $x \approx z$, for all $x, y, z \in S$ (transitivity)*
- *if $x \approx y$ then $y \approx x$, for all $x, y \in S$ (symmetry).*

Don't worry if this seems very abstract. We'll explore some examples to see how one gets and uses the equivalence relations.

Example 1 *(Single agent) Consider the Traffic System. Let's assume the driver waiting at the intersection has the following constraint on her capacity to obtain information about the system: she can see whether or not the traffic light for cross traffic is red, but cannot see until too late whether the oncoming car is going to stop. How do we represent this constraint as an equivalence relation?*

Suppose the system is really in the state 11*, i.e. the traffic light is red and the oncoming car is going to stop. Remember, the waiting driver does not know this. As far as the waiting driver is concerned, the actual state could be either* 11 *or* 10*, because all she can see is that the light is red. So for the driver, the states* 11 *and* 10 *are equivalent. Similarly, the states* 01 *and* 00 *are equivalent, because the agent (able to see only whether the light is on) cannot distinguish between them.*

This means that we should associate with our agent the equivalence relation that makes $11 \approx 10$ *(and thus by symmetry* $10 \approx 11$*) and makes* $01 \approx 00$ *(and thus also* $00 \approx 01$*). If we were to write out all the ordered pairs in the relation* $\approx$ *we would get*

$$11 \approx 11, 10 \approx 10, 01 \approx 01, 00 \approx 00, 11 \approx 10, 10 \approx 11, 01 \approx 00, 00 \approx 01.$$

In the example above we have a single agent, namely the driver waiting at the intersection. The equivalence relation $\approx$ that we have so carefully constructed is the accessibility relation of the agent. More generally, if we have a set I of agents, then the equivalence relation $\approx_i$ associated with agent i is the accessibility relation for agent i.

The essential idea is that if the system is actually in state $s \in S$ but states s, t, u are equivalent to s (i.e. accessible from s), then the agent will consider s, t and u all to be equally good candidates for being the actual state, because the information available to the agent is compatible with all three states and cannot decide between them. For example, in the Traffic System an agent who sees that the light is red considers the states 11 and 10 to be equally good candidates for being the real system, and (unlike the case of nonmonotonic logic) does not have heuristic information that may help to discriminate between them. They are entirely equivalent as far as the agent knows.

Collecting together states that are equivalent to one another gives us an equivalence class.

Handy tip: The easiest way to build the equivalence relation for an agent is to remember that the equivalence classes divide S up (technically, they *partition* S) into one or more pieces (technically, the pieces are disjoint subsets whose union is the whole of S). So we might say that $\approx$ is the equivalence relation that partitions $S = \{11, 10, 01, 00\}$ into the subsets $\{11, 10\}$ and $\{01, 00\}$, because this is just an easy way to say that $11 \approx 10$ and that $01 \approx 00$, from which we can get all the other pairs such that $x \approx y$ simply by remembering that $\approx$ has to be reflexive, transitive, and symmetric.

Example 2 *(Multiple agents) Let's look at modelling the 3 Card System in epistemic logic. Take* $I = \{1, 2, 3\}$ *and* $A = \{r_1, r_2, r_3, g_1, g_2, g_3, b_1, b_2, b_3\}$*. The semantics of* L_A^I *involves* $S = \{rgb, rbg, grb, gbr, brg, bgr\}$ *and the obvious labelling function. What accessibility relations should be associated with the agents? Well, that depends. What are the constraints (i.e. limitations) on the ability of each agent to get information from the environment (i.e. from the system of interest)?*

Suppose players 1 *and* 2 *are ordinary card players, each limited to seeing his or her own cards, whereas player* 3 *is a cheat using an elaborate spy system that allows him to see not only his own card but also what card each of the other players has been dealt.*

As far as player 1 *is concerned, the states* rgb *and* rbg *are indistinguishable, because she can see only the red card in her own hand. Similarly, for player* 1 *the states* grb *and* gbr *are equivalent, and the states* brg *and* bgr *are equivalent. Thus the equivalence relation associated with player* 1 *must divide* S *up into the three subsets* $\{rgb, rbg\}$*,* $\{grb, gbr\}$*, and* $\{brg, bgr\}$*. Hence* $rgb \approx_1 rbg$ *and* $grb \approx_1 gbr$ *and* $brg \approx_1 bgr$*. This tells us everything we need to know about* $\approx_1$*.*

In the case of player 2*, the states* rgb *and* bgr *are indistinguishable because he can see only the card in his own hand, and thus* $rgb \approx_2 bgr$*. Similarly,* $rbg \approx_2 gbr$ *and* $grb \approx_2 brg$*. Or to put it differently,* $\approx_2$ *partitions* S *into* $\{rgb, bgr\}$*,* $\{rbg, gbr\}$*, and* $\{grb, brg\}$*.*

In the case of player 3*, who can see precisely what card is in each hand, the set* S *is partitioned into the six subsets* $\{rgb\}$*,* $\{rbg\}$*,* $\{grb\}$*,* $\{gbr\}$*,* $\{brg\}$*,* $\{bgr\}$*. In other words, player* 3 *is able to distinguish between all the states, and can tell exactly what the actual state of the system is.*

We see that player 3 has the most information, and this is reflected by his accessibility relation having small equivalence classes. Players 1 and 2 have less information, which is reflected by their accessibility relations having larger equivalence classes — more things are equivalent because the agent can't tell the difference between them. In the next section, we will look more closely at what an agent knows. We will see that an agent able to extract lots of information from the environment (as shown by the accessibility relation having small equivalence classes) will indeed know a lot (as shown by the sentences in the agent's knowledge set).

Knowledge

To bring the the syntax and semantics of our logic together we need a notion of satisfaction. We have previously seen how to work out whether a state satisfies a sentence built up from the atomic sentences using only the propositional connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$. But what about sentences in which modal box operators $[i]$ occur? We need to add a new clause to our definition of satisfaction.

Definition 2 *(Satisfaction of $\Box$) A state w satisfies $[i]\varphi$ iff φ is satisfied by every state u such that $w \approx_i u$.*

In the single-agent case, a sentence of the form $\Box\varphi$ is true in a state w if every state u accessible from w satisfies φ, and there is only one equivalence relation to use for accessibility. With multiple agents, we use the equivalence relation matching the agent named in the box operator.

When we explored AGM belief change operations, we associated with every agent a belief set K, the set of sentences believed by that agent at that point. Now we can associated with each agent something similar — the agent's knowledge when the system is in a particular state.

Definition 3 *(Knowledge sets) The knowledge of agent i in state w is the set of all sentences φ such that w satisfies $[i]\varphi$.*

From the definition of satisfaction, we see that agent i will know that φ if φ is satisfied not only by w, the actual state of the system, but also by all the other states u that agent i cannot distinguish from w, i.e. by all the states u in the equivalence class of w. If the agent's accessibility relation has small equivalence classes, then it is much easier for a sentence to be satisfied by all the states in the given equivalence class. So the agent knows a lot. But if the agent's accessibility relation has large equivalence classes, then fewer sentences will manage to be satisfied by all the states in the equivalence class, so the agent will know fewer things.

Example 3 *(Single agent) Consider again the Traffic System (one agent waiting at the intersection, limited to seeing whether the traffic light is red, with accessibility relation $\approx$ given by $11 \approx 10$ and $01 \approx 00$). Recall that p stands for "The light is red" and q for "The oncoming car stops."*

Let's see what the agent knows. Intuitively, since the agent can see (only) whether the light is red, we should expect that if the system is actually in state 11, then the agent should know p but should not know q. Does our formal system give these expected results?

Yes indeed, the state 11 satisfies $\Box p$ since 11 and 10 are the two states accessible from 11 and both satisfy p. Thus the agent knows p when the system is in state 11.

However, state 11 does not satisfy $\Box q$ since 11 and 10 are the two states accessible from 11 and 10 fails to satisfy q. Thus the agent does not know q when the system is in state 11.

More generally, whatever the state is, the agent with the given accessibility relation $\approx$ should know whether the light is red, i.e. in states 11 *and* 10 *the agent should know p but in states* 01 *and* 00 *the agent should know $\neg p$. Does our logic bear out this intuition? Indeed it does.*

By similar reasoning to the above, state 10 *satisfies $\Box p$, indicating that when the system is in state* 10 *the agent knows the light is red.*

On the other hand, state 01 *fails to satisfy $\Box p$ since the states accessible from* 01 *are* 00 *and* 01*, and neither satisfies p. So when the actual state is* 01 *then it is not the case that the agent wrongly knows p. In fact,* 01 *satisfies $\Box\neg p$, telling us that in state* 01 *the agent knows $\neg p$, i.e. knows that the traffic light is not red.*

Similarly, state 00 *fails to satisfy $\Box p$, but* 00 *does satisfy $\Box\neg p$.*

Let us perform a further check of whether the agent's knowledge fits our intuition — in state 11*, where we have checked that the agent does not know q, does the agent erhaps wrongly know $\neg q$? Happily, no such error occurs. The two states accessible from* 11 *are* 11 *and* 10*, and* 11 *does not satisfy $\neg q$, so that* 11 *does not satisfy $\Box\neg q$.*

We have seen that the agent always knows whether p is true or not, but that the agent does not always know whether q is the case, because in state 11 *the agent does not know that q and also does not 'know' that $\neg q$.*

Similarly, the agent does not know whether q is the case in states 10*,* 01*, and* 00 *because none of these states satisfy $\Box q$ and none of them satisfy $\Box\neg q$.*

Incidentally, how does one express something like "The agent knows whether the car stops" in the logic? Well, if the agent knows whether the car stops, then the agent will know that the car is stopping or the agent will know that the car is not stopping. In other words, we can use the sentence $\Box q \vee \Box\neg q$ to express "whether".

Example 4 *(Multiple agents) Let the 3 Card System have $A = \{r_1, r_2, r_3, g_1, g_2, g_3, b_1, b_2, b_3\}$, $I = \{1, 2, 3\}$, and $S = \{rgb, rbg, grb, gbr, brg, bgr\}$ with the obvious labelling function. As before, let the accessibility relations be such that*

- *$\approx_1$ partitions S into $\{rgb, rbg\}$, $\{grb, gbr\}$, and $\{brg, bgr\}$*
- *$\approx_2$ partitions S into $\{rgb, bgr\}$, $\{rbg, gbr\}$, and $\{grb, brg\}$*
- *$\approx_3$ partitions S into $\{rgb\}$, $\{rbg\}$, $\{grb\}$, $\{gbr\}$, $\{brg\}$, $\{bgr\}$.*

Suppose the actual state of the system is rgb. What are some of the things the agents know or don't know in this state?

Agent 1 *knows that she has the red card because rgb satisfies $[1]r_1$ (since the states accessible from rgb via $\approx_1$ are rgb and rbg, and both of them satisfy r_1).*

But agent 2 *does not know that agent* 1 *has the red card, because rgb does not satisfy $[2]r_1$ (since $rgb \approx_2 bgr$ and bgr fails to satisfy r_1).*

On the other hand, agent 3 *does know that agent* 1 *has the red card, because the only state accessible from rgb via $\approx_3$ is rgb itself, and rgb satisfies r_1.*

So far we have spoken about the objective knowledge of the agents, i.e. what they know about the cards that were dealt. What about their subjective knowledge, i.e. what they know about the agents' knowledge?

Agent 1 *knows that she knows that she has the red card, because the actual state rgb satisfies* $[1][1]r_1$*. To see this, note that the states accessible from rgb via* $\approx_1$ *are rgb and rbg. Now rgb satisfies* $[1]r_1$ *as we saw above, and similarly rbg satisfies* $[1]r_1$ *because the states accessible from rbg via* $\approx_1$ *are rgb and rbg and both satisfy* r_1*. (What this has shown is that in the state rgb, the actual state of the system, agent 1 has introspective knowledge, i.e. knows what it is that she knows. We'll see later whether this sort of thing holds in general.)*

Does agent 2 *know that agent* 1 *knows that agent* 1 *has the red card? Well, does the actual state rgb satisfy* $[2][1]r_1$*? The states accessible from rgb via* $\approx_2$ *are rgb and bgr, and bgr does not satisfy* $[1]r_1$ *(because bgr* $\approx_1$ *bgr and bgr fails to satisfy* r_1*). Hence rgb does not satisfy* $[2][1]r_1$*, meaning that in state rgb agent* 2 *does not know that agent* 1 *knows that agent* 1 *has the red card. (So agents do not telepathically read one another's minds.)*

Surely the agents all know that agents can see their own cards? So we'd expect agent 2 *to know that agent* 1 *knows what her own card is, and in particular whether or not she has the red card. Let's check this. Does the actual state, rgb, satisfy the sentence* $[2]([1]r_1 \vee [1]\neg r_1)$*? Agent* 2 *considers rgb and bgr to be equivalent. In rgb,* $[1]r_1$ *is true because agent* 1 *considers rgb and rbg equivalent, and both states satisfy* r_1*. In bgr,* $[1]\neg r_1$ *is true because agent* 1 *considers bgr and brg to be equivalent, and both states fail to satisfy* r_1*, i.e. satisfy* $\neg r_1$*. Hence all states accessible from the actual state rgb via* $\approx_2$ *satisfy* $[1]r_1 \vee [1]\neg r_1$*, so that rgb satisfies* $[2]([1]r_1 \vee [1]\neg r_1)$*.*

Recall that the accessibility relation of agent 3 *was selected as an example of an agent who always knows exactly what the state of the system is. This results in an agent with a surprising degree of telepathic power. Agent* 3 *does know that agent* 1 *knows that agent* 1 *has the red card, because the actual state rgb satisfies* $[3][1]r_1$ *(since the only state accessible from rgb via* $\approx_3$ *is rgb itself, and rgb satisfies* $[1]r_1$ *as we saw previously). So although agents usually cannot telepathically read one another's minds, an agent may sometimes have enough information to be able to tell exactly what another agent knows. (What makes it a bit strange is that the accessibility relation of agent 3 just tells us agent 3 is able to detect which card is dealt to which player, i.e. is restricted to the agent's capacity to extract information about the objective system, and doesn't refer directly to other agents and what they may be thinking. So it's a bit mysterious that agent 3 ends up knowing what it is that agent 1 knows. I consider this a paradox of epistemic logic.)*

Metalogic

We start by proving a useful lemma.

Lemma 1 *For every sentence* $\varphi \in L_A^I$*, state* $w \in S$ *and agent* $i \in I$*,* w *satisfies* $[i]\varphi \to \varphi$*.*

(Note that we are considering the sentence $([i]\varphi) \to \varphi$*, not* $[i](\varphi \to \varphi)$*.)*

Proof. *The only way in which* w *can fail to satisfy* $[i]\varphi \to \varphi$ *is for* w *to satisfy the antecedent* $[i]\varphi$ *but fail to satisfy the consequent* φ*.*

Suppose w *satisfies* $[i]\varphi$*.*

Then every state accessible from w *via* $\approx_i$ *satisfies* φ*.*

Since $\approx_i$ *is an equivalence relation, it is reflexive, and so* $w \approx_i w$*.*

Hence w *must satisfy* φ*.*

Therefore w *satisfies* $[i]\varphi \to \varphi$*.* ■

The significance of the lemma is that it allows us to prove a theorem illustrating that epistemic logic ignores the defeasible beliefs which may be formed by agents on the basis of their heuristic information.

Theorem 1 *If in state w agent i knows φ, then φ is true in w.*

Proof. *Suppose agent i knows φ in state w*

i.e. suppose w satisfies $[i]\varphi$.

By the lemma, w satisfies $[i]\varphi \to \varphi$.

Therefore w satisfies φ. ■

From the perspective of everyday reasoning, this theorem points out a limitation of epistemic logic: it cannot represent a belief that is plausible but turns out not to be true. Nevertheless you should be aware that it is traditional in philosophy to pretend that this is actually a strength of epistemic logic, a desirable property, usually described pompously as asserting that *knowledge is veridical.* In other words, the property is taken as defining what knowledge is, so that we may contrast knowledge with mere belief.

Speaking plainly, epistemic logic can represent the reasoning of a bean-counter checking your tax return, since bookkeeping involves sticking very close to the arithmetical facts. Similarly, epistemic logic is very useful when dealing with computing machines where "knowing" something might correspond to performing some sort of Boolean test or looking up data in a table. Defeasible beliefs don't come into it because typically we haven't programmed the machines to incorporate heuristic information. Of course, the moment we become ambitious and try to design, say, a robot detective like Daneel Olivaw, able to solve murders by sifting through evidence and using qualitative impressions of the suspects, we'll want to incorporate heuristic information, and epistemic logic becomes inapplicable. Similarly, epistemic logic cannot represent the reasoning of a bridge or poker player, nor a doctor making a diagnosis, since heuristic information must be used in these activities as the basis for 'educated guesses' or 'best estimates'. And don't forget the simple case of a driver who wants to cross an intersection.

Could we modify epistemic logic so that it can represent both knowledge and potentially false beliefs? This has been attempted, and results in so-called *doxastic* logics. (In classical Greek, the word for 'belief' was *doxasia*.)

To build a doxastic logic we proceed as follows. Instead of writing $[i]\varphi$ for "Agent i knows that φ", let us use the symbol K for "knows" and write $K_i\varphi$. And let's add to the language a bunch of new modal box operators in the form of the symbol B for "believe", so that we may write $B_i\varphi$ for "Agent i believes that φ". Each knowledge operator K_i still has an equivalence relation as its accessibility relation. For each belief operator B_i we would use a different kind of relation (not an equivalence) as the accessibility relation. We would ensure that the accessibility relation for B_i is not reflexive. This means we won't have the awkward property that belief is veridical, i.e. that if an agent believes φ then φ is true (since one needs reflexivity to prove that property). The problem is that these hybrid systems still idealise knowledge and belief to an unreasonable extent. Specifically, properties such as positive and negative introspection apply to both knowledge and belief in these hybrid systems.

What is *positive introspection*? We have already encountered the idea of positive introspection, namely that if the agent knows φ then she knows that she knows it (see example 4). This is not very realistic as a general rule. We may know something without being conscious that we know it. Imagine you are reading Harry Potter for the first time, and you remark to a friend, in joyful amusement, "Oh look at this money! There are 17 sickles to a galleon, and 23 knuts to a sickle." Your friend replies: "Yes, and did you notice that those are prime numbers?" You exclaim: "Really? Hang on a mo." After a pause in which you do some quick calculations you say: "Oh yes, I see." One could judge that you knew, in some sense, that 17 and 23 were prime numbers, because you knew exactly how to work out whether they were prime, in other words their primeness was a logical consequence of the things you already knew about numbers. But not having ever worked it out before, it wasn't something you were aware of consciously. So if asked "Did you know that you knew that 17 and 23 were prime?" you would have had to respond: "No."

Similarly, *negative introspection* is the idea that if an agent doesn't know φ then she knows that she doesn't know it. This is even less realistic. Do you know exactly what it is that you don't know? Consider a house-cleaning robot. While the robot is cleaning the living-room downstairs, it doesn't know that the child who sleeps upstairs left her doll on the bedroom floor. Does the robot know that it doesn't know that? Clearly not. Even if we substitute a human caregiver/cleaner for the robot, it would be unreasonable to expect an agent in such a situation to possess the property of negative introspection. However, one can imagine restricted applications where an agent can spell out all the possibilities and say something along the lines of "Well, it must be one of those, but I don't know which". For such restricted applications, epistemic logic works. It's the broader application to everyday reasoning that makes it look silly.

Let's show that the agents modelled by epistemic logic have positive introspection.

Theorem 2 *If state w satisfies $[i]\varphi$ then w satisfies $[i][i]\varphi$.*

Proof. *Suppose w satisfies $[i]\varphi$. Let $w \approx_i u$.*

I claim that u satisfies $[i]\varphi$.

To see this, pick any v such that $u \approx_i v$.

Since $w \approx_i u$ and $u \approx_i v$, by transitivity $w \approx_i v$.

Since w satisfies $[i]\varphi$ and $w \approx_i v$, v satisfies φ.

Since v was arbitrary, u satisfies $[i]\varphi$.

Since u was arbitrary, w satisfies $[i][i]\varphi$. ■

Corollary 1 *Every $w \in S$ satisfies every sentence of the form $[i]\varphi \to [i][i]\varphi$.*

Note that positive introspection iterates. As soon as agent i knows φ, then she also knows $[i]\varphi$, and knows $[i][i]\varphi$, and knows $[i][i][i]\varphi$, and so forth.

Now let us consider a new question. We've seen it's not easy to incorporate defeasible beliefs into epistemic logic. However, there is a very weak form of speculative contemplation that may easily be incorporated. A sentence of the form $[i]\varphi$ says "Agent i knows that φ" and similarly $\neg[i]\varphi$ says that agent i doesn't know that φ. Now consider a sentence of the form

$$\neg[i]\neg\varphi.$$

This sentence says the agent doesn't know that φ is false. To put it differently, the agent considers it possible for φ to be true.

The construction used above is so useful that we introduce a new *diamond* operator to abbreviate it.

Definition 4 *We write $\langle i\rangle\varphi$ as abbreviation for $\neg[i]\neg\varphi$, and read it as "Agent i considers it possible that φ". In the single-agent case, $\neg\Box\neg\varphi$ is written $\Diamond\varphi$.*

Since we know how to figure out whether a state satisfies a sentence having a box operator, we can also figure out whether the state satisfies a sentence having a diamond operator, but it is convenient to give the criterion directly.

Definition 5 *(Satisfaction of $\Diamond$) A state w satisfies $\langle i\rangle\varphi$ iff φ is satisfied by at least one u such that $w \approx_i u$.*

To illustrate, let us examine a new example which also shows that in epistemic logic we may want the set S of states to be different from the set W_A of truth assignments.

Example 5 *The Weather System is very simple. Our language has one atomic sentence p standing for "It is raining in Dunedin" and we consider the case of two agents, so $I = \{1, 2\}$. The set W_A has two truth assignments: $f(p) = 1$ and $g(p) = 0$. But we take $S = \{s, t, u\}$ because we want s and t to be states of the system in which agent 1 is working in a windowless room so that she cannot see what the weather is and state u to be the state of the system in which the weather is the same as for s but agent 1 has left her windowless office and can see what the weather is. On the other hand, we'll suppose that agent 2 can see what the weather is at all times.*

So let $V : S \longrightarrow W_A$ be given by $V(s) = f, V(t) = g, V(u) = f$ which means that p is true in states s and u but false in state t. Let the accessibility relations of the agents be given by

- *$s \approx_1 t$ (so $\approx_1$ partitions S into $\{s, t\}$ and $\{u\}$*
- *$s \approx_2 u$ (so $\approx_2$ partitions S into $\{s, u\}$ and $\{t\}$.*

Although states s and u both make p true, they are different in other respects.

If the system is in state s, then it is raining, i.e. p is true, and agent 2 knows this, since the states accessible from s via $\approx_2$ are s and u, both of which satisfy p, so that s satisfies $[2]p$. But agent 1 doesn't know that it is raining: s fails to satisfy $[1]p$ since $s \approx_1 t$ and t fails to satisfy p.

On the other hand, if the system is in state u, then agent 1 does know it is raining: u satisfies $[1]p$ since the only state accessible from u via $\approx_1$ is u itself, and u does satisfy p.

This illustrates that two different states may be 'extensionally' the same, i.e. the same as far as the truth values of the atoms go, but may differ in what they offer an agent in information. Note that this is a different reason for having $S \neq W_A$ than we saw in the case of the 3 Card System. In the 3 Card System, we had many spurious truth assignments, and so we wanted S to pick out the six realisable truth assignments, in other words S essentially corresponded to a subset of W_A. But in the Weather System, S has two states that correspond to the same truth assignment in W_A. We shall see something similar when we look at temporal logic next time.

What about possibility? Well, in state s agent 1 does not know that it is raining (although it really is), but agent 1 does at least consider it possible that it may be raining: s satisfies $\langle 1\rangle p$ since $s \approx_1 s$ and s satisfies p, so that there is at least one state accessible from s that satisfies p.

It makes sense to ask whether diamonds have the same properties as boxes. The answer is, sometimes but not always.

Recall that box operators endowed agents with the property of positive introspection. Let's examine the analogous property for diamond operators.

Theorem 3 *Every $w \in S$ satisfies every sentence of the form $\langle i\rangle\varphi \to \langle i\rangle\langle i\rangle\varphi$.*

Proof. *Suppose w satisfies $\langle i\rangle\varphi$.*

Then u satisfies φ for some u such that $w \approx_i u$.

Since $u \approx_i u$ by reflexivity, it follows that u satisfies $\langle i\rangle\varphi$.

Thus w satisfies $\langle i\rangle\langle i\rangle\varphi$. ■

Now let's find a sensible and interesting property that holds for box operators but **not** for diamonds, and see how we construct a counterexample in the diamond case. The property we examine is basically an epistemic version of Modus Ponens. It says that if an agent knows that $\varphi \to \psi$ is the case and knows that φ is the case then that agent will know that ψ is also the case. The property relies on the fact that in order for an agent to *know* that something is the case, that something must be true in *all* states that are accessible from the actual state.

Theorem 4 *Every sentence of the form $[i](\varphi \to \psi) \to ([i]\varphi \to [i]\psi)$ is satisfied by each $w \in S$.*

Proof. *Suppose w satisfies $[i](\varphi \to \psi)$ but fails to satisfy $[i]\varphi \to [i]\psi$.*

By the former, u satisfies $\varphi \to \psi$ for every u such that $w \approx_i u$.

By the latter, w must satisfy $[i]\varphi$ but fail to satisfy $[i]\psi$.

So every u such that $w \approx u$ will satisfy φ

but one of these (say u') fails to satisfy ψ.

Since u' is accessible from w, we must have that u' satisfies $\varphi \to \psi$.

This is not possible, since u' satisfies φ but not ψ. ■

On the other hand, for diamonds, the analogous property doesn't hold. The agent might consider it possible for $\varphi \to \psi$ to be the case because there is some state in which, say, φ is false, and the agent might also consider it possible for φ to be the case because there is a different state in which φ is true. But there may be no state in which ψ is true. The proof we give below simply spells out an example of a Kripke model in which the states behave in this manner.

Theorem 5 *Sentences of the form $\langle i\rangle(\varphi \to \psi) \to (\langle i\rangle\varphi \to \langle i\rangle\psi)$ need not be satisfied by all states.*

Proof. *It is sufficient to consider the single-agent case.*

Take the language $L^{\Box}_{\{p,q\}}$ with $S = \{11, 10, 01, 00\}$.

Let $11 \approx 01$ and $10 \approx 00$.

Take $\varphi = p$ and $\psi = q$.

Now 10 satisfies $\Diamond(p \to q)$ because $10 \approx 00$ and 00 satisfies $p \to q$.

However, 10 fails to satisfy $\Diamond p \to \Diamond q$.

To see this, note that 10 satisfies $\Diamond p$ since $10 \approx 10$ and 10 satisfies p.

But 10 fails to satisfy $\Diamond q$ since 10 and 00 are accessible from 10 and both fail to satisfy q. ■

Dynamic epistemic logic

We have already mentioned some ways in which epistemic logic differs from belief change theory. In what follows, we want to say more. (Relax — this section is not for exam purposes.)

First, a bit of terminology for you. Given an epistemic language L_A^I, we don't actually have an epistemic logic until we decide on a semantics for L_A^I. This semantics has three main parts: the set S of states, a labelling function $V : S \longrightarrow W_A$ revealing which atomic sentences are true in which states, and a set of equivalence relations $\approx_i$ (one for each agent). It has become fashionable to call the semantics a *Kripke structure* for the epistemic language.[2]

Now let's think about the fact that an agent's internal representation of the environment may have to change.

When we explored AGM belief change theory, we restricted consideration to a single agent whose internal representation of the world included heuristic information and thus defeasible beliefs. In this context, we looked at three ways to change the belief set of the agent, namely expansion, contraction, and revision. Both contraction and revision made use of heuristic information (represented by a total preorder $\preccurlyeq$ on S). In other words, AGM belief change is essentially about an agent changing her belief set after she learns that one of her defeasible inferences was mistaken. AGM belief change is not really about the update of a belief set that would be necessitated by the system changing its state – instead one should think about the system remaining in the same state but the agent trying to build a more accurate picture of that state.

In the case of epistemic logic, we generalise to multiple agents (a set I of agents) but ignore heuristic information, and therefore cannot rely on the AGM postulates as a guide to the changes that may occur in the knowledge of agents. Nevertheless, it does make sense to think about how such knowledge might change.

The underlying assumption in epistemic logic is that an agent's knowledge (while it may be incomplete) is accurate. A change in an agent's knowledge may be the result of the agent expanding her knowledge by learning some new fact about the system (while the system remains in the same state) or the system may change its state (so that the agents may need to update their respective mental pictures of what the current state is).

Such changes are represented by changes in the Kripke structure associated with the language. Typically, an update would make it clear to all the agents that some state can be ruled out, and the new Kripke structure would use a set $S' \subseteq S$ in which the excluded state no longer occurs. The study of this sort of change, which might be the result of some agent making a public announcement (such as player 1 announcing to all the other players "I have the red card"), is called *dynamic epistemic logic*, with the word "dynamic" alerting us to the action (e.g. a public announcement) that forces the change of Kripke structure.

A recent account of the field is given in *Dynamic Epistemic Logic* by Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi, published by Springer Verlag. Hans was a member of our department for a number of years, Wiebe visited us several times, and Barteld spent an extended period with us too as a visiting researcher.

[2] Saul Kripke is a famous logician who at the age of 16 wrote a paper on this sort of semantics. He was the 6th logician to do so, after Tarski, Carnap, Prior, Hintikka and Kanger, so it's a bit strange to identify the semantics as Kripke semantics. Arthur Norman Prior was a New Zealand logician who invented temporal logic, so maybe we should start calling it Prior semantics.

Exercises

Quiz: The quiz at the start of lecture 8 will be based on exercise 1.

1. Consider the 3 Card System with language and semantics as earlier in this lecture. Again assume that the actual state is rgb. Answer each of the following by demonstrating that the state rgb satisfies an appropriate sentence.
   - Does player 1 know that the green card is in player 2's hand?
   - Does player 1 consider it possible that the green card is in player 2's hand?
   - Does player 1 know that player 3 knows that the green card is in player 2's hand?
   - Does player 1 know that player 2 doesn't know that the red card is in player 1's hand?
   - Does player 1 know that player 2 doesn't know that player 1 knows that the red card is in player 1's hand?
   - Does player 1 know that player 3 knows that the red card is in player 1's hand?
   - Does player 1 know that player 3 knows that player 1 knows that the red card is in player 1's hand?
   - Does player 1 consider it possible that the red card is in player 2's hand?
   - Does player 1 consider it possible that player 2 knows that the red card is in player 1's hand?
   - Does player 1 consider it possible that player 2 considers it possible that the red card is in player 1's hand?
2. Show that for every sentence $\varphi \in L_A^I$ and every state $s \in S$ and every agent $i \in I$, the state s satisfies
$$\neg[i]\varphi \to [i]\neg[i]\varphi.$$
   (This property is known as negative introspection, and tells us that the agents modelled by epistemic logic know what they don't know.)
3. Show that for all sentences $\varphi, \psi \in L_A^I$ and every state $s \in S$ and every agent $i \in I$, the state s satisfies
$$[i](\varphi \wedge \psi) \to ([i]\varphi \wedge [i]\psi).$$
4. Show that it is not the case that for all sentences $\varphi, \psi \in L_A^I$ and every state $s \in S$ and every agent $i \in I$, the state s satisfies
$$[i](\varphi \vee \psi) \to ([i]\varphi \vee [i]\psi).$$
5. Do you think it is the case that for all sentences $\varphi, \psi \in L_A^I$ and every state $s \in S$ and every agent $i \in I$, the state s satisfies
$$\langle i\rangle(\varphi \to \psi) \to ([i]\varphi \to \langle i\rangle\psi)?$$
   If so, prove it. If not, give a counter-example. (Note the occurrence of a box operator, so that this is a new problem.)