Epistemic Logic and Information Update

A. Baltag, H.P. van Ditmarsch, and L.S. Moss

January 14, 2009

1 Prologue

Epistemic logic investigates what agents know or believe about certain factual descriptions of the world, and about each other. It builds on a model of what information is (statically) available in a given system, and isolates general principles concerning knowledge and belief. The information in a system may well change as a result of various changes: events from the outside, observations by the agents, communication between the agents, etc. This requires *information updates.* These have been investigated in computer science via *interpreted systems*; in philosophy and in artificial intelligence their study leads to the area of *belief revision.* A more recent development is called *dynamic epistemic logic.* Dynamic epistemic logic is an extension of epistemic logic with dynamic modal operators for belief change (i.e., information update). It is the focus of our contribution, but its relation to other ways to model dynamics will also be discussed in some detail.

Situating the chapter This chapter works under the assumption that *knowledge* is a *variety of true justifiable belief.* The suggestion that knowledge *is nothing but* true *justified* belief is very old in philosophy, going back to Plato if not further. The picture is that we are faced with alternative "worlds", including perhaps our own world but in addition other worlds. To know something is to observe that it is true of the worlds considered possible. Reasoners adjust their stock of possible worlds to respond to changes internal or external to them, to their reasoning or to facts coming from outside them.

The *identity* of knowledge with true justified (or justifiable) belief has been known to be problematic in light of the Gettier examples (see also our discussion in Section 3.3). Being very short again, the point is that this simple picture ignores the *reasons* that one would change the collection of possibilities considered, and in particular it has nothing to say about an agent who made a good change for bad reasons and thereby ends up "knowing" something in a counter-intuitive way.

However, the work presented in this chapter is in no way dependent on this mistaken identity: while all the forms of knowledge presented here are forms of true justified belief, the converse does not necessarily hold. On the contrary, in all the logics in this chapter that are expressive enough to include *both* knowledge and belief operators, the above-mentioned identity is *provably wrong.*

We have already mentioned that we are interested in the broader concept of *justifiable belief.* This is broader in the sense that we consider an even more ideal agent, someone who reasons perfectly and effortlessly. (Technically, this means that we are going to ignore the fact that the agents we want to model are not *logically omniscient.* So justifiable belief can be regarded as a modeling of logically omniscient agents immune from Gettier-type problems.)

In addition, justifiable belief for us diverges from knowledge in the sense that it need not imply truth. As with Gettier examples, if one accepts and uses misinformation, then whatever conclusions are drawn are in some sense "justified." We postpone a fuller discussion of this point until later, but we wanted to alert the reader who expects us to write justifiable *true* belief for what we study.

Since the topic of the chapter is "epistemic logic", and since we were quick to point out that it is mistaken to identify knowledge with (true) justifiable belief, the reader may well wonder: why are they writing about it?

We claim that the study of justifiable belief is in itself illuminating with respect to the nature of knowledge, even if it fails as a satisfactory proposal for knowledge. This is the main reason why people have worked in the area. The main contributions are technical tools that allow one to make reliable predictions about complicated *epistemic scenarios*, stories about groups of agents which deal with who knows what about whom, etc. We shall go into more detail on what this means in Section 2 just below, and then in Section 5 we shall see how it works in detail.

Getting back to our study overall, one could think of it as a first approximation to the more difficult studies that would be desirable in a formal epistemology. It is like the study of motion on a frictionless plane: it is much easier to study the frictionless case than that of real motion on a real surface, but at the same time the easier work is extremely useful in many situations. We also point out that our subject has two other things going for it. First, it is a completely formalized subject with precise definitions, examples, and results. We know that not all readers will take this to be a virtue, and we have tried hard to introduce the subject in a way that will be friendly to those for whom logic is a foreign language. But we take the formalized nature of the subject to be an attraction, and so we aim to convey its nice results. Second, in recent years the subject has concentrated on two phenomena that are clearly of interest to the project of epistemology and which for the most part are peripheral in older treatments of epistemic logic. These are the *social* and *dynamic* sides of knowledge. The modeling that we present puts these aspects in the center.

At the time of this writing, it seems fair to say that the subject matter of the first part of our chapter, modeling based on justifiable belief, is fairly advanced. There are some open issues to be sure, and also outstanding technical questions. Many of the people involved in the area have gone on to adjacent areas where the insights and technical machinery may be put to use. Two of these are *combinations of logic and game theory* and *belief revision theory.* We are not going to discuss logic and game theory in this chapter, but the last section of our chapter does present proposals on the extension of dynamic epistemic logic to the area of belief revision.

In addition, as the subject moves closer to belief revision, it is able to question the notion of justifiable belief, to develop a more general notion of *conditional belief*, and also to meaningfully distinguish between *various types of "knowledge"* and characterize them in terms of conditional belief. These and similar formal notions should appeal to the epistemologist as well.

Overview Our overarching goal is that this chapter make the case for its subject to the uninitiated. We begin work in Section 2 with discussion of a series of epistemic scenarios.

These discussions illustrate the subject of the chapter by example, rather than by a direct discussion. They also are a form of "on-the-job-training" in the kinds of logical languages and representations that will be found later. This leads to a collection of issue areas for the subject that we present briefly in Section 3. Following that, we have some background in logic in Section 4. Even there, we are not only offering a catalog of a million logical systems: we attempt to say *why* the philosophically-minded reader might come to care about technical results on those systems. Dynamic epistemic logic (**DEL**, for short) is our next topic. This is the part of the chapter with the most sustained technical discussion. After this, we end the chapter with our look at belief revision theory. The proposals there are foreshadowed in Section 2, and a reader mainly interested in belief revision probably could read only Sections 2 and 7. It goes without saying that such readers should also read Hans Rott's Chapter 4c on the subject in this handbook. This is also true for readers whose main interest is in epistemology.

All readers would do well to consult Johan van Benthem and Maricarmen Martinez' Chapter 3b in order to situate our work even more broadly in studies of information modeling, especially the contrast and blending of the correlational and proof theoretic stances on the topic. In their terminology, however, the work in this chapter is squarely in the "information as range" paradigm.

The material in this chapter alternates between its main thrust, an examination of the philosophical and conceptual sides of the subject, and the technical aspects. We have chosen to emphasize the philosophical material because it is the subject of this handbook; also, there already are several introductions to the technical material. Historical pointers are mainly to be found at the ends of the sections.

2 Introduction: Logical Languages and Representations

As with many works in the area, we begin with an *epistemic scenario.* The one here is probably the simplest possible such scenario, an agent ignorant of which of two exclusive alternatives holds.

A person named Amina enters a room and is then shown a closed box on a table. She has been told that box contains a coin, and that the coin lies flat in the box. What she does not know is whether the coin lies heads up or tails up.

Our tasks as modelers are (1) to provide an adequate representation of this scenario; (2) to use the representation as part of a formal account of "knowledge" and related terms; (3) to see where the representation and formal account run into problems; (4) to then "scale up" all of the previous points by considering more complicated scenarios, models, and accounts, with the same goals in (1)–(3).

The most natural representation is simply as a set of two alternatives. In pictures, we have

The two circles are intended as abstract representations of the two states of the coin. There is no significance to the symbols H (for heads) and T (for tails, but please do not confuse it with *truth*). There is also no significance to the fact that the representation has heads on the left and tails on the right. There is a very real significance to the fact that each circle has exactly one symbol. There is *some* significance to the absolutely symmetric treatment of the two alternatives. Perhaps the most important aspect of the representation is that it leaves out everything to do with Amina's state of mind: why she thinks that heads and tails are the only ones possible, her prior experience with similar situations, her emotions, etc. For the most part, the formal work of this chapter will not help with proposals on any of these important matters precisely because the representations abstract from them.

We regard the symbols H and T as *atomic propositions* (We also call them *atomic sentences*, using the two terms interchangeably.) It is problematic at this point to speak of these as true or false in our scenario: since the story was silent on the matter of whether the coin was, in fact, lying heads up or tails up, it is debatable whether there is a "fact of the matter" here or not. No matter how one feels on where there is, in fact, a fact or not, it is less controversial to hold that either the coin lies heads up or tails up, and not both. (Recall that this is part of what Amina has been told at the outset.) It is natural to use standard propositional logic, and to therefore write $\mathsf{H} \leftrightarrow \neg\mathsf{T}$. (We may read this as "heads holds just in case tails fails".) We would like this sentence $\mathsf{H} \leftrightarrow \neg\mathsf{T}$ to come out true on our representation, and so clearly we need a semantics for sentences of this type.

Even before that, we need a formal language. We take propositional logic built over the atomic propositions H and T. So we have examples such as the one we just saw, H and T, and also $\neg\neg\mathsf{H}$, $\mathsf{H} \rightarrow (\mathsf{H} \rightarrow \mathsf{T})$, $\mathsf{H} \vee \mathsf{T}$, etc. The language is built by recursion in the way all formal languages are. We'll use letters like φ for propositions of this and other languages.

In order to give the semantics, it will be very useful to change the representation a little. We had used H and T inside the circles, but this will get in the way; also, as we shall see many times in this chapter, the states in our representations are not individuated by the atomic facts that they come with. So let us change our representation to

with the additional information that H holds at s and not at t, and T holds at t and not at s. We take this extra piece of information to be part of our representation. So we have a set $\{s,t\}$ of two *(abstract) states* and some extra information about them. The set $\{s,t\}$ has four subsets: $\emptyset$, $\{s\}$, $\{t\}$, and $\{s,t\}$ itself. We also have the usual set theoretic operations of the union of two subsets ($x \cup y$), intersection ($x \cap y$), and relative complement ($\overline{x}$). To spell out the details of relative complement in this example: $\overline{\emptyset} = \{x,y\}$, $\overline{\{s\}} = \{t\}$, $\overline{\{t\}} = \{s\}$, and $\overline{\{s,t\}} = \emptyset$.

Now it makes sense to formally interpret our language, assigning a set of states $\llbracket\varphi\rrbracket$ to a sentence φ as follows:

$$\begin{array}{lcl@{\qquad}lcl}
\llbracket\mathsf{H}\rrbracket & = & \{s\} & \llbracket\varphi\wedge\psi\rrbracket & = & \llbracket\varphi\rrbracket \cap \llbracket\psi\rrbracket \\
\llbracket\mathsf{T}\rrbracket & = & \{t\} & \llbracket\varphi\vee\psi\rrbracket & = & \llbracket\varphi\rrbracket \cup \llbracket\psi\rrbracket \\
\llbracket\neg\varphi\rrbracket & = & \overline{\llbracket\varphi\rrbracket} & \llbracket\varphi\rightarrow\psi\rrbracket & = & \overline{\llbracket\varphi\rrbracket} \cup \llbracket\psi\rrbracket \\
 & & & \llbracket\varphi\leftrightarrow\psi\rrbracket & = & (\llbracket\varphi\rrbracket \cap \llbracket\psi\rrbracket) \cup (\overline{\llbracket\varphi\rrbracket} \cap \overline{\llbracket\psi\rrbracket})
\end{array}$$

The reader will know that we could have given only a few of these, leaving the rest to reappear as derived properties rather than the official definition. The choice is immaterial. What counts is that we have a precise definition, and we can verify important properties such as $\llbracket\mathsf{H} \leftrightarrow \neg\mathsf{T}\rrbracket = \{s,t\}$. The reason is that

$$(\llbracket\mathsf{H}\rrbracket \cap \llbracket\neg\mathsf{T}\rrbracket) \cup (\overline{\llbracket\mathsf{H}\rrbracket} \cap \overline{\llbracket\neg\mathsf{T}\rrbracket}) \quad = \quad (\{s\} \cap \overline{\{t\}}) \cup (\overline{\{s\}} \cap \overline{\overline{\{t\}}}) \quad = \quad \{s,t\}.$$

We'll use S to refer to our set of states, both in this discussion and in later ones. And we shall say that φ is *valid in a model* if its semantic interpretation $[\![\varphi]\!]$ is the full set S of states, not merely a proper subset.

We have reliable and consistent intuitions concerning *knowledge*. Surely one feels that upon walking into the room, Amina does not know whether the coin lies heads or tails up: she was informed that there is a coin in the box, but so without further information to the contrary, she should not know which alternative holds. We expand our language by adding a knowledge operator K as a sentence-forming operation, making sentences from sentences the way $\neg$ does. We thus have sentences like $K\mathsf{H}$, $K\neg K\mathsf{T}$, etc. The semantics is then given by

$$[\![K\varphi]\!] \quad = \quad \begin{cases} S & \text{if } [\![\varphi]\!] = S \\ \emptyset & \text{if } [\![\varphi]\!] \neq S \end{cases}$$

Notice that this modeling makes knowledge an "all-or-nothing" affair. One can check that $[\![K\mathsf{H}]\!] = \emptyset$, matching the intuitions that Amina does not know that the coin lies heads up. But also $[\![K\neg\mathsf{H}]\!] = \emptyset$. In contrast, $K(\mathsf{H} \vee \neg\mathsf{H})$ is valid on this semantics: its interpretation is the entire state set.

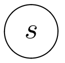
"Knowing that" and "knowing whether" Up until now, all of our modeling of knowledge is at the level of *knowing that* a given proposition, say φ, is true or false. We have no way of saying *knowing whether* φ holds or not. The easiest way to do this in our setting is to identify *knowing whether* φ with the disjunction *knowing that* φ or *knowing that* $\neg\varphi$. It will turn out that in this example and all of our other ones, this "or" is automatically an exclusive disjunction. That is, our modeling will arrange that no agents know both a sentence and its negation.

Iterated knowledge One should note that our formal semantics gives a determinate truth value to sentences with *iterated knowledge* assertions. For example, $K\neg K\mathsf{H}$ comes out true. (The reason: we saw above that $[\![K\mathsf{H}]\!] = \emptyset$. Therefore $[\![\neg K\mathsf{H}]\!] = \{s, t\}$, and so also $[\![K\neg K\mathsf{H}]\!] = \{s, t\}$.) Translating this back to our original scenario, this means that we are predicting that Amina knows that she doesn't know that the coin lies heads up. For a real person, this *introspectivity* is clearly false in general: though sometimes people can and do introspect about their knowledge, it seems that only a tiny amount of what we talk about people knowing is even susceptible to introspection. However, given our take on knowledge as justifiable belief (and with justifications modeled as surveys of all relevant possibilities), it fits. The upshot is that in the case of iterated knowledge assertions, the kind of modeling that we are doing gives predictions which are at odds with what real people do (though they are the acid test) but seem to work for ideal agents.

Note, however, that justifiable knowledge is a kind of *potential knowledge*: we would not feel that a reasoner who exhibited it was making a mistake. We would be more likely to commend them. Thus, the modeling that uses it is of value in *adversarial situations* of the kind found in game theory. When you reason about your opponent in a game (or war), you should not assume him to be stupid, but on the contrary: the safe option is to assume that he already knows everything that he could possibly know; i.e, to model him as a logically omniscient, fully introspective ideal agent. This is because you want to make sure your strategy works no matter how smart or how resourceful your opponent happens to be. (On the contrary, when reasoning about your own, or your allies', knowledge, it is safer not to idealize it, but to take into account possible failures of introspection.)

2.1 Learning

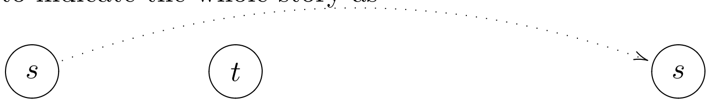
Suppose next that Amina opens the box and sees that the coin lies heads up. It is natural to assume that after she looks, she *knows* that the coin lies heads up. Furthermore, and in support of this, consider the model



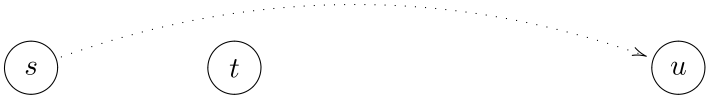
along with the information that s is a state where H is true. This model reflects the intuition that she considers only one state to be possible. Recall from Section 1 that our work here is mainly about knowledge as justifiable belief. It takes knowledge to result from a survey of the possible. The model above reflects this choice: it is a survey of the justifiable possible. (But a one-point model is so simple that it reflects other intuitions as well.)

What interests us most is that we have a *change of model*. In this case, the change was to throw away one state. Taking seriously the idea of change leads to *dynamics*, a key point of our study. Sometimes people with a little exposure to logic, or even a great deal of it, feel that logic is the study of eternal certainties. This is not the case at all. In the kinds of settings we are interested in here, we move from single models to *sequences* of them, or structures of some other kind. The particular kind of structure used would reflect intuitions about time, causality and the like. These matters are largely orthogonal to the epistemic modeling. But the important point for us now is that we can "add a dimension" to our models to reflect *epistemic actions*.

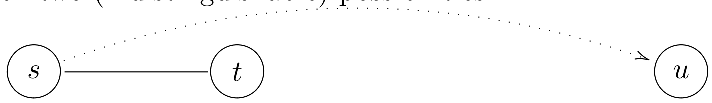
We would like to indicate the whole story as



We think of this as the two representations from before (along with the information that the picture suppresses, that s is a state where the coin lies heads up, and t the same for tails) connected by the dotted arrow. Amina discards the state t because it does not reflect what she learned. But s persists, and the dotted arrow indicates this. As it happens, it will again be confusing to use the same letter s for both the "before" and "after" states. This is not strictly needed here, but later in the paper it will be needed. So we would prefer to illustrate the story as



Again, we would supplement the picture with a description of which states have which coin faces showing: H is true at s and u, while T is true at t. Note that we *can no longer use the "all-or-nothing" notion of knowledge* from the previous section: in the original state s, Amina knows the state cannot be u (since she knows that she doesn't yet know the face of the coin); while in the new state u, Amina knows the state is neither s or t anymore. In other words, Amina *cannot distinguish* between the initial states s and t, but *can* distinguish between them and the new state u. We illustrate this by using lines to represent the agent's *indifference* between two (indistinguishable) possibilities:



The way to read an assertion like "there is a line between s and t" is as follows: Amina is indifferent in u between the world being as described in s and being described as in t. The agent is unable to tell apart these two descriptions: for all she knows, either of them can be a correct description of the real world. So in s, Amina thinks that the world might be t, or again it might be s itself. (This is not shown in the picture, but it is tacitly assumed.) In u, Amina thinks that u itself is the only possible state, and so she knows there that the coin lies heads up.

But what are these "states"? The above model has also more states. Since we have been using the word "state" quite a bit, a word is in order on this usage. Our states are the same as *possible worlds* in explanations of modality. That is, one should regard them as theoretical primitives that have an overall use in the modeling. They are abstract objects that we as outsiders use to discuss the examples and to further build a theory. Using them does not involve a commitment to their ontological or psychological reality. There is also a tradition of possible worlds as *(maximal consistent) sets of propositions*, and we also think of Carnap's state descriptions. But we do not follow either of these in our modeling, preferring to keep states as the primitive. It would be possible to change what we do to render states as maximal consistent sets, however, if one took the underlying language to be the full modal language which we are describing, not simply propositional logic. For our states are not individuated by the propositional facts holding in them: as we shall see shortly in (6) below, to separate states one needs the information contained in the arrows.

Knowledge The heart of the matter is the proposal for the semantics of knowledge. Building on our explanation of what the worlds and the lines represent, the idea is that Amina "knows" (in our sense) φ in a world x just in case the following holds: φ is true in all worlds that she cannot tell apart from x. In symbols, the semantics is given by:

$$[\![K\varphi]\!] \quad = \quad \{s : \text{whenever Amina is indifferent between } s \text{ and } t,\ t \in [\![\varphi]\!]\} \tag{1}$$

In other words: we relativize the previous "all-or-nothing" definition to the set of worlds that are indistinguishable from the real one. Using this modified definition, we can see that, in the initial state s of the above model, Amina doesn't know the face of the coin is H, while in the new state u she knows the face is H.

Comparison with probabilistic conditioning It will be useful at this point to note a similarity between the kind of updating that models Amina's learning in this very simple setting and what happens all the time in *probability*. Suppose that we have a *probability space*. This is a set S of *simple events*, together with a *probability* $p : S \to [0,1]$ with the property that $\sum_{s\in S} p_s = 1$. The probability space as a whole is $S = (S, p)$; that is, the set S together with the function p.

The subsets of S are called *events*. Every event X gets a probability $p(X)$, defined by $p(X) = \sum_{s\in X} p_x$. We should think of S as the states of some background system $\mathcal{S}$. An event X is then like a property of the system, and p_X as the probability that a randomly observation of $\mathcal{S}$ will have the property X. Not only this, but every event X whose probability is nonzero gives a probability space on its own. This time, the sample space is X, and the probability $p|X$ is given by $(p|X)(s) = p(s)/p(X)$. This formula reflects the re-normalization of the probability p. We'll call this new space $S|X = (X, p|X)$. It is a *subspace of the original* S *called* S *conditioned on* X. The idea is that if we again start with a system $\mathcal{S}$ whose set of states is modeled by the probability space S, and if we obtain additional information to the effect that the background system definitely has the property X, then we should revise our modeling, and instead take use $S|X$:

$$(S, p) \xrightarrow{\text{learning that } X} (S|X, p|X) \ \ .$$

The sentences in our formal language are notations for extensional properties (subsets) of the overall state space. Adding new information, say by a direct observation, corresponds to *moving to a subspace*, to changing the representation.

2.2 Another agent enters

Let us go back to our first scenario, where Amina walks into the room in ignorance of whether the coin lies heads or tails up. (We therefore set aside Section 2.1 right above.) Being alone in a room with a closed box is not much fun. Sometime later, her friend Bao walks over. The door is open, and someone tells Bao the state of the coin and does it in a way that makes it clear to Amina that Bao now knows it. At the same time, Bao is in the dark about Amina's knowledge.

The natural representation here uses four states.

$$\boxed{u{:}\mathsf{H}} \overset{b}{\text{——}} \boxed{v{:}\mathsf{H}} \overset{a}{\text{——}} \boxed{w{:}\mathsf{T}} \overset{b}{\text{——}} \boxed{x{:}\mathsf{T}} \qquad (2)$$

Note that now we have *labeled* the indifference lines with the names of the agents. In this case, we have four worlds called u, v, w, and x. The atomic information is also shown, and we have made the picture more compact by eliminating redundant information. (So we intend u to be a world where H is true and T is false, even though the second statement is not explicit.)

As before, the way to read an assertion like "there is a line labeled b between u to v" is as follows: Bao is indifferent in u between the world being as described in u and being described as in v. So in v, Amina thinks that the world might be w, or again it might be v itself. In u, Amina thinks that u itself is the only possible state, and so she knows there that the coin lies heads up.

The world u is the *real world*, and so once we have a formal semantics, we check our intuitions against the formal semantics at u.

We begin, as we did in Section 2, with the propositional language built from symbols H and T. In our current model, we have semantics for them:

$$[\![\mathsf{H}]\!] = \{u, v\} \qquad\qquad [\![\mathsf{T}]\!] = \{w, x\}.$$

We clearly need now two knowledge operators, one for a (Amina) and one for b. We shall use K_a and K_b for those, and we define by taking for each agent the appropriate indifference lines in the definition (1):

$$\begin{array}{lcl} [\![K_a\varphi]\!] & = & \{s : \text{whenever Amina is indifferent between } s \text{ and } t,\ t \in [\![\varphi]\!]\} \\ [\![K_b\varphi]\!] & = & \{s : \text{whenever Bao is indifferent between } s \text{ and } t,\ t \in [\![\varphi]\!]\} \end{array} \qquad (3)$$

We can check whether our formal semantics matches our intuitions about our model. The way we do this is by translating a sentence A of English into a sentence φ in our formal language, and evaluating the semantics. We want to be sure that $x \in [\![\varphi]\!]$, where x is the "real" or "actual" world in the model at hand.

Here are some examples:

English	Formal rendering	Semantics
the coin shows heads	H	$\{u, v\}$
a knows the coin shows heads	$K_a\mathsf{H}$	$\emptyset$
a knows whether the coin shows heads	$K_a\mathsf{H} \vee K_a\neg\mathsf{H}$	$\emptyset$
b knows that the coin shows heads	$K_b\mathsf{H}$	$\{u, v\}$
b knows whether the coin shows heads	$K_b\mathsf{H} \vee K_b\neg\mathsf{H}$	$\{u, v, w, x\}$
a knows that b knows whether the coin shows heads	$K_a(K_b\mathsf{H} \vee K_b\neg\mathsf{H})$	$\{u, v, w, x\}$

In our model, u is the "real world". In all of the examples above, the intuitions match the formal work.

But does there have to be a real world? Our representation in (2) and our semantics in (3) did not use a designated real world. So mention of a real world could be dropped. However, doing so would mean that we have less of a way to check our intuitions against the formalism (since our intuitions would be less sharp). But one who doesn't like the idea of a "real world" would then look at all the worlds in a representation, take intuitive stories for them, and then check intuitions in all cases.

Announcements We have already begun to discuss dynamics in connection with these simple scenarios. Here are ways to continue this one. Suppose that Amina and Bao go up and open the box. We again would like to say that the resulting model has a single world, and in that world both agents consider that world to be the only possible one. In a word (picture), we expect

$$\boxed{u{:}\mathsf{H}} \tag{4}$$

However, this is not quite what we get by the world-elimination definition which we have already seen. What we rather get is

$$\boxed{u{:}\mathsf{H}} \overset{b}{\text{——}} \boxed{v{:}\mathsf{H}} \tag{5}$$

(We have dropped the worlds w and x from the picture in (2), and all the lines pertaining to them.)

So we have a question at this point: can we say that the two models are equivalent, and can we do so in a principled way? We return to this point at the end of Section 4.4.

As an alternative, suppose someone outside simply shouts out "The coin lies Heads up." Again, on the modeling so far, we have the same state at the end. We thus conclude *our representations cannot distinguish sources of information.*

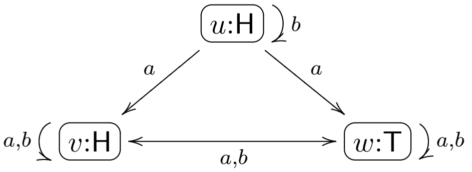
2.3 Another agent enters, take II

At this point, we want an alternative to the story in Section 2.2. Amina is again in the room in ignorance of the state of the coin. Bao walks over, but this time, the door is shut. Outside the door, some trusted third party says to him, "I'll tell *you* that the coin lies heads up." Then Bao walks in.

We naturally have some intuitions about what is going on. Bao should know that the coin lies heads up, and he should also know that Amina doesn't know whether the coin lies heads up or tails up. What about Amina? What should she think about Bao's knowledge? We don't know enough about her to say for sure, but to be definite, let us assume that she (falsely) believes that Bao is as ignorant as she. Moreover, let us assume that Bao believes *this* about Amina.

Belief Since we have mentioned belief, a comment is in order. We continue to deal with "knowledge" as justifiable belief. This is the only notion at play at the moment. We have used *belief* in the previous paragraph only to emphasize that the proposition believed is actually false.

At this point, we return to our scenario and to a model of these intuitions. The model we have in mind is



(6)

We shall shortly go into details about why this model works. Before that, a comment: If one tries to come up with a model by hand which reflects correctly the intuitions that we spelled out above, he or she will see that it is not a straightforward task. It is hard to argue that something is not easy to do, so we encourage readers to try it. This will also provide experience in the important task of examining putative models with the eye towards seeing whether they match intuitions or not.

This model generalizes the two-directional lines that we saw above to one-directional arrows. The way to read an assertion like "there is an arrow labeled a from u to v" is as follows: if the situation were modeled by u, then Amina would be justified in considering v a possibility.

In our example, u is the "real world". In that world, Bao is countenances no others. Amina, on the other hand, thinks that the world is either v or w. (But she does not think that the world u itself is even possible. This reflects the intuition that Amina doesn't think it possible that Bao knows the state of the coin.) The worlds she thinks are possible are ones pretty much like the two we saw earlier for her alone, except that we have chosen to put in arrows for the two agents.

Note that u and v in (6) have the same atomic information. However, they are very different because what counts is not just the information "inside" but also the arrows. Now given our explanation of what the epistemic arrows are intended to mean, we can see that there is some circularity here. This is not a pernicious circularity, and the use of the logical language makes this evident. Once we take the representation as merely a *site for the evaluation of the logical language*, the problematic features of the models lose their force. We turn to that evaluation now.

Building on our explanation of what the worlds now represent, we say that a *believes* φ in a world x just in case the following holds: φ is true in all worlds that she would think are possible, if x were the actual world. Formally:

$$\begin{array}{lcl} [\![B_a\varphi]\!] & = & \{s : \text{whenever } s \xrightarrow{a} t,\, t \in [\![\varphi]\!]\} \\ [\![B_b\varphi]\!] & = & \{s : \text{whenever } s \xrightarrow{b} t,\, t \in [\![\varphi]\!]\} \end{array} \tag{7}$$

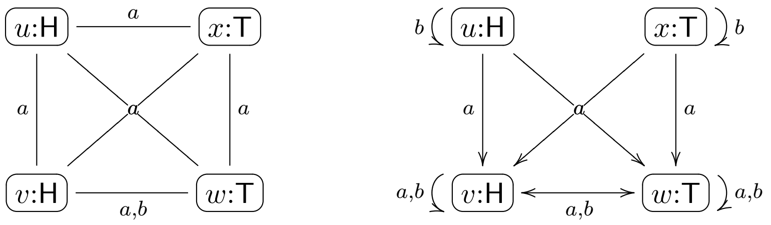
We again check that our semantics and model are sensible, going via examples.

English	Formal rendering	Semantics
the coin shows heads	H	$\{u, v\}$
a knows (believes) the coin shows heads	$B_a\mathsf{H}$	$\emptyset$
a believes the coin shows tails	$B_a\mathsf{T}$	$\emptyset$
b believes the coin shows heads	$B_b\mathsf{H}$	$\{u\}$
b believes that a doesn't know (believe) it's heads	$B_b\neg B_a\mathsf{H}$	$\{u, v, w\}$
b believes that a believes that b doesn't know (believe) it's heads	$B_bB_a\neg B_b\mathsf{H}$	$\{u, v, w\}$

In our model, u is the "real world". In all of the examples above, the intuitions match the formal work.

Knowledge and Belief What does Amina actually *know* in the scenario above? She believes that Bao doesn't know the face of the coin, but this belief is not true. Is Amina aware of the possibility that her belief is false? Let us assume so: in other words, although she *believes* that Bao does not know the face, she at least countenances the possibility that he does. Note that the states u, v and w are *indistinguishable* for her: she "sees" the same things, and has the same information and the same beliefs in all these states. But then these states also are indistinguishable for her from a *fourth* possibility, namely the one in which Bob knows the face but the coin shows tails. All the information that Amina has is consistent with this fourth possibility, so she cannot exclude it either.

To distinguish between belief and (truthful) knowledge, we supplement what we so far have, in a few ways. First, we need to add to the state space a fourth state, representing the fourth possibility just mentioned. Second, we consider *two models* on the same state set. The one on the left below is intended to model *knowledge*, while the one on the right is intended for *belief*:



(8)

The real world is u. On the side of knowledge, Amina is indifferent between all the states. The difference between the belief model and the one in (6) is that we have added a fourth state, x. This state is *inaccessible* from u in the belief model, and so it will turn out to be irrelevant from the point of view of the agent's beliefs in the actual world; however, x will be crucial in dealing with knowledge and *conditional belief* in the real world. (Concerning knowledge, leaving x off would mean that Amina in the real world knows there is no world in which Bao knows that the coin is tails up. This would give her knowledge beyond what is justified by our story.) Recall that the lines (as on the left) are the same as two-way arrows. So we can see that all of the arrows in the right diagram (for belief) are also present in the left diagram (for knowledge). This is good: it means that everything the agents know in a model will also be believed by them. To make this precise, we of course need a formal semantics. Let us agree to write $\overset{a}{\sim}$ for the knowledge arrows (indifference lines), and $\xrightarrow{a}$ for the belief ones.

In fact, it is natural to consider loops as being implicitly present in the knowledge model, so we put $s \overset{a}{\sim} t$ iff either $s = t$ or there is an indifference line between them. The relevant definitions (stated only for Amina) will then be

$$\begin{array}{lcl} [\![K_a\varphi]\!] & = & \{s : \text{whenever } s \overset{a}{\sim} t,\, t \in [\![\varphi]\!]\} \\ [\![B_a\varphi]\!] & = & \{s : \text{whenever } s \xrightarrow{a} t,\, t \in [\![\varphi]\!]\} \end{array} \tag{9}$$

On this semantics, then, Amina will believe that Bao does not know the state of the coin, but also, she does not *know* this.

Observe now that the two models for this situation are not independent: *the belief model contains all the information about the knowledge model.* Indeed, we can recover the knowledge model from the belief model by *closing the belief arrows under reflexivity, symmetry and transitivity.* Visually, this amounts to replacing in the model on the right all the one-way arrows by lines, adding loops everywhere and adding lines between any two states that are connected by a chain of lines. A simpler alternative procedure is to connect any two states by lines if and only if the same states are reachable from both via (one-step) belief arrows. This gives us the knowledge model on the left.

We know that there are issues involved in the translation that we are ignoring. One point worth mentioning is that translating beliefs regarding *conditionals* is problematic (and this is why none of the sentences in the table are conditionals). The reason is that the formal language suggests the material conditional, and so the mismatches between any natural language conditional and the material conditional are highlighted by the process we are suggesting.

2.4 Conditional beliefs

Consider again the belief-knowledge model (8). Where this model goes wrong is in dealing with conditional assertions that are *counterfactual with respect to the agents' beliefs.* Consider the following statements:

1. If Bao knows the state of the coin, then the coin lies either heads up or tails up.

2. If Bao knows the state of the coin, then the coin lies heads up.

3. If Bao knows the state of the coin, then Amina does, too.

We are interested in whether *Amina believes any of these statements.* As such, these are *conditional belief* assertions. Intuitively, she should believe the first statement, but not the second and third. Yet, they are all true on the definition of belief in (9), if we interpret conditional beliefs as beliefs in the conditional and we interpret the conditionals as material conditionals.

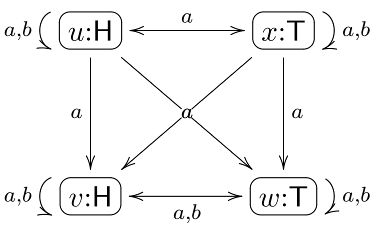
In fact, the problem is not simply the use of the material conditional: no other "belief-free" notion of conditional would do either! As argued in e.g. Leitgeb [54], it is not possible to separate a conditional belief into a doxastic part (the "belief") and a belief-free part (the "conditional" that forms the content of the "belief"). Gärdenfors' Impossibility Theorem[1] can be understood as showing that, under reasonable assumptions, *conditional beliefs are not equivalent to beliefs in conditionals, for any belief-free notion of conditional.* As a consequence, we have to treat conditional belief as one indivisible operator $B_a^\alpha \varphi$ instead of a composed

[1] See Section 7 and Hans Rott's Chapter 4c in this handbook for.

expression $B_a(\alpha \Rightarrow \varphi)$.[2] But the above models are not adequate to give a semantics to this operator. On a technical level, the problem is that to make the sentence *Amina believes that Bao doesn't know* come out true we need a belief model (such as the one above (8)) in which u and x are *not* to be accessible for Amina from u; but at the same time, for evaluating hypothetical statements like the conditionals above, we need to use a different belief model, one in which u and x become accessible from u.

There are several ways of getting an appropriate model, but all of them involve going beyond simple belief models. We are going to present one common modeling, essentially derived from work of Lewis, Grove, and others. We supplement our model with a *Grove system of spheres*, exactly as discussed and pictured in Section 2.3 of Chapter 4c. We need one for each agent at each state. We'll only spell out what Amina's system looks like at u. It would have v and w in the center, since she is most committed to them. In the next ring, we put u and (crucially) x. As with all such systems of spheres, the definition gives us notions of which worlds are *at least as plausible* as others, and *strictly more plausible* as others. According to the above system of spheres, states v and w are equally plausible for Amina, and they are strictly more plausible than the other two states u, x, which are also themselves equally plausible.

If we draw arrows from any given state to all the states that are at least as plausible as it, we obtain the following diagrammatic representation:



(10)

This is a *plausibility model*: it doesn't directly capture knowledge or beliefs, but only doxastic plausibility. However, *this plausibility model contains all the information about the belief and knowledge models.* Indeed, we can recover the belief model by looking at the *most plausible states* (for each agent); i.e., the states which can be reached via some (one-step) plausibility arrow from any other state that is reachable from them (via a one-step arrow). To obtain the belief model in (8), we keep for each agent only the arrows pointing to that agent's most plausible states. We can then recover the knowledge model from the belief model as in the previous section. Or we can directly obtain the knowledge model in (8) from the above plausibility model, by simply replacing all the arrows by lines (and deleting the loops).

We now *rework the definition of conditional belief.* Actually, we keep track of the antecedent of the conditional in a special way, and write $B_a^\alpha \chi$ to symbolize *Amina believes that were α to be true, χ would have been true as well.* The formal definition is:

$$\begin{aligned} [\![B_a^\alpha \chi]\!] \quad = \quad & \{s : t \in [\![\chi]\!], \text{ for all } t \in [\![\alpha]\!] \text{ such that } s \overset{a}{\sim} t \\ & \quad \text{and such that there is no } u \in [\![\alpha]\!], \text{ such that } s \overset{a}{\sim} u \\ & \quad \text{and } u \text{ is strictly more plausible than } t \text{ for Amina}\} \end{aligned} \tag{11}$$

(And similarly for $B_b^\alpha \chi$, of course.)

[2] In Chapter 3b, this expression would be written $B_a(\varphi|\alpha)$.

The idea is that in evaluating a conditional whose antecedent α contradicts the current beliefs, one has to discard the most plausible worlds, and instead to "fall back" to the worlds that are most plausible given α.

Let us see how this works in our example. For Amina, x is at least as plausible as the real world u. So she should use this world in evaluating conditionals, along with others. This easily shows why sentences 2 and 3 in the beginning of this subsection come out false.

Incidentally, one desirable property of this kind of modeling is that an agent's *knowledge* (as opposed to belief) should not be overridden, even in hypothetical contexts. (This will not be suitable to modeling conditionals which are *counterfactual with respect to knowledge.*) To arrange this, we should require that the union of all spheres for a given agent in a given state coincides with the $\sim$-equivalence class of the agent there.

Modern epistemic logic started to flourish after modal logic (with its roots in Aristotle) was formalized and given a possible world semantics. It is hard to track down the exact origins of this semantics, but it is widely known as Kripke semantics, after Kripke, who devoted a number of early papers to the semantics of modal logic [50]. A contemporary and thorough reference for modal logic is the monograph [20].

Observations on how epistemic states change as a result of new information have been around since the Hintikka founded the field of epistemic logic in his 1962 book *Knowledge and Belief* [45] (republished in 2005 by King's College, London). Hintikka is broadly acknowledged as the father of modern epistemic logic, and his book is cited as the principal historical reference. Hintikka himself thinks that von Wright [117] deserves these credits.

From the late 1970s, epistemic logic became subject of study or applied in the areas of artificial intelligence (as in R.C. Moore's early work [71] on reasoning about actions and knowledge), philosophy (as in Hintikka's [46]), and game theory (e.g. Aumann [6]). In the 1980s, computer scientists became interested in epistemic logic. In fact, the field matured a lot by a large stream of publications by Fagin, Halpern, Moses and Vardi. Their important textbook *Reasoning about Knowledge* [30] which appeared in 1995, contains the contents of many papers co-authored by (subsets of) them over a period of more than ten years. Another important textbook in both 'pure' and 'applied' epistemic logic is Meyer and van der Hoek [66]. These both should be consulted in connection with Sections 1–4 of this chapter. The work from Section 5 onward (on dynamic epistemic logic and its extensions) is mainly newer than those books. A brief, but very good, introduction to the history, the philosophical importance and some of the technical aspects of epistemic logic is the chapter "Epistemic Logic", by Gochet and Gribomont, in the *Handbook of History of Logic* [40]. It also gives a very brief look at some of the older work in dynamic epistemic logic.

At the same time as computer scientists became interested in the topic, linguistic semanticists were also discovering many of the basic issues, such as effects of public announcements and the problem of belief revision. Of special mention here is the long work of Robert Stalnaker, whose longstanding involvements with knowledge and belief revision theory include publications such as [86, 87, 88].

3 Further Issues and Areas in Epistemic Logic

At this point, we have said a little about the subject matter of the chapter. Before we go further, we mention a few issues, problems, puzzles, and discussion topics in the area. We especially comment on how they relate to the examples in the previous sections.

3.1 The Muddy children

Perhaps the most common of epistemic puzzles is one known in various guises and under names like *the muddy children, the wise men* or *the unfaithful spouses* [34, 72]. Here is one version of it. A number of children have been playing outside. After some time, some of them might have mud on their foreheads; however, they don't discuss this with one another. But along comes one of their fathers, and says:

"At least one of you has mud on his/her forehead. Do you know if you are muddy?"

Let n be the number of who have muddy foreheads. If $n = 1$, the one muddy one sees the clean heads of the others and deduce that she herself is muddy.

Otherwise, all reply (in unison) "No, I don't know." At this point, the father again asks the same question. If $n = 2$, the two muddy ones would know see each other and know that $n \geq 1$, simply because the other did not answer Yes the first time. So they would know on the second questioning.

The story continues in this way. The mathematical upshot is that if there are n muddy children to start, then after the father asks his question n times, the muddy ones will know their status; and before the nth time, nobody will know it. The essential feature for us of the story is that it illustrates that *statements about ignorance can lead to knowledge.*

Comparing to the content of this chapter, it is not hard to draw the representations for this problem and for variations, and to watch the process of announcement (as we have seen it in Section 2.2). Indeed, we have found these to be excellent sources of exercises in modal logic. We shall see the formal logical systems related scenarios like this.

Incidentally, we saw above that statements can change an agent's knowledge. It is even possible to find a setting where an agent can believe something at the outset, then someone else's statement causes them to lose this belief, and then a third statement to regain it. We present an example in Section 6.1.

3.2 Logical omniscience

Logical omniscience is the phenomenon whereby an agent's beliefs or knowledge are modeled in such a way that they are closed under logical deduction. So the agent knows (or believes) all the consequences of their knowledge, and in particular knows infinitely many sentences, theorems whose length or complexity are absurdly great, etc. Logical omniscience is thus a *complaint* against all of the kinds of models we are considering in this chapter. To avoid the complaint, one must adopt much more fine-grained models. A number of different such models have been proposed: a logic of *awareness* ([58], further extended by Fagin and Halpern), multi-valued epistemic logic (A. Wisniewski [122]), doxastic linear logic (D'Agostino, Gabbay and Russo [25]), resource-bounded belief revision (R. Wassermann [119], [120]) etc. A solution using a new type of dynamic-epistemic logic was proposed by Ho Ngoc Duc [28].

3.3 The Gettier challenge

Gettier [39] pointed out examples that effectively jettisoned the justified true belief analysis of knowledge. The ensuing discussions are central to modern epistemology. For an overview of the area, see, e.g., Steup [90].

Perhaps the easiest related example in epistemic logic is the following. Consider a muddy children scenario with two children, say A and B. A is muddy and B clean. A parent announces that at least one is muddy, asks if the two know their state. Usually, A would announce affirmatively and B negatively, but this time let A lie and say that she does not; B of course truthfully replies that he doesn't know. Then on second round, both announce that they do know. The point is, that B's announcement is truthful: taking knowledge to be justifiable true belief, he will have some knowledge of his state after hearing A once, no matter what she says. B's announcement is also justified, being based on A's first declaration. At that point, B has a justified true belief that he knows his state. But we would not judge B to actually know whether he is dirty or not. This would mean either knowing that he is dirty, or knowing that he is clean: he thinks he knows the former and denies he knows the latter.

3.4 Other notions of knowledge

The Gettier examples have been used, among other things, to deny the validity of the Negative Introspection axiom for knowledge: in the example in Section 3.3, B *thinks that he knows* his state, but intuitively speaking we can't agree that he actually knows it. So agents may not know something, while believing that they know it.

Various authors proposed dropping the characteristic $S5$ axiom (Negative Introspection), and sticking with the system $S4$ instead. For instance, the $S4$ principles were as far as Hintikka [45] was willing to go. This may also be appropriate in an *intuitionistic context*, and also fit well with a *topological interpretation* of knowledge. For other work on appropriate axioms, see Lenzen [56, 57].

The defeasibility analysis of knowledge We only look here at one of the alternative proposals for a knowledge concept, that fits well with our discussion of conditional beliefs in Section 2.4. This is the "defeasibility strategy", followed by many of those who attempted to respond to Gettier's challenge, see e.g Lehrer and Paxson [53], Swain [91], Stalnaker [87, 88]. To quote Stalnaker [88], "the idea was that the fourth condition (to be added to justified true belief) should be a requirement that there would be no 'defeater' - no true proposition that, if the knower learned that it was true, would lead her to give up the belief, or to be no longer justified in holding it". One way to do this is to add to the semantics of belief a theory of *belief revision*, and then define knowledge as belief that is stable under any potential revision by a true piece of information. But as we shall see, *conditional beliefs* and plausibility models, introduced in Section 2.4, give us a semantics for belief revision. So it is not surprising that defeasible knowledge was formalized using a logic of conditional beliefs, as in [21] and [17], or a logic of conditionals [87].

Knowledge and "safe belief" However, the notion of knowledge defined on plausibility models in Section 2.4 is *stronger* than the one of (true, justifiable) defeasible belief. As we shall see, it corresponds to a belief that is "absolutely unrevisable": it cannot even be defeated by revising with *false* information. Since we followed the common usage in Computer Science and called "knowledge" this strong, absolute notion, we shall follow Baltag and Smets [16, 17] and call *safe belief* the weaker notion resulting from the defeasibility analysis. In [87, 16, 17], this concept is applied to reasoning about solution concepts in Game Theory.

3.5 Moore sentences

By a *Moore sentence* we mean one of the form 'p is true and I don't believe that', or 'p is true and I don't know that'. Moore's "paradox" is that such a sentence may well happen to be *true*, but it can never be *truthfully asserted*: a person uttering this sentence *cannot believe it*. As this example crops up in very different settings, and as it is so crucial for a proper understanding of dynamic epistemics, we discuss its origin in some detail, as a proper historical primer to the subject area. In this discussion, $B\varphi$ means "I believe φ " and $K\varphi$ means "I know φ ".

Moore writes that if I assert a proposition φ, I *express* or *imply* that I *think* or *know* φ, in other words I express $B\varphi$ or $K\varphi$. But φ cannot be said to *mean* $B\varphi$ [68, p.77] as this would cause, by substitution, an infinite sequence $BB\varphi$, $BBB\varphi$, ad infinitum. "But thus to believe that somebody believes, that somebody believes, that somebody believes ... quite indefinitely, without *ever* coming to anything which is what is believed, is to believe nothing at all" [68, p.77]. Moore does not state in [68] (to our knowledge) that $\varphi \wedge \neg B\varphi$ cannot be believed. In Moore's "A reply to my critics", a chapter in the 'Library of Living Philosophers' volume dedicated to him, he writes " 'I went to the pictures last Tuesday, but I don't believe that I did' is a perfectly absurd thing to say, although *what* is asserted is something which is perfectly possibly logically" [69, p.543]. The absurdity follows from the implicature 'asserting φ implies $B\varphi$' pointed out in [68]. In other words, $B(p \wedge \neg Bp)$ is 'absurd' for the example of factual information p. As far as we know, this is the first full-blown occurrence of a Moore-sentence. Then in [70, p.204] Moore writes " 'I believe he has gone out, but he has not' is absurd. This, though absurd, is not self-contradictory; for it may quite well be true."

Hintikka [45] mentions the so-called 'Moore'-problem about the *inadequacy of information updates with such sentences*. This leads us to an interesting further development of this notion, due to Gerbrandy [36], van Benthem [97] and others. This development, addressed in our contribution, firstly puts Moore-sentences in a *multi-agent* perspective of announcements of the form 'I say to you that: p is true and that *you* don't believe that', and, secondly, puts Moore-sentences in a *dynamic* perspective of announcements that cannot be believed after being announced. This analysis goes beyond Moore and makes essential use of the tools of dynamic epistemic logic. The dynamic point of view asks how an agent can possibly *come to believe* (or know) that a Moore sentence φ is true. The only way to achieve this seems to be by *learning* φ, or by learning some other sentence that implies φ. But one can easily see that, when φ is a Moore sentence, *the action of learning it changes its truth value*: the sentence becomes false after being learned, though it may have been true before the learning! The same applies to any sentence that implies φ. In terms of [36], an update with a Moore sentence can never be "successful": indeed, in Section 5.2, a *successful formula* is defined as one that is *always* true after being announced. Observe that Moore sentences have the opposite property: they are "strongly un-successful", in the sense that they are *always* false after being announced. As a consequence, they are *known* to be un-successful: once their truth is announced, their negation is known to be true. Van Benthem [97] calls such sentences *self-refuting*.

There is nothing inherently paradoxical about these properties of Moore sentences: the "world" that a Moore sentence is talking about is not simply the world of facts, but a "world" that comprises the agent's own beliefs and knowledge. In *this* sense, *the world is always changed by our changes of belief*. Far from being paradoxical, these phenomena can in fact be formalized within a consistent logic, using e.g. the logic of public announcements in Section

5.1: using the notation introduced there, $!\varphi$ is the action of learning (or being announced) φ. If φ is a Moore sentence of the form $p \wedge \neg Kp$, it is easy to check the validity of the dynamic logic formulas $[!\varphi]\neg\varphi$ and $[!\varphi]K\neg\varphi$. The first says that Moore sentences are strongly un-successful; the second says that Moore sentences are self-refuting. As argued in [97], self-refuting sentences are essentially *un-learnable*. This explains why a Moore sentence can never be known or believed: because it can never be learned.

A similar analysis applies to the doxastic versions $p \wedge \neg Bp$ of Moore sentences. But, in the case of belief, this phenomenon has even more far-reaching consequences: as pointed out by van Ditmarsch [107] and others, the un-successfulness of Moore sentences shows that the standard **AGM** postulates for belief revision (in particular, the "Success" postulate) cannot accommodate higher-order beliefs. This observation leads to the distinction, made in the dynamic-epistemic literature [98, 15, 19], between "static" and "dynamic" belief revision. As shown in Section 7.2, in the presence of higher-order beliefs the **AGM** postulates (even in their multi-agent and "knowledge-friendly" versions) apply only to *static* belief revision.

3.6 The Knower paradox

Related to the phenomenon of Moore-sentences is what comes under the name of 'paradox of the knower', also known as Fitch's paradox [23]. The general verification thesis states that *everything that is true can be known* to an agent; formally, if we introduce a modal possibility $\Diamond\varphi$ to express the fact that something *can* be achieved (by an agent), this says that the implication $\varphi \to \Diamond K\varphi$ is true (in our world), for all formulas φ. The following argument, due to Fitch, appears to provide a refutation of verificationism on purely logical grounds. Take a true Moore sentence φ, having the form $\psi \wedge \neg K\psi$. By the "verificationist" implication above, $\Diamond K\varphi$ must be true. But then $K\varphi$ must be true at some possible world (or some possible future "stage", achievable by the agent). But, as we have already seen in the previous subsection, this is impossible for Moore sentences: $K(\psi \wedge \neg K\psi)$ is inconsistent, according to the usual laws of epistemic logic. The only possible way out is to conclude that *there are no true Moore sentences*; in other words, the implication $\psi \to K\psi$ holds for all formulas. This simply trivializes the verificationist principle, by collapsing the distinction between truth and knowledge: all truths are already known!

Numerous solutions for this paradox have been proposed; see [118, 92], for example. In particular, Tennant [92] argues persuasively that the verificationist principle should be weakened, by restricting its intended applications only to those sentences φ for which $K\varphi$ is consistent. In other words: if φ is true and if it is logically consistent to know φ, then φ can be known. This excludes the principle's application to Moore sentences of the usual type.

An interesting take on this matter is proposed by van Benthem in [97]: one can interpret the modal possibility operator $\Diamond\varphi$ above in a dynamic sense, namely as the 'ability' to achieve φ by performing some learning action, e.g. an announcement in the technical sense of Section 5, to follow. In other words, 'φ is knowable' is identified with 'a true announcement can be made after which the agent knows φ.' In this interpretation, the above-mentioned verificationist thesis reads "*what is true may come to be known* (after some learning)", while its Tennant version restricts this to sentences φ such that $K\varphi$ is consistent. The Moore-sentences are obviously unknowable (by the agent to whose knowledge they refer). But van Benthem [97] shows that this interpretation is also incompatible with Tennant's weakened verificationist principle: in other words, there are sentences φ such that $K\varphi$ is consistent but still, $\varphi \to \Diamond K\varphi$ does not hold. A counterexample is the formula $(p \wedge \Diamond\neg p) \vee K\neg p$. The dynamic epistemic logic of the 'ability' modality $\Diamond$ is completely axiomatized and thoroughly studied in [7].

3.7 The Hangman paradox

The Hangman paradox, also known as the Surprise Examination paradox, has a relatively short history of about sixty years. Apparently the Swedish mathematician Lennart Ekbom heard a message on the radio during the second world war announcing a civil defense exercise, which was to take place in the next week. It was also announced that this exercise would be a surprise. Then he noticed that there was something paradoxical about this announcement. [51, 84, pp.253]. The paradox was first published by O'Conner in 1948.

> Consider the following case. The military commander of a certain camp announces on a Saturday evening that during the following week there will be a "Class A blackout". The date and time of the exercise are not prescribed because a "Class A blackout" is defined in the announcement as an exercise which the participants cannot know is going to take place prior to 6.00 p.m. on the evening in which it occurs. It is easy to see that it follows that the exercise cannot take place at all. It cannot take place on Saturday because if it has not occurred on the first six days of the week it must occur on the last. And the fact that the participants can know this violates the condition which defines it. Similarly, because it cannot take place on Saturday, it cannot take place on Friday either, because when Saturday is eliminated Friday is the last available day and is, therefore, invalidated for the same reason as Saturday. And by similar arguments, Thursday, Wednesday, etc., back to Sunday are eliminated in turn, so that the exercise cannot take place at all. [74]

Many authors proposed various solutions to this paradox. Williamson [121] analyzes it as an epistemic variety of the Sorites paradox. The first analysis that uses dynamic epistemic logic was presented in [36], and found its final form in [108] and [37]. According to Gerbrandy, the commander's statement is ambiguous between two possible readings of what "Class A" means: the first reads "You will not know (before 6:00 on the evening of the blackout) when the blackout will take place, *given* (the information you have in) *the situation as it is at the present moment*", while the second reads "You will not know (before 6:00 of that evening) when it will take place, *even after you hear my announcement.*" Gerbrandy chooses the first reading, and shows that in fact there is no paradox in this interpretation, but only a Moore-type sentence: in this reading, the property of being "Class A" cannot remain true after the announcement. Unlike the previous puzzles however, there is also a more complex temporal aspect that needs to be modeled by sequences of such announcements. As for the second reading, Gerbrandy considers it to be genuinely paradoxical, similar to more standard self-referential statements, such as the Liar Paradox.

3.8 Common knowledge

One of the important concepts involved in the study of social knowledge is that of *common knowledge.* The idea is that common knowledge of a fact by a group of people is more than just the individual knowledge of the group members. This would be called *mutual knowledge.* Common knowledge is something more – what, exactly, is an issue, as is how to model it in the kinds of models we are dealing with.

Countries differ as to which side of the road one drives a car; the matter is one of social and legal convention. As it happens, at the time of this writing all three co-authors are living in countries in which people drive on the left. Suppose that in one of those, the government decides to change the driving side. But suppose that the change is made in a quiet way, so that only one person in the country, say Silvanos, finds out about it. After this, what should Silvanos do? From the point of view of safety, it is clear that he should not obey the law: since others will be disobeying it, he puts his life at risk. Suppose further that the next day the government decides to make an announcement to the press that the law was changed. What should happen now? The streets are more dangerous and more unsure this day, because many people will still not know about the change. Even the ones that have heard about it will be hesitant to change, since they do not know whether the other drivers know or not. Eventually, after further announcements, we reach a state where:

$$\text{The law says } \textit{drive on the right} \text{ and everyone knows (12).} \tag{12}$$

Note that (12) is a circular statement. The key point is not that everyone know what the law says, but that they in addition know *this very fact*, the content of the sentence you are reading.

This is an intuitive conception of common knowledge. Obviously it builds on the notion of knowledge, and since there are differing accounts of knowledge there will be alternative presentations of common knowledge. When we turn to the logical formalism that we'll use in the rest of this chapter, the alternative presentations mostly collapse. The key here is to unwind a sentence like (12) into an infinite list:

α_0: The law says *drive on the right*.

α_1: α_0, and everyone knows α_0.

α_2: α_1, and everyone knows α_1.

...

Each of these sentences uses knowledge rather than common knowledge. Each also implies its predecessors. Taking the *infinite conjunction*

$$\alpha_0 \wedge \alpha_1 \wedge \alpha_2 \wedge \cdots \tag{13}$$

we arrive at a different proposal for common knowledge. As it happens, given the kind of modeling that we are doing in this chapter, the *fixed point account* in (12) and the *infinite iteration account* in (13) agree.

History An early paper on common knowledge is Friedell [33]. This paper contains many insights, both mathematical and social. In view of this, it is even more noteworthy that the paper is not common knowledge for people in the field. Probably the first commonly-read source in the area is David Lewis' 'Convention' [59]. Heal's 1978 paper [44] came a decade later and is still a good source of examples. In the area of game theory, Aumann's [6] gives one of the first formalizations of common knowledge. McCarthy formalizes common knowledge in a rather off-hand way when solving a well-known epistemic riddle, the Sum and Product-riddle [65] (although at the time it was unknown to him that this riddle originated with the Dutch topologist Freudenthal [32]) as an abstract means towards solving the Sum and Product-riddle. McCarthy's work dates from the seventies but was only published later in a collection of his work that appeared in 1990.

Concerning the formalizations, here is how matters stand in two textbooks on epistemic logic: Fagin et al. [30] defines common knowledge by transitive closure, whereas Meyer and van der Hoek [66] define it by reflexive transitive closure. There is a resurgence of interest in variants of the notion, e.g., Artemov's evidence-based common knowledge, also known as justified common knowledge [3]. Another interesting variant is *relativized (or conditional) common knowledge*, which came to play an important role in some recent developments of dynamic epistemic logic [102, 49].

4 Epistemic Logic: What and Why?

We have introduced logical systems along with our representations in Section 2. We have presented a set of logical languages and some semantics for them. One of the first things one wants to do with a language and a semantics is to propose one or another notion of *valid* sentences; informally, these are the sentences true in all intended models. In all the settings of this paper, this information on validity includes the even more useful information of which sentences semantically imply which others. Then one wants to present a *proof system* for the valid sentences. There are several reasons why one would want an axiomatic system in the first place. One might wish to compare alternative presentations of the same system. One might also want to "boil a system down" to its essentials, and this, too, is accomplished by the study of axiomatic systems. Finally, one might study a system in order to see whether it would be feasible (or even possible) for a computer to use the system. We are not going to pursue this last point in our chapter, but we instead emphasize the "boiling down" aspect of *logical completeness results.*

4.1 Logic for ignorance of two alternatives

We return to our earliest scenario of Section 2, one person in a room with a concealed coin. We have a language including a knowledge operator K and a semantics for it using just one model. We write $\models \varphi$ to say that φ is true in that model. The object of our logical system is to give an alternative characterization of the true sentences.

Figure 1 contains our logical system truth. This paper is not the place to learn about logical systems in a detailed and deep way, but in the interests of keeping the interest of philosophers who may not know or remember the basics of logic, we do hope to provide a refresher course.

We say that φ is provable in our system if there is a sequence of sentences each of which is either an axiom or follows from previous sentences in the sequence by using one of the two rules of inference, and which ends in φ. In this case, one would often write $\vdash \varphi$, but to keep things simple in this chapter we are not going to use this notation.

all sentential validities	
$\mathsf{H} \leftrightarrow \neg\mathsf{T}$	exclusivity
$\neg K\mathsf{H}, \neg K\mathsf{T}$	basic ignorance axioms
$K\neg\varphi \to \neg K\varphi$	consistency of knowledge
$K(\varphi \to \psi) \to (K\varphi \to K\psi)$	distribution
$K\varphi \to \varphi$	veracity
$K\varphi \to KK\varphi$	positive introspection
$\neg K\varphi \to K\neg K\varphi$	negative introspection
From φ and $\varphi \to \psi$, infer ψ	modus ponens
From φ, infer $K\varphi$	necessitation

Figure 1: A logical system for valid sentences concerning two exclusive alternatives and a very simple semantics of K. The axioms are on top, the rules of inference below.

Here is a simple deduction in the logic, showing that $\vdash K\neg K\mathsf{T}$:

1. $\mathsf{H} \leftrightarrow \neg\mathsf{T}$
2. $(\mathsf{H} \leftrightarrow \neg\mathsf{T}) \to (\neg\mathsf{T} \to \mathsf{H})$
3. $\neg\mathsf{T} \to \mathsf{H}$
4. $K(\neg\mathsf{T} \to \mathsf{H})$
5. $K(\neg\mathsf{T} \to \mathsf{H}) \to (K\neg\mathsf{T} \to K\mathsf{H})$
6. $K\neg\mathsf{T} \to K\mathsf{H}$
7. $\neg K\mathsf{H}$
8. $\neg K\mathsf{H} \to ((K\neg\mathsf{T} \to K\mathsf{H}) \to \neg K\neg\mathsf{T})$
9. $(K\neg\mathsf{T} \to K\mathsf{H}) \to \neg K\neg\mathsf{T}$
10. $\neg K\neg\mathsf{T}$
11. $\neg K\neg\mathsf{T} \to K\neg K\neg\mathsf{T}$
12. $K\neg K\neg\mathsf{T}$

Line 1 is our exclusivity axiom and line 7 the basic ignorance axiom. A distribution axiom is found in line 5, and negative introspection in 11. This deduction uses propositional tautologies in lines 2 and 8, modus ponens in 3, 6, 9, 10, and 12, and necessitation in 4.

We mentioned before that there are several different reasons why one would construct a logical system to go along with a particular semantics. The first, perhaps, is that by *formulating sound principles, one uncovers (or highlights) hidden assumptions*. In this case, we can see exactly what the assumptions are in this example: they are the principles in the figure. The interesting point is that these assumptions are *all there is*: if one reasons with the system as above, then they will obtain *all* the truths.

Proposition 1 The logical system above is *sound and complete*: $\vdash \varphi$ iff φ is true in the model.

The point of this completeness theorem is that we have isolated all the assumptions in the scenario.

One remark which is of only minor significance in this discussion is that the veracity axioms $K\varphi \to \varphi$ may be dropped from the system. That is, all instances of them are provable anyway from the other axioms. The reason for including them is that they will be needed in all of the future systems.

Recall that the system here is based on our discussion at the beginning of Section 2. We then went on in Section 2.1 to the situation after Amina looks. It is not hard to re-work the logical system from Figure 1 to handle this second situation. We need only discard the basic ignorance axioms $\neg K\mathsf{H}$ and $\neg K\mathsf{T}$, and instead take $K\mathsf{H}$ so that we also get H. In particular, all of the sound principles that we noted for the earlier situation continue to be sound in the new one.

$\mathit{flip}\,\mathsf{H} \leftrightarrow \mathsf{T}$	flipping
$\varphi \leftrightarrow \mathit{flip}\,\mathit{flip}\,\varphi$	involution
$\mathit{flip}\,\neg\varphi \leftrightarrow \neg\mathit{flip}\,\varphi$	determinacy
$\mathit{flip}\,(\varphi \rightarrow \psi) \rightarrow (\mathit{flip}\,\varphi \rightarrow \mathit{flip}\,\psi)$	normality
$K\varphi \leftrightarrow \mathit{flip}\,K\varphi$	invariance
From φ, infer $\mathit{flip}\,\varphi$	necessitation

Figure 2: The logical system for knowledge and flipping. We also need everything from Figure 1, except the basic ignorance axioms (these are derivable).

4.2 Logic can change the world

There is another intuition about knowledge pertinent to the simple scenario that we have been dealing with. It is that *what Amina knows about a coin in a closed box is the same as what she would know if the box were flipped over.* In this section, we show what this means.

We consider the same language as before, except we add an operator *flip* to indicate the flipping the box over. For the semantics, let us begin with two models on the same state set.

1. M, a model with two states s and t, with the information that H is true at s and false at t, and T is true at t and false at s.

2. N, a model with two states s and t, with the information that H is true at t and false at s, and T is true at s and false at t.

Then we define $M, u \models \varphi$ and $N, u \models \varphi$ in tandem, the main points being that

$$\begin{array}{lcl} M, u \models \mathit{flip}\,\varphi & \text{iff} & N, u \models \varphi \\ N, u \models \mathit{flip}\,\varphi & \text{iff} & M, u \models \varphi \end{array}$$

Finally, we say that φ is *valid* if it holds at both states in both models. (This turns out to be the same as φ holding in any one state in either model.)

We present a logical system for validity in Figure 2. Perhaps the first exercise on it would be to prove the other flipping property: $\mathit{flip}\,\mathsf{T} \leftrightarrow \mathsf{H}$. We did not include in the figure the general principles of knowledge that we have already seen, but we intend them as part of the system. (These are the consistency, distribution, veracity, and introspection axioms; and the necessitation rule.) Note that the invariance axiom $K\varphi \leftrightarrow [\mathit{flip}]K\varphi$ is exactly the opening of our discussion. We refrain from presenting the completeness proof for this system, but it does hold. One thing to note is that the invariance axiom of this system makes the earlier ignorance axioms $\neg K\mathsf{H}$ and $\neg K\mathsf{T}$ unnecessary: they are derivable in this system.

4.3 Modal logics of single-agent knowledge or belief

At this point, we review the general topic of logics of knowledge. The basic language begins with a set P of atomic propositions. From these sets, a language $\mathcal{L}$ is built from the atomic propositions using the connectives of classical logic and also the knowledge operator K. We get a language which is basically the same as what we saw in Section 2, except that our set of atomic propositions is taken to be arbitrary, not just $\{\mathsf{H}, \mathsf{T}\}$.

Semantics We interpret this language on *relational models*. These are tuples $M = \langle S, R, V\rangle$ consisting of a *domain* S of *states* (or 'worlds'), an *accessibility relation* $R \subseteq S \times S$, and a *valuation* (function) $V : P \to \mathcal{P}(S)$. We usually write V_p instead of $V(p)$. We also write $s \to t$ instead of $R(s,t)$. We call these semantic objects *relational models*, but they are more often called *Kripke models*. We call a tuple of the form (M, s), where M is a model and s is a state in it, is an *epistemic state*.

We then define the interpretation of each sentence φ on an epistemic state (suppressing the name of the underlying model):

$$\begin{array}{lcl} [\![p]\!] & = & V_p \\ [\![\neg\varphi]\!] & = & \overline{[\![\varphi]\!]} \\ [\![\varphi \wedge \psi]\!] & = & [\![\varphi]\!] \cap [\![\psi]\!] \\ [\![K\varphi]\!] & = & \{s : \text{whenever } s \to t,\ t \in [\![\varphi]\!]\} \end{array}$$

It is more common to find this kind of definition "re-packaged" to give the interpretation of a sentence *at a given point*. This rephrasing would be presented as follows:

$$\begin{array}{lll} s \models p & \text{iff} & s \in V_p \\ s \models \neg\varphi & \text{iff} & s \not\models \varphi \\ s \models \varphi \wedge \psi & \text{iff} & s \models \varphi \text{ and } s \models \psi \\ s \models K\varphi & \text{iff} & \text{for all } \ t \in S : s \to t \text{ implies } t \models \varphi \end{array}$$

The two formulations are completely equivalent. At the same time, using one notation over another might lead different people to different insights or problems.

Further semantic definitions A sentence φ is *valid* on a model M, notation $M \models \varphi$, if and only if for all states s in the domain of M: $s \models \varphi$. A formula φ is *valid*, notation $\models \varphi$, if and only if for all models M (of the class of models for the given parameters of A and P): $M \models \varphi$. That is, φ holds on all epistemic states.

Logic and variations on it The logical system for validity is a sub-system of one which we already have seen. Look back at Figure 1, and take only the propositional tautologies, modus ponens, the distribution axiom, and the necessitation rule. This logical system is called K. Every book on modal logic will prove its completeness: a sentence is valid in the semantics just in case it can be proved from the axioms using the rules.

What's the point? For the modeling of knowledge, all we have so far is a spare definition and logic: An agent lives in a world and can see others. What it knows in a given world is just what is true in the worlds it sees. This seems a far cry from a full-bodied analysis of knowledge. The logical completeness result underscores the point. Any agent who "knew" things in the manner of this semantics would exemplify properties indicated by the logic. In particular, it would act as if the distribution axiom and necessitation rule held. The former, turned around a bit, says that the agent would be a *perfect reasoner*: if it knows φ and also knows $\varphi \to \psi$, then it automatically and effortlessly knows ψ. Necessitation says that it also knows all the general features of this logic. Thus, the agent is *logically omniscient*. And one would be hard-pressed to maintain that such an agent "knew" in the first place. For it is even possible that in some situations (models) the agent would "know" things which are false:

Ax	formal statement	property of R	interpretation
K	$K(\varphi \to \psi) \to (K\varphi \to K\psi)$	(none)	closed under modus ponens
T	$K\varphi \to \varphi$	reflexive	veracity
D	$K\varphi \to \neg K\neg\varphi$	serial	consistency
4	$K\varphi \to KK\varphi$	transitive	positive introspection
5	$\neg K\varphi \to K\neg K\varphi$	Euclidean	negative introspection

Figure 3: Axiom schemes of modal logic with their relational correspondents and epistemic interpretations.

this might well happen the agent lived in a world which was not among those it considered possible.

This leads to our next point. If one wants to model agents with certain epistemically-desirable properties (see below for these), one can impose mathematical conditions on the accessibility relation of models under consideration. Then one changes the definition of *valid* from *true in all epistemic states* to *true in all epistemic states meeting such-and-such a condition.*

To see how this works, we need a definition. A *frame* is the same kind of structure as what we are calling a *model*, but it lacks the valuation of atomic sentences. So it is just a pair $F = \langle S, R\rangle$, with R a relation on S. (In other terms, a frame is a *graph.*) Given a sentence φ in our logic, we say that $F \models \varphi$ if for all valuations $V : P \to \mathcal{P}(S)$, every $s \in S$ satisfies φ in the model $\langle S, R, V\rangle$.

Figure 3 presents well-known *correspondences* between conditions on a frame and properties of the logic. One example: a frame F satisfies each instance of D (say) iff F meets the condition listed that every point in it has a successor. is related by R to *some* point or other. Reflexivity means that every point is related to itself. Transitivity means that if x is related to y, and y to z, then x is related to z. The Euclidean condition mentioned in connection with the 5 axioms is

$$(\forall x)(\forall y)(\forall z)((xRy \land xRz) \to yRz).$$

There is a further aspect of the correspondence. If one wants to study the sentences which are valid on, say, transitive models, then one need only add the corresponding axiom (in this case $K\varphi \to KK\varphi$) to the basic logic that we mentioned above. On a conceptual level, we prefer to turn things around. If one wants to model agents which are positively introspective in the sense that if the know something, then they know that they know it, then one way is to assume, or argue, that they work with transitive models. The same goes for the other properties, and for combinations of them.

There are many modal systems indeed, but we wish to mention only a few here. We already have mentioned K. If we add the axioms called T and 4, we get a system called S4. It is complete for models which are reflexive and transitive, and intuitively it models agents who only know true things and which are positively introspective. If one adds the negative introspection axioms 5, one gets a system called S5. The easiest way to guarantee the S5 properties is to work with relations which are reflexive, symmetric, and transitive (*equivalence relations*), for these are also Euclidean.

Turning from knowledge to belief, the T axiom is highly undesirable. So logics appropriate for belief will not have T. But they often have D, the seriality axioms. These may be interpreted as saying that if an agent believes φ, then it does not at the same time believe

$\neg\varphi$. Probably the most common logic of belief is KD45, obtained by adding to K the other axioms in its names. KD45 is complete with respect to models which have the properties listed above. When studying belief, one usually changes the name of the modality from K to B, for obvious reasons.

We wish to emphasize that all of the intuitive properties of knowledge and belief discussed in this section, and indeed in this chapter as a whole, are highly contestable. Using logic does not commit one to any of those properties. But logic can help to clarify the commitments in a given form of modeling. For example, any modeling of knowledge using relational semantics will always suffer from problems having to do with logical omniscience, as we have seen.

4.4 Multi-agent epistemic logic

The formal move from the basic modal logic of the last section its multi-agent generalization is very easy. One begins with a set A of *agents* in addition to the set of atomic propositions. Then the syntax adds operators K_a, and we read $K_a\varphi$ as "a knows φ." The semantics then moves from the models of the last section to what we shall call *epistemic models*. These are tuples $\langle S, R, V\rangle$ as before, except now R is an *accessibility* (function) $R\ : A \to \mathcal{P}(S \times S)$. That is, it is a family of accessibility relations, one for each agent. We usually write R_a instead of $R(a)$. Further, we write $s \xrightarrow{a} t$ instead of $(s,t) \in R_a$.

Example 2 We are going to present an example which hints at the applicability of our subject to the modeling of games.

Consider three players Amina, Bao, and Chandra (a, b, c). They sit in front of a deck consisting of exactly three cards, called clubs, hearts, and spades. Each is dealt a card, they look, and nobody sees anyone else's card. We want to reason about knowledge in this situation. It makes sense to take as atoms the nine elements below

$$\{\mathit{Clubs}_a, \mathit{Clubs}_b, \ldots, \mathit{Spades}_b, \mathit{Spades}_c\}.$$

The natural model for our situation has six states. It is shown in a slightly re-packaged form in Figure **??**. We call the model *Hexa*. The states are named according to who has what. For example, $\clubsuit\heartsuit\spadesuit$ is the state where Amina has clubs, Bao hearts, and Chandra spades. The lines between the states have labels, and these indicate the accessibility relations. So Bao, for example, cannot tell the difference between $\heartsuit\clubsuit\spadesuit$ and$\spadesuit\clubsuit\heartsuit$: if the real deal were one of those, he would think that it could be that same deal, or the other one (but no others).

We mentioned that the picture of *Hexa* differs slightly in its form from what the official definition calls for. That version is mathematically more elegant but does not immediately lend itself to a picture. It would have, for example,

$$\begin{array}{lcl}
V(\mathit{Clubs}_a) & = & \{\clubsuit\heartsuit\spadesuit, \clubsuit\spadesuit\heartsuit\} \\
 & \cdots & \\
V(\mathit{Spades}_c) & = & \{\clubsuit\heartsuit\spadesuit, \heartsuit\clubsuit\spadesuit\} \\
R_a & = & \{(\clubsuit\heartsuit\spadesuit, \clubsuit\spadesuit\heartsuit), (\heartsuit\clubsuit\spadesuit, \heartsuit\spadesuit\clubsuit), (\spadesuit\clubsuit\heartsuit, \spadesuit\heartsuit\clubsuit)\} \\
 & \cdots &
\end{array}$$

We can then evaluate sentences in English by translating into the formal language and using the semantics. Here are some examples.

Amina knows she has the heart card. We translate to $K_a\mathit{Hearts}_a$. The semantics in *Hexa* is $[\![K_a\mathit{Hearts}_a]\!] = \{\heartsuit\clubsuit\spadesuit, \heartsuit\spadesuit\clubsuit\}$. (In more detail: in each of the two worlds $\heartsuit\clubsuit\spadesuit$ and $\heartsuit\spadesuit\clubsuit$, every world that Amina thinks is possible belongs to $V(\mathit{Hearts}_a)$. And if s is one of the four other worlds, there is some world accessible from s for Amina which does not belong to $V(\mathit{Hearts}_a)$. For example, in $s = \clubsuit\heartsuit\spadesuit$, the world s itself is accessible for Amina, and she does not have hearts there.) That is, our sentence is true exactly at $\clubsuit\heartsuit\spadesuit$ and $\spadesuit\heartsuit\clubsuit$. Note that this is precisely the set of worlds where Amina indeed has hearts. So the sentence *Amina has hearts if and only if she knows she has hearts* comes out true at all states.

If Bao has spades, Chandra has clubs. This is $\mathit{Spades}_b \to \mathit{Clubs}_c$. The semantics is

$$\{\clubsuit\heartsuit\spadesuit, \heartsuit\clubsuit\spadesuit, \heartsuit\spadesuit\clubsuit, \spadesuit\clubsuit\heartsuit, \spadesuit\heartsuit\clubsuit\}.$$

If Amina has hearts, then she knows that if Bao has spades, Chandra has clubs. The translation is

$$\mathit{Hearts}_a \to K_a(\mathit{Spades}_b \to \mathit{Clubs}_c).$$

This true at all states.

Bao considers it possible that Amina has spades but actually Amina has clubs. We translate "consider it possible that φ" by "does not know that φ is false." So our sentence here is $\neg K_b \neg \mathit{Spades}_a \wedge \mathit{Clubs}_a$. Usually one prefers to avoid negation by introducing an abbreviation. So if we say that $\hat{K}_b\varphi$ abbreviates $\neg K_b \neg\varphi$, then we may read this as *Bao considers φ possible* and translate our sentence as above. Its semantics is $\{\clubsuit\heartsuit\spadesuit\}$.

The last sentence shows the *duality* between "consider possible" and "know". This phenomenon of dual definitions is something we shall see later as well.

Logic We define *validity* exactly as in the last section, but using the generalized language and semantics. Given a language and a semantics, one task for logic is to determine the set of sentences which are valid. In particular, we might ask for a nice logical system for the valid sentences, since this may well give insight into the underlying assumptions in the model. Now in what we have been discussing in this section, we might seek at least three logical systems:

1. A system for validity on the model *Hexa*.

2. A system for validity on all models whatsoever.

3. A system for validity on all models that are "basically similar" to *Hexa*.

For the first question, we modify the system of Figure 1 for ignorance of two alternatives. The axioms that actually pertain to H and T must be replaced, of course. Instead of exclusivity, we take an axiom that says, informally, for exactly one of the six states s of *Hexa*, all atoms true in s hold, and all not true in s do not hold. We also add an axiom saying that *If Amina has clubs, then she knows it*, and similarly for the other players and cards, and also that *Amina does not know which card any other player holds.* All of the rest of the axioms are valid in this model, as are the rules. Each statement of the old system would be replaced by three, one for each player. So one of the introspection axioms would be $K_b\varphi \to K_bK_b\varphi$. In particular, the introspectivity and necessitation principles are valid.

In passing, we note that this logical system has no *interaction properties* between the knowledge of different agents. That is, none of the axioms mix K_a and K_b for different a and b. Mathematically, this means that the generalization of single-agent knowledge to the multi-agent case will be almost trivial. But there are two caveats: first, the phenomenon of *common knowledge* does involve discussions of different agents' knowledge, and so it turns out to be harder to study. And second, it really is possible to have situations where interaction properties make sense. For example, suppose one wants to model situations where *everything Bao knows Chandra also knows.* In the semantics, one would want $R_c \subseteq R_b$. And then in the logic one could add $K_b\varphi \to K_c\varphi$.

For the second question above, the fact that we are dealing with a larger class of models means that fewer logical principles are sound. The only sound principles would be the propositional tautologies and the distribution axioms, and the rules of modus ponens and necessitation.

The last question is obviously not precise. The point of raising it is that one can make precise the sense in which games are, or are not, similar by using the logical principles that hold. For example, one often simplifies matters by assuming that adversaries are perfect reasoners, and in this setting it is natural to *assume* the introspectivity principles in the modeling. That is, one works only with models where those principles turn out to be sound. The easiest way to arrange this is to look at the chart in Figure 3. If each accessibility relation $\xrightarrow{a}$ is reflexive, transitive, and euclidean, then the model will satisfy both introspectivity assertions. (This holds no matter what the valuation V happens to do.) It turns out that a relation with these properties is automatically *symmetric* and hence an equivalence relation. Moreover, an equivalence relation on a set is equivalently formulated as a *partition* of the set. So one often finds the tacit assumption in much of the game theory/economics literature that the models used are *partition models* consisting of a set S and a *partition* of S for each player.

Identity conditions on models We first looked at announcements in Section 2.2. In discussing (4) and (5), we noted the need for principled identity conditions on relational models. Fortunately, the general study of relational models and their logical systems gives us a usable condition, called *bisimulation.* This condition is coarser than *isomorphism*, does what we want (in particular, it tells us that (4) and (5) ought to be identified), and has an intuitive resonance as well. For the formal definition and much more, see any text on modal logic, for example Blackburn et al. [20].

4.5 Common knowledge

We have discussed the idea of common knowledge in Section 3.8. We turn now to the formalization on top of what we saw in Section 4.4 above. We present a generalization of our previous concept, however. For each group $B \subseteq A$, we want notions of *group knowledge for the set B* and *common knowledge for the set B*. This last notion has the same intuitive basis as common knowledge itself. For example, it is common knowledge among Englishmen that one drives on the left, but this common knowledge does not hold for the entire world.

For the syntax of group knowledge, we add operators E_B to the language of multi-agent epistemic logic. The semantics is given by

$$[\![E_B\varphi]\!] = \bigcap_{a\in A} [\![K_a\varphi]\!].$$

This means that we (straightforwardly) translate $E_B\varphi$ as *Everyone knows φ*. In the case of finitely many agents (the assumption in practically all papers on the topic), $E_B\varphi$ may be regarded as an abbreviation.

Example 3 We return to the model *Hexa* from Section 4.4 (see Figure **??**). We have

$$♣♡♠ \models E_{\{a,b\}}\neg(\mathit{Spades}_a \wedge \mathit{Clubs}_b \wedge \mathit{Hearts}_c).$$

That is, both Amina and Bao know that the deal of cards is *not* ♠♣♡. However, despite this, each does not know that the other knows this fact.

We next turn to common knowledge. The syntax adds operators $C_B\varphi$ to the language, exactly as with group knowledge. The semantics is more problematic, and indeed there are differing proposals in the literature. We follow the most common treatment. One essentially takes the unwinding of the fixed point that we saw in (13) in Section 3.8 as the definition, and then the fixed point property becomes a semantic consequence later on.

For the semantics, for each group B we pool all the accessibility relations for the members of B together, and then take the *reflexive-transitive closure*:

$$R^*_B \equiv (\bigcup_{a\in B} R_a)^*.$$

(see below for an explanation). Then we interpret via

$$s \models C_B\varphi \quad \text{iff} \quad \text{for all } \ t \in S : R^*_B(s,t) \text{ implies } t \models \varphi$$

Alternatively said, $C_B\varphi$ is true in s if φ is true in any state s_m that can be reached by a (finite) path of zero or more states $s_1, \ldots, s_m$ such that, for not necessarily different agents $a, b, c \in B$: $R_a(s_1, s_2)$, $R_b(s_2, s_3)$, and …, and $R_c(s_{m-1}, s_m)$. A path of zero states is just a single state alone. Hence if $C_B\varphi$ is true at s, then automatically φ is true at s as well.

As an example of both the formal and informal concepts, we consider an n-person muddy children scenario (see Section 3.1), before any announcement that at least one agent is muddy. It is easy to describe the model: it has 2^n states with the rest of the structure determined in the obvious way. Then it is common knowledge at all states that no agents know their own state. More interesting is the comment that in this model, if s is a state and every agent knows φ at s, then φ is already common knowledge at all states.

The logic of common knowledge adds two principles to the basic multi-agent epistemic logic. Those are the *Mix Axiom*:

$$C_B\varphi \to \varphi \wedge E_B C_B\varphi$$

(so-called because it deals with the interactions of the the two operators of this section) and the *induction rule*:

$$\text{from } \chi \to \psi \text{ and } \chi \to K_a\chi \text{ for all } a \in B, \text{ infer } \chi \to C_B\psi.$$

Using this logic, one can prove the important properties of common knowledge. For example, it is *idempotent*:

$$C_B\varphi \leftrightarrow C_B C_B\varphi.$$

The interesting direction here says that if φ is common knowledge in a group, then the fact of its being common knowledge is itself common knowledge in the group.

4.6 Belief-knowledge logic

Intuitively, *belief* and *knowledge* are related but different. However, we have heretofore conflated the two notions. We present here the simplest possible system which can sensibly separate the two by incorporating both at the same time with different semantics. It and the logical axioms are taken from Meyer and van der Hoek [66].

We fix sets A of agents and P of atoms. To separate the two notions, we need a language with different operators K and B.

A *knowledge-belief model* (KB-model) is a Kripke model of the form $\langle S, R_a^K, R_a^B, V\rangle_{a\in A}$, where S is set of states, R_a^K and R_a^B are binary accessibility relations in $\mathcal{P}(S \times S)$, and V is a function from P to $\mathcal{P}(S)$. We write $s \overset{a}{\sim} t$ instead of $(s,t) \in R_a^K$, and $s \xrightarrow{a} t$ instead of $(s,t) \in R_a^B$. As the letters K and B indicate, the first relation $\overset{a}{\sim}$ is meant to capture the *knowledge* of agent a, while the second $\xrightarrow{a}$ captures the agent's *beliefs*.

A KB model is required to satisfy the following conditions: $\overset{a}{\sim}$ is an equivalence relation; $\xrightarrow{a}$ is serial; if $s \overset{a}{\sim} t$ and $s \xrightarrow{a} w$, then $t \xrightarrow{a} w$; and finally, $\xrightarrow{a}$ is included in $\overset{a}{\sim}$. So the modeling reflects the following intuitions: the *truthfulness and introspection of of knowledge*, full belief introspection (agents *know their own beliefs*), *beliefs are consistent*, and *knowledge implies belief.* It is not necessary to assume that $\xrightarrow{a}$ is transitive and Euclidean, since these properties immediately follow from the above conditions. So we also have for free that belief is introspective, in the usual sense.

Notice also that, as observed on the example in Section 2.3, the knowledge relation $\overset{a}{\sim}$ is recoverable from the belief relation $\xrightarrow{a}$, via the following rule:

$$s \overset{a}{\sim} t \quad \text{iff} \quad (\forall w)(s \xrightarrow{a} w \text{ iff } t \xrightarrow{a} w). \tag{14}$$

To see this, first assume that $s \overset{a}{\sim} t$. Then also $t \overset{a}{\sim} s$. From this and one of the conditions in a KB model, we get the right-hand side of (14). And if the right-hand side holds, we show that $s \overset{a}{\sim} t$. First, there must be some w so that $s \xrightarrow{a} w$. For this w we also have $t \xrightarrow{a} w$. And then using the fact that $\overset{a}{\sim}$ is an equivalence relation included in $\xrightarrow{a}$, we see that $s \overset{a}{\sim} t$.

So, in fact, one could present KB-models simply as *belief models* $\langle S, \xrightarrow{a}, V\rangle$, where $\xrightarrow{a}$ is transitive, serial and Euclidean, and one can take the knowledge relation as a defined notion, given by the rule (14) above. We interpret the logical system in KB-models via

$$\begin{array}{lcl} \llbracket K_a\varphi\rrbracket & = & \{s \in S : t \in \llbracket\varphi\rrbracket, \text{ for all } t \text{ such that } s \overset{a}{\sim} t\} \\ \llbracket B_a\varphi\rrbracket & = & \{s \in S : t \in \llbracket\varphi\rrbracket, \text{ for all } t \text{ such that } s \xrightarrow{a} t\} \end{array} \tag{15}$$

Example 4 The easiest example is a two-state model for ignorance of two alternatives, say heads and tails, together with a belief in one of them, say heads. Formally, we have one agent, so we drop her from the notation. There are two states s and t, for H and T. The relation $\to$ is $\mathsf{H} \to \mathsf{H}$ and $\mathsf{T} \to \mathsf{H}$. The relation $\sim$ therefore relates all four pairs of states. Then at both states $B\mathsf{H} \wedge \neg K\mathsf{H}$. In particular, the agent believes heads at the tails state t. Hence we have our first example of a *false belief.* But agents in KB-models are not so gullible that they believe absolutely *anything*: $\neg B(\mathsf{H} \wedge \mathsf{T})$, for example. And indeed, the seriality requirement prohibits an agent from believing a logical inconsistency.

The logic is then axiomatized by the S5 system for knowledge, the KD45 system for belief, and two connection properties: First, $B_a\varphi \to K_aB_a\varphi$. So an agent may introspect on her own beliefs. (It also follows in this logic that $\neg B_a\varphi \to K_a\neg B_a\varphi$.) We should mention that *introspection about beliefs* is less controversial than *introspection about knowledge.* If we take knowledge to be a relation between an agent and an external reality, then it is as problematic to account for an agent's knowledge of their own knowledge as it is to account for any other type of knowledge. But to the extent that belief is an "internal" relation, it seems easier to say that fully-aware agents should have access to their own beliefs.

The second logical axiom connecting knowledge and belief is $K_a\varphi \to B_a\varphi$. This reiterates an early point: we are thinking of knowledge as a strengthening of belief. It is sound due to the requirement that $\xrightarrow{a}$ be included in $\overset{a}{\sim}$.

Variations There are some variations on the soundness and completeness result which we have just seen. Suppose one takes an arbitrary relation $\xrightarrow{a}$, then defines $\overset{a}{\sim}$ from it using (14), and then interprets our language on the resulting structures by (15). Then $\overset{a}{\sim}$ is automatically an equivalence relation, and so the S5 axioms for knowledge will be sound. Further, the two connection axioms automatically hold, as does negative introspection. Continuing, we can add impose conditions on $\xrightarrow{a}$ (such as seriality, transitivity, the Euclidean property, or subsets of these), and then study validity on that class. In the logic, one would take the corresponding axioms, as we have listed them in Figure 3. In all of the cases, we have completeness results.

4.7 Conditional doxastic logic

We now re-work the logical system of Section 4.6, so that it can handle conditionals in the way that we did in Section 2.4. The logical system is based on Board [21], and Baltag and Smets [15, 16, 17]. Following the latter, we'll call it *conditional doxastic logic* (**CDL**).

Its syntax adds to propositional logic statements of the form $B_a^\alpha\varphi$, where a is an agent and α and φ are again sentences in the logical system. This should be read as "If a were presented with evidence of the assumption α in some world, then she should believe φ describes the world (as it was before the presentation)."

The simplest semantics for this logic uses *plausibility models.* These are also special cases of the *belief revision structures* used in Board [21]. (We shall see the general notion at the end of this section.) Plausibility frames are Kripke structures of the form $(S, \leq_a)_{a\in A}$, consisting of a set S endowed with a family of *locally pre-wellordered* relations $\leq_a$, one for each agent a. When S is *finite*, a locally pre-wellordered relation on S is just one that is *reflexive, transitive and weakly connected both forwards and backwards*[3], i.e. $s \leq_a t$ and $s \leq_a w$ implies that either $t \leq_a w$ or $w \leq_a t$, and also $t \leq_a s$ and $w \leq_a t$ implies that either $t \leq_a w$ or $w \leq_a t$. Equivalently, we have a *Grove system of spheres*, just as in Section 2.3, consisting of a number of (disjoint) "smallest spheres" (listing the worlds "in the center"), then surrounding them the next smallest spheres (containing worlds a little less plausible than these central ones), then the next ones, having worlds a little less plausible than these, etc. To match the notion of local pre-wellorder above (again, in the finite case), we need to assume that every world belongs to some sphere, and that if two spheres intersect or are both included in a larger sphere, then one is included in the other.

As for Kripke models in general, a *plausibility model* is a tuple $M = (S, \leq, V)$, where $(S, \leq)$ is a plausibility frame and V is a valuation on it. A *doxastic state* is a tuple of the form (M, s), where M is a plausibility model and $s \in M$.

[3] This last property is also known as *no branching to the left or to the right.*

Remark For readers of van Benthem and Martinez' Chapter 3b, we mention that our orderings go the other way from theirs: for them, "more plausible" is "upward" in the ordering, and for us it is "in the center" or "lower down".

As in the example in Section 2.3, we define a *knowledge (indifference) relation* by putting

$$s \overset{a}{\sim} t \text{ iff either } s \leq_a t \text{ or } t \leq_a s.$$

Plausibility models for only one agent have been used as models for *AGM* belief revision in [35, 81, 85]. The additional indifference relations turn out to be useful in modeling, as we indicate shortly.

Similarly, we define a *belief relation* by putting:

$$s \xrightarrow{a} t \quad \text{iff} \quad s \overset{a}{\sim} t \text{ and } (\forall u)(s \overset{a}{\sim} u \to t \leq_a u) \ .$$

It is easy to see that $\overset{a}{\sim}$ and $\xrightarrow{a}$ satisfy all the postulates of a KB model. More generally, we obtain a *conditional belief* relation by putting, for any set $X \subseteq S$ of states:

$$s \xrightarrow{a,X} t \quad \text{iff} \quad s \overset{a}{\sim} t, t \in X, \text{ and } (\forall u)(s \overset{a}{\sim} u \ \& \ u \in X \to t \leq_a u) \ .$$

In other words, $s \xrightarrow{a,X} t$ if a considers t to be possible in s, if $t \in X$, and if t is among the most plausible worlds for a with these two conditions. So here we see the basic idea: reasoning about a's hypothetical beliefs assuming α involves looking at the relation $\xrightarrow{a,X}$ where $X = [\![\alpha]\!]$.

Observe that the belief relation $\xrightarrow{a}$ is the same as the conditional belief relation $\xrightarrow{a,S}$, where S is the set of all states. As a passing note, for each $X \subseteq S$ we can use the relations $\xrightarrow{a,X}$ to make a structure which is *almost* a KB-model in the sense of Section 4.6. Take S for the set of worlds, $\xrightarrow{a,X}$ for the belief relation for each agent a, take the knowledge relation $\overset{a}{\sim}$ for a as above, and use the same valuation in S^X as in S. The only property of KB-models lacking is that the belief relations might fail to be serial: a state s has no $\xrightarrow{a,X}$-successors if X is disjoint from the $\overset{a}{\sim}$-equivalence class of s. This makes sense: seriality corresponds to consistency of belief; but X being disjoint from the $\overset{a}{\sim}$-equivalence of s corresponds to conditionalizing on a condition X that is *known* (with absolute certainty) to be false: the resulting set of conditional beliefs should be inconsistent, since it contradicts (and at the same time it preserves) the agent's knowledge.

For the semantics, we say

$$[\![B_a^\alpha \varphi]\!] \quad = \quad \{s \in S : t \in [\![\varphi]\!], \text{ for all } t \text{ such that } s \xrightarrow{a,X} t, \text{ where } X = [\![\alpha]\!]\} \tag{16}$$

In other words, to evaluate a doxastic conditional, a looks at which of her possible worlds are the most plausible given the antecedent α, and then evaluates the conclusion on all of those worlds. If they all satisfy the conclusion φ of the conditional, then a believes φ conditional on α. It is important that the evaluation take place in the original model.

Example 5 The model S from Section 2.4 had four worlds u, v, w, x. Amina's plausibility relation $\leq_a$ is essentially the ordered partition: $\{v, w\} < \{u, x\}$. Bao's plausibility relation is the reflexive closure of the relation $\{(v, w), (w, v)\}$. We are interested in sentences $B_a^\alpha \varphi$, where $\alpha = K_b\mathsf{H} \vee K_b\neg\mathsf{H}$. Let $X = [\![\alpha]\!] = \{u, x\}$. Then $\xrightarrow{a,X}$ relates all four worlds to u and x. Our semantics in (16) is equivalent to what we used in (11). Note as well that the right-hand model in (8) shows $\xrightarrow{a,S}$ and $\xrightarrow{b,S}$.

Knowledge in plausibility models There are two equivalent ways to define knowledge in plausibility models. One can use the definition (15) applied directly to the $\overset{a}{\sim}$ relations introduced above, saying that

$$s \models K_a\varphi \text{ iff } s \overset{a}{\sim} t \text{ implies } t \models \varphi.$$

Alternatively, one can use the following observation to get an intuitively appealing reformulation.

Proposition 6 Let S be a plausibility model and $s \in S$. The following are equivalent:

1. $s \models K_a\varphi$.
2. $s \models B_a^{\neg\varphi}\bot$, where $\bot$ is a contradiction such as $p \wedge \neg p$.
3. $s \models B_a^{\neg\varphi}\varphi$.
4. $s \models B_a^{\alpha}\varphi$, for *every* sentence α.

Here is the reasoning: If (1) holds, then s has no $\xrightarrow{a,X}$ -successors at all, where $X = [\![\neg\varphi]\!]$. This means that the next two assertions hold vacuously. Also, notice that, for *every* set X, $s \xrightarrow{a,X} t$ implies $s \overset{a}{\sim} t$. Hence, if (1) holds then, for any sentence α, φ is true at all $\xrightarrow{a,X}$ -successors of s, where $X = [\![\alpha]\!]$. Therefore (4) holds as well. Conversely, (4) clearly implies (3); also, if either of (2) or (3) hold, then the semantics tells us that s has no $\xrightarrow{a,X}$ -successors, so every $t \overset{a}{\sim} s$ must satisfy φ; i.e., (1) must hold.

A proposal going back to Stalnaker [86] *defines* "necessity"[4] in terms of conditionals, via the clause (3) above. This contains an idea concerning our notion of "knowledge": what it means to know φ (in this strong sense) is that one would still believe φ even when hypothetically assuming $\neg\varphi$.

Finally, (4) can be related to the "defeasibility analysis" discussed in Section 3.4. Indeed, what is says is that our notion of knowledge satisfies a *stronger* version of this analysis than the original one: our knowledge is the same as "*absolutely unrevisable*" belief. One "knows" φ (in this absolute sense) if giving up one's belief in φ would *never* be justified, under *any* conditions (even when receiving false information).

The logic We list a complete axiomatization in Figure 4. Note that in the syntax, we take conditional belief as the only basic operator, and define knowledge via (3). Verifying the soundness of most of the axioms is easy, and we discuss only the principle of minimality of revision. Let $X = [\![\alpha]\!]$. Let $Y = [\![\alpha \wedge \varphi]\!]$, so $Y \subseteq X$. Suppose that $s \models \neg B_a^{\alpha}\neg\varphi$. So there is some t so that $s \xrightarrow{a,X} t$ and $t \models \varphi$. This means that, for all w, we have $s \xrightarrow{a,X} w$ iff $s \xrightarrow{a,Y} w$. We first show that if $s \models B_a^{\alpha}(\varphi \to \theta)$, then $s \models B_a^{\alpha\wedge\varphi}\theta$. To see this, let w be such that $s \xrightarrow{a,Y} w$. Then $s \xrightarrow{a,X} w$, and since $w \models \varphi$, we have $w \models \theta$. For the second half, one checks directly that $B_a^{\alpha\wedge\varphi}\theta$ implies $s \models B_a^{\alpha}(\varphi \to \theta)$.

Incidentally, the axioms of the logic have interpretations in terms of *belief revision*, as we shall see. In particular, the last axiom ("minimality of revision") corresponds to the conjunction of the **AGM** principles of Subexpansion and Superexpansion (principles (7) and (8) in Section 7).

[4] This is denoted in [86] by $\Box\varphi$. Note that it corresponds in our notation to $K\varphi$, and should not be confused with our notation for "safe belief" $\Box\varphi$ in the next subsection.

Syntax	$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid B_a^\alpha \varphi$	
Definition	$K_a\varphi := B_a^{\neg\varphi}\varphi$	(knowledge)
Main Axioms	$B_a^\alpha \alpha$	hypothetical acceptance
	$K_a\varphi \to \varphi$	veracity
	$K_a\varphi \to B_a^\alpha\varphi$	persistence of knowledge
	$B_a^\alpha\varphi \to K_a B_a^\alpha\varphi$	positive belief introspection
	$\neg B_a^\alpha\varphi \to K_a\neg B_a^\alpha\varphi$	negative belief introspection
	$\neg B_a^\alpha\neg\varphi \to (B_a^{\alpha\wedge\varphi}\theta \leftrightarrow B_a^\alpha(\varphi \to \theta))$	minimality of revision

Figure 4: Syntax and axioms for conditional doxastic logic. We also assume Modus Ponens, as well as Necessitation and the (K) axiom for B_a^α.

Adding common knowledge and belief It is also possible to expand the language with operators Cb_B^α and Cb_B^α, reflecting *conditional versions of common belief and knowledge* in a group B. For the proof systems and more on these systems, cf. Board [21], and Baltag and Smets [15, 16, 17]. Board's paper also contains interesting variations on the semantics, and additional axioms. Baltag and Smets offer a generalization of the notion of a plausibility model to that of a *conditional doxastic model*. Both authors considers applications to modeling in games.

Belief revision structures The plausibility models that we have been concerned with in this section may be generalized in a number of ways. First of all, the pre-wellorders for each agent might be *world-dependent*. This would be important to model agents with incorrect beliefs about their own beliefs, for example.

In this way, we arrive at what Board [21] calls *belief revision structures*. For more on this logic, including completeness results, see Board [21].

4.8 The logic of knowledge and safe belief

As we saw, Stalnaker's defeasibility analysis of knowledge asks a *weaker* requirement than the one satisfied by our notion of "absolutely unrevisable knowledge": namely, it states that φ is known (in the weak, defeasible sense) if there exists no *true* piece of information X such that the agent would no longer be justified to believe φ after learning X. Following Baltag and Smets [16, 17], we call *safe belief* this weak notion of defeasible "knowledge", and we use the notation $\Box_a\varphi$ to express the fact that a safely believes φ. We can immediately formalize this notion in terms of conditional belief arrows:

$$s \models \Box_a\varphi \text{ iff for all } t \in S, \text{ and all } s \in X \subseteq S : s \xrightarrow{a,X} t \text{ implies } t \models \varphi.$$

To our knowledge, the first formalization of safe belief (under the name of "knowledge") was due to Stalnaker [87], and used the above clause as a definition. Observe that it uses quantification over propositions (sets of states). It was only recently observed, in [16, 17, 88], that this second-order definition is equivalent to a simpler one, which takes safe belief as the

Kripke modality associated to the relation "at most as plausible as":

$$[\![\Box_a\varphi]\!] \quad = \quad \{s \in S : t \in [\![\varphi]\!], \text{ for all } t \leq_a s\} \tag{17}$$

This last condition was adopted by Baltag and Smets [19] as the definition of safe belief. The same notion was earlier defined, in this last form, by van Benthem and Liu [100], under the name of "preference modality".

Example 7 The situation described in Section 2.4 provides us with examples of safe and unsafe beliefs. In the model 10, Amina believes (though she doesn't know) that Bao doesn't know that the face of the coin is tails. If the real state is v, then this belief is true, and moreover it is *safe*: this is easy to see, since it is true at both v and w, which are the only two states that are at least as plausible for Amina as v. This gives us an example of a *safe belief which is not knowledge.* If instead the real state is u, then the above belief is still true (though is not known). But it is *not safe* anymore: formally, this is because at state x (which for Amina is at least as plausible as u), Bao does know the face is tails. To witness this unsafety, consider the sentence $\alpha := B_b\mathsf{H} \vee B_b\mathsf{T}$, saying that Bao knows the face of the coin. At the real world u, the sentence α is true; but, at the same world u, Amina does not believe that, if α were true then Bao wouldn't know that the face was tails: $u \models \alpha \wedge \neg B_a^\alpha \neg B_b\mathsf{T}$. This shows that Amina's belief, though true, can be defeated at state u by learning the true sentence α.

Stalnaker [88] observes that *belief can be defined in terms of safe belief*, via the logical equivalence: $B_a\varphi \leftrightarrow \neg\Box_a\neg\Box_a\varphi$, and that *the complete logic of the safe belief modality $\Box_a$ is the modal logic $S4.3$.*[5] Baltag and Smets [16, 17] observe that by combining safe belief $\Box_a\varphi$ with the "absolute" notion of knowledge $K_a\varphi$ introduced in the previous section, one can define conditional belief via the equivalence[6]:

$$B_a^\alpha\varphi \;\;\leftrightarrow\;\; (\neg K_a\neg\alpha \to \neg K_a\neg(\alpha \wedge \Box_a(\alpha \to \varphi))).$$

The logic of knowledge and safe belief is then axiomatized by the S5 system for knowledge, the S4 system for safe belief, and two connection properties: First, $K_a \to \Box_a\varphi$. This reiterates the earlier observation: knowledge (in our absolute sense) is a strengthening of safe belief. The second axiom says that the plausibility relation $\leq_a$ is connected within each $\stackrel{a}{\sim}$ -equivalence class:

$$K_a(\varphi \vee \Box_a\psi) \wedge K_a(\psi \vee \Box_a\varphi) \to K_a\varphi \vee K_a\psi.$$

Belief and conditional belief are derived notions in this logic, defined via the above logical equivalences.

4.9 Propositional Dynamic Logic

This section of our chapter mainly consists of brief presentations of logical systems which are intended to model notions of importance for epistemic or doxastic logic. The current subsection is an exception: *propositional dynamic logic* (**PDL**) is a system whose original motivations and main uses come from a different area, semantics studies programming languages. We are not concerned with this here, and we are presenting **PDL** in a minimum of

[5] $S4.3$ is the logic of reflexive transitive frames with no branching to the right.

[6] Van Benthem and Liu [100] use another logical equivalence to similarly define conditional belief.

Syntax	**Sentences** φ	p_i	$\neg\varphi$	$\varphi \wedge \psi$	$[\pi]\varphi$		
	Programs π	a	$?\varphi$	$\pi;\sigma$	$\pi \cup \sigma$	π^*	
Semantics	**Main Clauses**	$[\![[\pi]\varphi]\!] = \{s : \text{if } s[\![\pi]\!]t, \text{ then } t \in [\![\varphi]\!]\}$ $[\![?\varphi]\!] = \{(s,s) : s \in [\![\varphi]\!]\}$ $[\![\pi;\sigma]\!] = [\![\pi]\!] \ ; \ [\![\sigma]\!]$ $[\![\pi \cup \sigma]\!] = [\![\pi]\!] \cup [\![\sigma]\!]$ $[\![\pi^*]\!] = ([\![\pi]\!])^*$					

Figure 5: The language of Propositional Dynamic Logic (**PDL**)

detail, so that the reader who has not seen it will be able to see what it is, and how it is used in systems which we shall see later in the chapter.

The syntax of **PDL** begins with atomic sentences p, q, $\ldots$, (also called atomic propositions), and also *atomic programs* a, b, $\ldots$ (sometimes called *actions*). From these atomic sentences and programs, we build a language with two types of syntactic objects, called sentences and programs. The syntax is set out in Figure 5. The sentence-building operations include those of standard logical systems. In addition, if φ is a sentence and π a program, then $[\pi]\varphi$ is again a sentence. The intended meaning is "no matter how we run π, after we do so, φ holds." This formulation hints that programs are going to be non-deterministic, and so one of the syntactic formation rules does allow us to take the *union* (or *non-deterministic choice*) of π and σ to form $\pi \cup \sigma$. The other formation rules include composition (;), testing whether a sentence is true or not ($?\varphi$), and iteration (π^*).

The basic idea in the semantics is that we have state set S to start, and programs are interpreted in the most extensional way possible, as relations over S. So we are identifying the program with its input-output behavior; since we are thinking of non-deterministic programs, this behavior is a relation rather than a function. Atomic programs are interpreted as relations which are given as part of a model, and the rest of the programs and sentences are interpreted by a simultaneous inductive definition given in Figure 5. With this interpretation, each program π turns into a sentence-forming operation $[\pi]$; these then behave exactly as in standard relational modal systems. The clause for program composition uses composition of relations, and the one for iteration uses the reflexive-transitive closure operation.

PDL turns out to be decidable and to have a nice axiom system. The system resembles modal logic, and indeed one takes the basic axioms and rules for the operators $[\pi]$ given by programs. The other main axioms and rules are

(Test)	$\vdash [?\varphi]\psi \leftrightarrow (\varphi \to \psi)$
(Composition)	$\vdash [\pi;\sigma]\varphi \leftrightarrow [\pi][\sigma]\varphi$
(Choice)	$\vdash [\pi \cup \sigma]\varphi \leftrightarrow ([\pi]\varphi \wedge [\sigma]\varphi)$
(Mix)	$\vdash [\pi^*]\varphi \to \varphi \wedge [\pi][\pi^*]\varphi$

One also has an Induction Rule: From $\chi \to \psi$ $\chi \to [\pi]\chi$, infer $\chi \to [\pi^*]\psi$. The treatment of iteration is related to what we saw for common knowledge in Section 4.5; there is a common set of mathematical principles at work. For more on **PDL**, see, e.g., Harel, Kozen, and Tiuryn [43].

PDL and epistemic updates One important observation that links **PDL** with epistemic logic is that the changes in agents' accessibility relations as a result of an epistemic action of some sort or other are often given as *relation-changing* programs. We want to spell this out in detail, because it will be important in Section 7.2.

The semantics of a **PDL** sentence φ in a given model M may be taken to be a subset $[\![\varphi]\!] \subseteq M$, namely the set of worlds making φ true. Similarly, the semantics of a **PDL** program π may be taken to be a relation $[\![\pi]\!]$ on M. Now given a **PDL** program $\pi(r)$ with a *relation variable* r, we can interpret $\pi(r)$ by a function $[\![\pi(r)]\!]$ from relations on M to relations on M (a *relation transformer*): for each $R \subseteq M \times M$, we use R for the semantics of r and the rest of the semantics as above.

We shall see a simple example of this shortly, in Example 9 of Section 5.1.

5 Dynamic Epistemic Logic

We now move on to a different form of dynamics related to the topic of our chapter. Starting from the perspective of epistemic logic, knowledge and belief change can also be modeled by expanding the logic with dynamic modal operators to express such changes. The result is known as *Dynamic Epistemic Logic(s)*, or **DEL** for short. The first and the simplest form of dynamics is that associated with *public announcements*. It is simple from the perspective of change, but not particularly simple seen as an extension of epistemic logic. Public announcement logic is discussed in Section 5.1, and some related technical results of philosophical interest are presented in Section 5.2. Next, we move on to various forms of *private announcements* and to the associated dynamic logics, presented in Section 5.3. Even more complex types of dynamics, induced by various types of *epistemic actions*, are treated in Section 5.4. Finally, in Section 5.5 we briefly introduce logical languages and axioms for epistemic actions.

5.1 Public announcements

We first saw public announcements in Section 2.1. The example there was very simple indeed, and so to illustrate the phenomenon further, it will be useful to have a more complicated scenario. This discussion in this section is based on Example 2 in Section 4.4. Assume for the moment that the card deal is described by ♣♡♠: Amina holds clubs, Bao hearts, and Chandra spades. Amina now says ('announces') that she does not have the hearts card. Therefore she makes public to all three players that all deals where Hearts_a is true can be eliminated from consideration: everybody knows that everybody else eliminates those deals, etc. They can therefore be *publicly* eliminated. This results in a restriction of the model *Hexa* as depicted in Figure **??**.

At this point, we only consider announcements like this in states where the announcement is true. We view the public announcement "I do not have hearts" as an 'epistemic program'. We interpret it as an *state transformer* just as flipping the box was so interpreted in Section 4.2. This program is interpreted as an 'epistemic state transformer' of the original epistemic state, exactly as we saw in Section 4.9 for **PDL**. We want $[!\varphi]\psi$ to mean that after (every) truthful announcement of φ, the sentence ψ holds. Continuing to borrow terminology from dynamic logic, state transformers come with *preconditions*. In this case, we want the precondition to be $\neg \mathit{Hearts}_a$, so that we set aside the matter of false announcements.

The effect of such a public announcement of φ is the restriction of the epistemic state to all worlds where φ holds. So, 'announce φ' can indeed be seen as an epistemic state transformer, with a corresponding dynamic modal operator $[!\varphi]$.

We *appear* to be moving away slightly from the standard paradigm of modal logic. So far, the accessibility relations were between states in a given model underlying an epistemic state. But all of a sudden, we are confronted with an accessibility relation *between* epistemic states as well. "I do not have hearts" induces a(n) (epistemic) state transition such that the pair of epistemic states in Figure **??** is in that relation. The epistemic states take the role of the points or worlds in a seemingly underspecified domain of 'all possible epistemic states'. By lifting accessibility between points in the original epistemic state to accessibility between epistemic states, we can get the dynamic and epistemic accessibility relations 'on the same level' again, and see this as an 'ordinary structure' on which to interpret a perfectly ordinary multimodal logic. (There is also a clear relation here with interpreted systems, which will be discussed in Subsection 6.3, later.) A crucial point is that this 'higher-order structure' is induced by the initial epistemic state and the actions that can be executed there, and not the other way round. So it is standard modal logic after all.

Amina's announcement "I do not have hearts" is a simple epistemic action in various respects. *It is public.* A 'private' event would be when she learns that Bao has hearts without Bao or Chandra noticing anything. This required a more complex action description. *It is truthful.* She could also have said "I do not have clubs." She would then be lying, but, e.g., may have reason to expect that Bao and Chandra believe her. This would also require a more complex action description. *It is deterministic.* In other words, it is a state transformer. A non-deterministic action would be that Amina whispers into Bao's ear a card she does not hold, on Bao's request for that information. This action would have two different executions: "I do not have hearts", and "I do not have spades." Such more complex actions can be modeled in the action model logic presented in Section 5.4.

Language and semantics Add an inductive clause $[!\varphi]\psi$ to the definition of the language. For the semantics, add the clause:

$$M, s \models [!\varphi]\psi \quad \text{iff} \quad M, s' \models \varphi \text{ implies } M|\varphi, s \models \psi$$

where $M|\varphi = \langle S', R', V' \rangle$ is defined as

$$\begin{array}{lcl} S' & \equiv & \{s' \in S \mid M, s' \models \varphi\} \\ R'_a & \equiv & \{(s', t') \in S' \times S' : (s, t) \in R_a\} \\ V'_p & \equiv & \{s' \in S' : s \in V_p\} \end{array}$$

In other words: the model $M|\varphi$ is the model M restricted to all the states where φ holds, including access between states (a submodel restriction in the standard meaning of that term). It might be useful to look back at Section 2.1 for a discussion of the parallel case of probabilistic conditioning.

The language described above is called the language of public announcement, or *public announcement logic* (**PAL**).

Example 8 After Amina's announcement that she does not have hearts, Chandra knows that Amina has clubs (see Figure **??**). We can verify this with a semantic computation as follows: In order to check that $Hexa, ♣♡♠ \models [!\neg Hearts_a]K_c Clubs_a$, we have to show that $Hexa, ♣♡♠ \models \neg Hearts_a$ implies $Hexa|\neg Hearts_a, ♣♡♠ \models K_c Clubs_a$. The antecedent of this conditional being true, it remains to show that $Hexa|\neg Hearts_a, ♣♡♠ \models K_c Clubs_a$. The state $Hexa|\neg Hearts_a, ♣♡♠$ is shown in Figure **??**. Clearly, at the world ♣♡♠ in it, $K_c Clubs_a$.

Example 9 We mentioned at the end of Section 4.9 that program terms in **PDL** with variables may be used to specify the actions of epistemic actions. Here is how this works for public announcement of a sentence φ. For each agent a, the program $\pi(r)$ we want is $?\varphi; r; ?\varphi$. As previously explained, this defines a relation transformer $[\![\pi(r)]\!]$ on the underlying model. Then if R_a is agent a's accessibility relation before the announcement, $[\![\pi(r)]\!](R_a)$ is her accessibility relation afterwards. In more detail,

$$\begin{array}{lcl} [\![?\varphi; r; ?\varphi]\!](R_a) & = & \{(w,w) : w \in [\![\varphi]\!]\}; \{(u,v) : u\, R_a, v\}; \{(w,w) : w \in [\![\varphi]\!]\} \\ & = & \{(u,v) : u, v \in [\![\varphi]\!] \text{ and } u\, R_a, v\} \end{array}$$

The dual operators $\langle !\varphi\rangle$ Most people prefer to consider the *dual* $\langle !\varphi\rangle$ of $[!\varphi]$. That is, we take $\langle !\varphi\rangle\psi$ to an abbreviation for $\neg[!\,\varphi]\neg\psi$. This is equivalent to saying that $M, s \models \langle !\varphi\rangle\psi$ if and only if $M, s \models \varphi$ *and* $M|\varphi, s \models \psi$.

The point is that statements of the form $[!\varphi]\psi$ are conditionals and therefore are taken to be true when their antecedents are false; the duals are conjunctions. To see the difference, $\langle !\neg \mathit{Hearts}_a\rangle K_c \mathit{Clubs}_a$ is *false* at $(\mathit{Hexa}, \clubsuit\heartsuit\spadesuit)$.

Announcement and knowledge In general, $[!\varphi]K_a\psi$ is not equivalent to $K_a[!\varphi]\psi$. The easiest way to see this in our running example is to note that

$$\mathit{Hexa}, \clubsuit\heartsuit\spadesuit \not\models K_c[!\mathit{Hearts}_a]\mathit{Clubs}_a.$$

The correct equivalence in general requires that we make the truth of $[!\varphi]K_a\psi$ conditional on the truth of the announcement. So we get the following:

$$[!\varphi]K_a\psi \text{ is equivalent to } \varphi \to K_a[!\varphi]\psi.$$

Announcement and common knowledge Incidentally, the principle describing the interaction between common knowledge and announcement is rather involved. It turns out to be an *inference rule* rather than an axiom scheme. (One may compare it to the Induction Rule of **PDL** which we saw in Section 4.9; the rule here generalizes that one.) We therefore turn next to the proof system for validity in this logic.

A logical system See Figure 6 for a a proof system for this logic, essentially taken from [12]. It has precursors (namely completeness results for the logic *with* announcements but *without* common knowledge) in [77] and [38]; technically, this works out easier because the rules of the logic allow one to rewrite all sentences in a way that eliminates announcements altogether, and in this situation we may appeal to the known completeness result for epistemic logic. Thus the main point in the axiomatic work of [12] was the formulation of the Announcement Rule relating announcements and common knowledge, and the resulting completeness theorem.

Announcements are functional If an announcement can be executed, there is only one way to do it. So the partial functionality axiom in Figure 6 is sound. It is also convenient to write this as

$$\langle !\varphi\rangle\psi \to [!\varphi]\psi.$$

This is a simple consequence of the functionality of the state transition semantics for announcement. One might also say (from a programming perspective) that announcements are *deterministic.*

$[!\varphi]p \leftrightarrow (\varphi \rightarrow p)$	atomic permanence
$[!\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$	partial functionality
$[!\varphi]K_a\psi \leftrightarrow (\varphi \rightarrow K_a[!\varphi]\psi)$	announcement-knowledge

From $\chi \rightarrow [!\varphi]\psi$ and $\chi \wedge \varphi \rightarrow K_a\chi$ for all $a \in B$, infer $\chi \rightarrow [!\varphi]C_B\psi$
(the Announcement Rule)

Figure 6: The main points of the logic of public announcements. We have omitted the usual modal apparatus for modalities $[!\varphi]$.

Sequence of announcements A sequence of two announcements can always be replaced by a single, more complex announcement. Instead of first saying 'φ' and then saying 'ψ' you may as well have said for the first time 'φ, and after saying that ψ'. This is expressed in

$$[!\langle!\varphi\rangle\psi]\chi \leftrightarrow [!\varphi][!\psi]\chi.$$

This is useful when analyzing announcements that are made with specific intentions; or, more generally, conversational implicatures à la Grice. Intentions can be postconditions ψ that should hold after the announcement. So the (truthful) announcement of φ with the intention of achieving ψ corresponds to the announcement $\langle!\varphi\rangle\psi$.

If a sentence is common knowledge to all agents, there is no point in announcing it It will not change anyone's knowledge state:

$$C_A\varphi \rightarrow (K_a\psi \leftrightarrow [!\varphi]K_a\psi).$$

Here A is our set of all agents.

What can be achieved by public announcements? An interesting question, related to van Benthem's [97] discussion of Fitch's paradox presented in Section 3.6, is to characterize the sentences that *can come to be true through some (sequence of) public announcement(s)*, at a given state (in a given model). One answer is offered by the "ability" modality $\Diamond\varphi$, informally introduced in Section 3.6. Now we can formally define the semantics of $\Diamond\varphi$, by saying it is true at a state iff there exists some epistemic sentence ψ such that $\langle!\psi\rangle\varphi$ is true at that state. So the "ability" modality can be obtained by quantifying over public announcements (for all epistemic sentences). The above observation on the fact that a sequence of public announcements can be simulated by a single announcement shows that we do not have to iterate the defining clause of $\Diamond\varphi$: it already captures what can be achieved by any iteration. It also means that this modality satisfies the axioms of the system S4. Balbiani et al [7] call $\Diamond\varphi$ the "arbitrary announcement" modality, and give a complete axiomatization of the logic of arbitrary announcements, as well as studying its expressivity.

Alternative semantics There are some alternative semantics for public announcements. Gerbrandy [36, 37] and Kooi [49] propose a different semantics for announcements in a setting possibly more suitable for 'belief'. The execution of such announcements is not conditional to the truth of the announced formula.

vision	$\bigwedge_{a\in A}\bigwedge_{b\neq a}((\mathsf{d}_b \to K_a\mathsf{d}_b) \ \wedge\ (\neg\mathsf{d}_b \to K_a\neg\mathsf{d}_b))$
at least one	$\bigvee_{a\in A}\mathsf{d}_a$
background	$C_A(\text{vision} \wedge \text{at least one})$
nobody knows	$\bigwedge_{a\in A}(\neg K_a\mathsf{d}_a \wedge \neg K_a\neg\mathsf{d}_a)$
somebody knows	$\neg$nobody knows

Figure 7: Abbreviations in the discussion of the Muddy Children scenario, following [38].

Yet another semantics has recently been proposed by Steiner [89], to solve the problem of inconsistent beliefs that may be induced by public announcements. The semantics presented in this section has the disadvantage that updates induced by announcements do not necessarily preserve the seriality property (axiom D above): agents who have wrong (but consistent) beliefs may acquire inconsistent beliefs after a truthful public announcement. Steiner's alternative semantics solves this by proposing a modified semantics in which the new information is rejected if not consistent with prior beliefs. Yet another possible solution would be to incorporate some mechanism for belief revision, along the lines we discuss in Section 7.

Relativized common knowledge Recent developments in the area use a different modal notion, 'relativized common knowledge', of which standard common knowledge can be seen as a special case [102, 49]. Here is the idea. Add to the syntax an operation $C_B(\varphi, \psi)$. The semantics is

$$w \models C_B(\varphi,\psi) \qquad \text{iff} \qquad \begin{array}{l}\text{every path from } w \text{ using } \bigcup_{a\in B} \xrightarrow{a} \\ \text{consisting entirely of worlds where } \varphi \text{ holds} \\ \text{ends in a world where } \psi \text{ holds}\end{array}$$

This results in more a expressive logic. At the same time, the relation between announcements and relativized common knowledge turns into an axiom:

$$(E_B(\varphi \to \psi) \to C_B(\varphi, \psi \to E_B(\varphi \to \psi))) \to C_B(\varphi, \psi).$$

van Benthem, van Eijck and Kooi [102] contains the completeness proofs for this logic and others, and also various expressivity results.

Iteration The language of **PDL** has an iteration operator on actions, but this has not been reflected in any of our example scenarios. However, there are scenarios and protocols whose natural description uses action iteration. One example is the general form of the Muddy Children-type scenario, as we described it in Section 3.1. We discuss this in connection with the sentences in Figure 7. These are based on sentences in Gerbrandy and Groeneveld [38]. In them, d_a is an atomic sentence asserting that child a is dirty, and similarly for the other children. Informally, the sentence vision says that every child a can see and therefore knows the status of all other children. Note that vision is a (finite) sentence since the set A of agents (the children here) is finite. background says that it is common knowledge that vision and at least one hold. The intuition is that this is the background that the children have after the adult's announcement that at least one of them is dirty.

The sentence background is much weaker than what one would usually take to be the formalization of the overall background assumptions in the Muddy Children scenario.

However, it is enough for the following result:

$$\varphi_A \quad \equiv \quad \textsc{background} \rightarrow \langle \textsc{nobody knows}^* \rangle \textsc{somebody knows}. \tag{18}$$

Note the $*$ in (18). The formal semantics would make this equivalent to the infinitary sentence

$$\textsc{background} \rightarrow \bigvee_n \langle \textsc{nobody knows}^n \rangle \textsc{somebody knows}.$$

Either way, φ_A says that given the background assumption, some finite number of public announcements of everyone's ignorance will eventually result in the opposite: someone knowing their status.

Proposition 10 For each finite set A of children, $\models \varphi_A$.

For a proof, see Miller and Moss [67]. The point of Proposition 10 is that the statements φ_A are natural *logical validities.* So it makes sense to ask for a logical system in which such validities coincide with the provable sentences. The basic logic of announcements and common knowledge is known to be decidable, and indeed we have seen the axiomatization in Figure 6. However, it was shown in [67] that adding the *iterated announcement* construct that gives us the $\langle \textsc{nobody knows}^* \rangle$ operation results in logical systems whose satisfiable sentences are not decidable. The upshot is that (unfortunately) there is no hope of a finitely axiomatized logical system for the validities in a language which includes sentences like (18).

Notes The logic of multi-agent epistemic logic with public announcements and without common knowledge has been formulated and axiomatized by Plaza [77]. For the somewhat more general case of introspective agents, this was done by Gerbrandy and Groeneveld [38]; they were not aware of Plaza's work at the time. In [77], public announcement is seen as a binary operation $+$, such that $\varphi + \psi$ is equivalent to $\langle ! \varphi \rangle \psi$. The logic of public announcements *with* common knowledge was axiomatized by Baltag, Moss, and Solecki [12], see also [13, 8, 11], in a more general setting that will be discussed in Section 5.4: the completeness of their proof system is a special case of the completeness of their more general logic of action models. A concise introduction into public announcement logic (and also some of the more complex logics presented later) is found in [112]. A textbook presentation of the logic is [113]. This also contains a more succinct completeness proof than found in the original references. Results on complexity of the logic are presented by Lutz in [64].

There are a fair number of precursors of these results. One prior line of research is in dynamic modal approaches to semantics, not necessarily also epistemic: 'update semantics'. Another prior line of research is in meta-level descriptions of epistemic change, not necessarily on the object level as in dynamic modal approaches. This relates to the temporal epistemics and interpreted systems approach for which we therefore refer to the summary discussion in the previous section.

The 'dynamic semantics' or 'update semantics' was followed in van Emde Boas, Groenendijk, and Stokhof [114], Landman [52], Groeneveld [41], and Veltman [116]. However, there are important philosophical and technical differences between dynamic semantics and dynamic epistemic logic as we present it here. The main one is that update semantics interprets meaning (in natural language) as a relation between states, and so it departs from standard accounts. Nevertheless, the "dynamic" feature is common to both. Work taking propositional dynamic logic (**PDL**) in the direction of natural language semantics and related areas was initiated by van Benthem [93] and followed up in de Rijke [26] and Jaspars [47]. As background literature to various dynamic features introduced in the 1980s and 1990s we recommend van Benthem [93, 95, 94]. More motivated by runs in interpreted systems is van Linder, van der Hoek, and Meyer [115]. All these approaches use dynamic modal operators for information change, but (1) typically not (except [115]) in a multi-modal language that also has epistemic operators, (2) typically not for more than one agent, and (3) not necessarily such that the effects of announcements or updates are defined given the update formula and the current information state: the **PDL**-related and interpreted system related approaches *presuppose* a transition relation between information states, such as for atomic actions in **PDL**. We outline, somewhat arbitrarily, some features of these approaches. Groeneveld's approach [41] is typical for dynamic semantics in that is has formulas $[!\varphi]_a\psi$ to express that after an update of agent a's information with φ, ψ is true. His work was later merged with that of Gerbrandy, resulting in the seminal [38]. Gerbrandy's semantics of public announcements is given in [36], in terms of the *universe V_{AFA} of non-wellfounded sets*: this is a kind of "universal Kripke model"; i.e., a class in which every Kripke model can be embedded in a unique manner (up to bisimilarity; see the end of Section 4.4). In this way, one can avoid changing the initial model (by eliminating states and arrows) after a public announcement: instead, one just moves to another state in the same huge, all-encompassing Kripke super-model V_{AFA}. It was later observed in [73] that one can do with ordinary models.

De Rijke [26] defines theory change operators $[+\varphi]$ and $[*\varphi]$ with a dynamic interpretation that link an enriched dynamic modal language to AGM-type theory revision [1] (see also Section 7 addressing dynamic epistemics for belief revision). In functionality, it is not dissimilar to Jaspars' [47] φ-addition (i.e., expansion) operators $[!\varphi]_u$ and φ-retraction (i.e., contraction) operators $[!\varphi]_d$, called updates and downdates by Jaspars. Van Linder, van der Hoek, and Meyer [115] use a setting that combines dynamic effects with knowledge and belief, but to interpret various action operators they assume an explicit transition relation as part of the Kripke structure interpreting such descriptions.

As somewhat parallel developments to [36], we also mention Lomuscio and Ryan [63]. They do not define dynamic modal operators in the language, but they define epistemic state transformers that clearly correspond to the interpretation of such operators: $M * \varphi$ is the result of refining epistemic model M with a formula φ, etc. Their semantics for updates is only an *approximation* of public announcement logic, as the operation is only defined for *finite* (approximations of) models.

5.2 Sentences true after being announced

Moore sentences, revisited Recall that Moore sentences are *strongly unsuccessful*: they are always false after being announced. In terms of our public announcement logic, we can define strongly unsuccessful formulas φ as the ones such that the formula $[!\varphi]\neg\varphi$ is valid. An interesting open problem is to give a syntactic characterization of strongly unsuccessful sentences.

Successful formulas A more interesting and natural question is to characterize syntactically the *successful* formulas: those φ such that $[!\varphi]\varphi$ is valid. That is, whenever φ holds and is announced, then φ holds after the announcement. In our setting, it is easy to see that a successful formula has also the property that $[!\varphi]C_A\varphi$ is valid.

For example, the atomic sentences p are successful, as are their boolean combinations and also the sentences Kp. *Logically inconsistent formulas* are also trivially successful: they can never be truthfully announced, so after their truthful announcement everything is true (including themselves). *Public knowledge formulas* are also successful: $[!C_A\varphi]C_A\varphi$ is valid. This follows from bisimulation invariance under point-generated submodel constructions. On the negative side, even when both φ and ψ are successful, $\neg\varphi$ may be unsuccessful (for $\varphi = \neg p \vee Kp$), $\varphi \wedge \psi$ may be unsuccessful (for $\varphi = p$ and $\psi = \neg Kp$), and as well $[!\varphi]\psi$ and $\varphi \rightarrow \psi$ may be unsuccessful.

In its general form, the question of syntactically characterizing successful sentences remains *open*. But we present now two results on this problem.

Preserved formulas One successful fragment form the *preserved formulas* (introduced for the language without announcements by van Benthem in [96]) that are inductively defined as

$$\varphi ::= p \mid \neg p \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid K_a\varphi \mid C_B\varphi \mid [!\neg\varphi]\psi$$

(where $B \subseteq A$). From $\varphi \rightarrow [!\psi]\varphi$ for arbitrary ψ, follows $\varphi \rightarrow [!\varphi]\varphi$ which is equivalent to $[!\varphi]\varphi$; therefore preserved formulas are successful formulas. The inductive case $[!\neg\varphi]\psi$ in the 'preserved formulas' may possibly puzzle the reader. Its proof [108] is quite elementary (and proceeds by induction on formula structure) and shows that the puzzling *negation* in the announcement clause is directly related to the truth of the announcement as a *condition*:

Let $M, s \models [!\neg\varphi]\psi$, and $M' \subseteq M$ such that $s \in M'$. Assume $M', s \models \neg\varphi$. Then $M, s \models \neg\varphi$ by contraposition of the inductive hypothesis for φ. From that and $M, s \models [!\neg\varphi]\psi$ follows $M|\neg\varphi, s \models \psi$. From the inductive hypothesis for ψ follows $M'|\neg\varphi, s \models \psi$. Therefore $M', s \models [!\neg\varphi]\psi$ by definition.

Universal formulas A different guess would be that φ is successful iff φ is equivalent to a sentence in the *universal* fragment of modal logic, the fragment built from atomic sentences and their negations using K, $\wedge$, and $\vee$. However, this is not to be. We discuss work on single agent models whose accessibility relation is an *equivalence relation* in this discussion. It remains open to weaken this assumption and obtain similar results.

Suppose that φ and ψ are non-modal sentences (that is, boolean combinations of atomic sentences). Suppose that $\models \psi \rightarrow \varphi$. Consider $\varphi \vee \hat{K}\psi$. (Again, $\hat{K}$ is the dual of K, the "possibility" operator.) This is clearly not in general equivalent to a sentence in our fragment. Yet we claim that

$$\models [!(\varphi \vee \hat{K}\psi)](\varphi \vee \hat{K}\psi).$$

To see this, fix a state model M and some state s in it. If $s \in [\![\varphi]\!]$ in M, then since φ is non-modal, s "survives the announcement" and satisfies $\varphi \vee \hat{K}\psi$ in the new model. On the other hand, suppose that $s \in [\![\hat{K}\psi]\!]$ in M. Let $s \rightarrow t$ with $t \in [\![\psi]\!]$ in M. Then again, t survives and satisfies ψ and even $\varphi \vee \hat{K}\psi$. Hence, s satisfies $\varphi \vee \hat{K}\psi$ in the new model.

This example is due to Lei Qian. He also found a hypothesis under which the "first guess" above indeed holds. Here is his result. Let T_0 be the set of non-modal sentences. Let

$$T_1 \quad = \quad T_0 \cup \{K\varphi : \varphi \in T_0\} \cup \{\hat{K}\varphi : \varphi \in T_0\}.$$

Finally, let T_2 be the closure of T_1 under $\wedge$ and $\vee$.

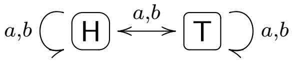
Theorem 11 (Qian [78]) Let $\varphi \in T_2$ have the property that $\models [!\varphi]\varphi$. Then there is some ψ in the universal fragment of modal logic such that $\models \varphi \leftrightarrow \psi$.

5.3 Varieties of privacy

As a warm up before meeting the general notion of "epistemic actions" in the next section, we present here two generalizations of public announcements: the first, called *fully private announcements*, is essentially due (modulo minor differences[7]) to Gerbrandy [38, 36], while the second, which we call *fair-game announcements*, is due to van Ditmarsch [105, 106]. Both can be regarded as forms of private announcements: some information is broadcast to an agent, or a group of agents, while being withheld from the outsiders. But there are important differences: a fully private announcement is so secret that the outsiders do not even suspect it is happening; while a fair-game announcement is known by outsiders to be possible, among other possible announcements.

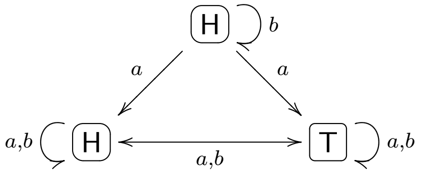
Fully Private Announcements to Subgroups For each subgroup B of agents, $!_B\varphi$ is the action of secretely broadcasting φ to all the agents in the group B, in a way that is completely oblivious to all outsiders $a \notin B$: they do not even suspect the announcement is taking place. An example of fully private announcement was encountered in Section 2.3: Bao is informed that the coin lies Heads up, but in such a way that Amina does not suspect that this is happening. The announcement is *truthful* (as in the previous section) but completely private: after that, Amina still believes that Bao doesn't know the state of the coin.

Assuming that before Bao entered, it was common knowledge that nobody knew that state of the coin, the belief/knowledge model *before the announcement* is a multi-agent version of the model



(19)

The situation after the fully private announcement (by which Bao is secretely informed that the coin lies Heads up) is given by the model (6) from Section 2.3. To recall, this was:



(20)

We can see that unlike the case of public announcements, the number of states *increases* after a fully private announcement. In fact, one can think of the model in the above picture as being obtained by putting together the initial model (19) and the model obtained from it by doing a public announcement, with the outsider (Amina) having doxastic arrows between the two submodels. In other words, the state transformer for a fully private announcement combines features of the original model with the one given by the state transformer for a public announcement.

Language and semantics Add an inductive clause $[!_B\varphi]\psi$ to the definition of the language. For the semantics, add the clause:

$$M, s \models [!_B\varphi]\psi \quad \text{iff} \quad M, s \models \varphi \text{ implies } M!_B\varphi, s' \models \psi$$

[7] As for public announcements, Gerbrandy's private announcements are not necessarily truthful. We present here a slightly modified version, that assumes truthfulness, in order to be able to subsume public announcements (as presented in Section 5.1) as a special case.

where $M!_B\varphi = \langle S' \cup S, R', V' \rangle$ is defined as

$$\begin{array}{rcl} S' & \equiv & \{s' \in S \mid M, s' \models \varphi\} \\ R'_a & \equiv & R_a \cup \{(s', t') \in S' \times S' : (s, t) \in R_a\} \\ V'_p & \equiv & V(p) \cup \{s' \in S' : s \in V(p)\} \end{array}$$

The language described above is called the logic of fully private announcements to subgroups. The axioms and rules are just as in the logic of public announcements, with a few changes. We must of course consider the relativized operators $[!_B\varphi]$ instead of their simpler counterparts $[!\varphi]$. The most substantive change which we need to make in Figure 6 concerns the Action-Knowledge Axiom. It splits into two axioms, noted below:

$$\begin{array}{ll} [!_B\varphi]K_a\psi \leftrightarrow (\varphi \rightarrow K_a[!_B\varphi]\psi) & \text{for } a \in B \\ [!_B\varphi]K_a\psi \leftrightarrow (\varphi \rightarrow K_a\psi) & \text{for } a \notin B \end{array}$$

The last equivalence says: assuming that φ is true, then after a private announcement of φ to the members of B, an outsider knows ψ just in case she knew ψ before the announcement.

Fair-game Announcements In a fair-game announcement, some information is privately learned by an agent or a group of agents, but the outsiders are aware of this possibility: it is publicly known that the announcement is one of a given list of possible alternatives, although only the insiders will known which one.

We illustrate fair-game announcements with two examples. Let us reconsider the epistemic state (*Hexa*, ♣♡♠) wherein Amina holds clubs, Bao holds hearts, and Chandra holds spades. It is shown in Figure **??**. Consider the following scenario:

> *Amina shows (only) Bao her clubs card. Chandra cannot see the face of the shown card, but notices that a card is being shown.*

It is assumed that it is publicly known what the players can and cannot see or hear. Call the action we are discussing showclubs. The epistemic state transition induced by this action is depicted in Figure **??**. Unlike after public announcements, in the showclubs action we cannot eliminate any state. Instead, all *b*-links between states have now been severed: whatever was the actual deal of cards, Bao now knows that card deal and cannot imagine any alternatives. We hope to demonstrate the intuitive acceptability of the resulting epistemic state. After the action showclubs, Amina considers it possible that Chandra considers it possible that Amina has clubs. That much is obvious, as Amina has clubs anyway. But Amina also considers it possible that Chandra considers it possible that Amina has hearts, because Amina considers it possible that Chandra has spades, and so does not know whether Amina has shown clubs or hearts. It is even the case that Amina considers it possible that Chandra considers it possible that Amina has spades, because Amina considers it possible that Chandra does not have spades but hearts, in which case Chandra would not have known whether Amina has shown clubs or spades. And in all those cases where Amina shows her card, Bao obviously would have learned the deal of cards. Note that, even though for Chandra there are only two possible actions—showing clubs or showing hearts—none of the *three* possible actions can be eliminated from public consideration.

But it can become even more complex. Imagine the following action, rather similar to the showclubs action:

> *Amina whispers into Bao's ear that she does not have the spades card, given a (public) request from Bao to whisper into his ear one of the cards that she does not have.*

This is the action whispernospades. Given that Amina has clubs, she *could* have whispered "no hearts" or "no spades". And whatever the actual card deal was, she could always have chosen between two such options. We obtain a model that reflects all possible choices, and therefore consists of $6 \times 2 = 12$ different states. It is depicted in Figure **??** (wherein we assume transitivity of the accessibility relation for c). There is a method of calculating complex representations like this one, and we shall discuss this particular model in Example 13 in the next section. But for now, the reader may look at the model itself to ascertain that the desirable postconditions of the action whispernospades indeed hold. For example, given that Bao holds hearts, Bao will now have learned from Amina what Amina's card is, and thus the entire deal of cards. So there should be no alternatives for Bao in the actual state (the underlined state ♣♡♠ 'at the back' of the figure—for convenience, different states for the same card deal have been given the same name). But Chandra does not *know* that Bao knows the card deal, as Chandra considers it possible that Amina actually whispered "no hearts" instead. That would have been something that Bao already knew, as he holds hearts himself—so from that action he would not have learned very much. Except that Chandra could then have imagined him to know the card deal ... Note that in Figure **??**, there is also another state named ♣♡♠, 'in the middle', so to speak, that is accessible for Chandra from the state ♣♡♠ 'at the back', and that witnesses that Bao doesn't know that Amina has clubs.

Notes The logic of fully private announcements has been first formulated and axiomatized by Gerbrandy [38, 36], in a slightly different version: as for public announcements, Gerbrandy's private announcements are not necessarily truthful. Also, Gerbrandy's semantics of fully private announcements, as the one of public announcements, is given in terms of non-wellfounded sets, rather than Kripke models. The version presented here (which assumes truthfulness and uses Kripke semantics) was formulated in Baltag and Moss [11], as a special case of a "logic of epistemic programs". Gerbrandy [36] considers more general actions: fully private announcements are only a special case of his operation of *(private) updating with an epistemic program.*

The logic of fair-game announcements is a special case of the work by van Ditmarsch [105, 106] and by Baltag, Moss, and Solecki [12, 11]; the latter call it "the logic of common knowledge of alternatives".

5.4 Epistemic actions and the product update

As we saw in the previous section, some epistemic actions are more complex than public announcements, where the effect of the action is always a restriction on the epistemic model. As in the previous examples, the model may grow in complex and surprising ways, depending on the specific epistemic features of the action. Instead of computing by hand the appropriate state transformer for each action, it would be useful to have a general setting, in which one could input the specific features of any desired action and compute the corresponding state transformer in an automatic way.

Action models We present a formal way to model such actions, and a large class of similar events, via the use of 'action models', originating in [12]. The basic idea is that the agents' uncertainty about actions can profitably be modeled by putting them in relation to other possible actions, in a way similar to how the agents' uncertainty about states was captured in a Kripke model by relating them to other possible states. When Amina shows her clubs card to Bao, this is indistinguishable for Chandra from Amina showing her hearts card to Bao—if she were to have that card. And, as Amina considers it possible that Chandra holds hearts instead of spades, Amina also considers it possible that Chandra interprets her card showing action as yet a third option, namely showing spades. These three different card showing actions are therefore, from a public perspective, all indistinguishable for Chandra, but, again from a public perspective, all different for Amina and Bao.

We can therefore visualize the 'epistemic action' of Amina showing clubs to Bao as some kind of Kripke structure, namely with a domain of three 'action points' standing for 'showing clubs', 'showing hearts', and 'showing spades', and accessibility relations for the three players corresponding to the observations above. We now have what is called an *action model*. What else do we need? To relate such 'action models' to the preconditions for their execution, we associate to each action point in such a model a formula in a logical language: the precondition of that action point.

To execute an epistemic action, we compute what is known as the *restricted modal product* of the current epistemic state and the epistemic action. The result is 'the next epistemic state'. It is a *product* because the domain of the next epistemic state is a subset of the cartesian product of the domain of the current epistemic state and the domain of the action model. It is *restricted* because we restrict that full product to those (state, action) pairs such that the precondition for the action of the pair is satisfied in the state of the pair. Two states in the new epistemic state are indistinguishable (accessible), if and only if the states in the previous epistemic state from which they evolved were already indistinguishable (accessible), and if the two different actions executed there were also indistinguishable. For example, Chandra cannot distinguish the result of Amina showing clubs in state ♣♡♠ from Amina showing hearts in state ♡♣♠, because in the first place she could not distinguish those two card deals, and in the second place she cannot distinguish Amina showing clubs from Amina showing hearts.

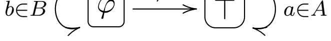
Remark This is perhaps a good point to make a comment on the terminology. What we are calling "action models" involve "actions" in an abstract sense, and so some of the important features of real actions are missing. For example, there is no notion of *agency* here: events like public announcement are modeled without reference to any agent(s) whatsoever as their source. Further, they may well be complex (many-step) actions, and for this reason they are also called *programs* in work such as [11]. So other authors have called our "action models" *event models*. We maintain the older terminology mainly because this is how it has appeared in the literature.

We now formally define action models and their execution, for any given logical language. We leave for later the problem of finding a good such language for describing epistemic actions and their effects. As usual, we assume background parameters in the form of a set of agents A and a set of propositional variables P.

Definition [Action model] Let $\mathcal{L}$ be a logical language. An *action model over* $\mathcal{L}$ is a structure $\mathsf{U} = \langle \mathsf{S}, \mathsf{R}, \mathsf{pre} \rangle$ such that S is a domain of *action points*, such that for each $a \in A$, R_a is an accessibility relation on S, and such that $\mathsf{pre} : \mathsf{S} \to \mathcal{L}$ is a precondition function that assigns a *precondition* $\mathsf{pre}(\sigma) \in \mathcal{L}$ to each $\sigma \in \mathsf{S}$. An *epistemic action* is a pointed action model (U, σ), with $\sigma \in \mathsf{S}$.

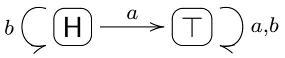
Example 12 The *public announcement* of φ is modeled by a singleton action model, consisting of only one action point, accessible to all agents, and having φ as its precondition. We call this action model *Pub* φ, and denote by $!\varphi$ the (action corresponding to the) unique point of this model. A more concrete example is the action $!\neg Hearts_a$ in Section 5.1 in which Amina publicly announces that she does not have the hearts card: the action model is $Pub \neg Hearts_a$.

A *fully private announcement* of φ to a subgroup B is modeled by a two-point action model, one point having precondition φ (corresponding to the private announcement) and the other point having precondition $\top := p \vee \neg p$ (corresponding to the case in which no announcement is made):



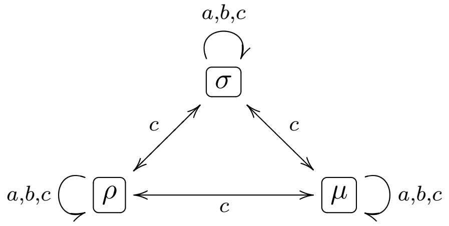
We call this action model $Pri_B \varphi$. Again, the action point on the left represents the fully private announcement of φ. This action will be denoted by $!_B \varphi$. The action point on the right has as precondition some tautology $\top$, and represents the alternative action in which *no announcement* is made: essentially nothing is happening. This action will be denoted by τ.

A more concrete example of a fully private announcement model is the action considered in Section 2.3, in which Bao was secretely informed that the coin lay heads up, without Amina suspecting this to be happening. This corresponds to the right-hand point in the action model Pri_b H:



Fair-game announcements with n commonly-known alternatives can be modeled using an action model with n points, having the corresponding announcements as preconditions. For the "insiders", the accessibility relation is the identity relation, while the accessibility for the "outsiders" is the universal relation (linking every two points).

As a concrete example of fair-game announcement, the action model U for the $\mathsf{showclubs}$ action in the previous section has three action points σ, ρ, and μ, with preconditions $\mathsf{pre}(\sigma) = Spades_a$, $\mathsf{pre}(\rho) = Hearts_a$, and $\mathsf{pre}(\mu) = Clubs_a$. The epistemic structure of this model is:



(21)

The pointed action model of interest is (U, μ). In it, the action μ which really happened is one where Amina and Bao come to share the knowledge that she has clubs: no other options are available to the two of them. Chandra, on the other hand, is in the dark about which of three announcements is taking place, but she does know the three possible messages: $Clubs_a, Spades_a, Hearts_a$.

Similarly, the action model U' for the action $\mathsf{whispernospades}$ in the previous section has the same structure as the model U above, except that we take: $\mathsf{pre}(\sigma) = \neg Clubs_a$, $\mathsf{pre}(\rho) = \neg Hearts_a$, $\mathsf{pre}(\mu) = \neg Spades_a$. The pointed action model of interest is (U', μ).

Definition [Execution, Product Update] Given an epistemic state (M, s) with $M = \langle S, R, V\rangle$ and an epistemic action (U, σ) with $\mathsf{U} = \langle \mathsf{S}, \mathsf{R}, \mathsf{pre}\rangle$. The result of executing (U, σ) in (M, s) is only defined when $M, s \models \mathsf{pre}(\sigma)$. In this case, it is the epistemic state $((M \otimes \mathsf{U}), (s, \sigma))$ where $(M \otimes \mathsf{U}) = \langle S', R', V'\rangle$ is a restricted modal product of M and U defined by

$$\begin{array}{lll} S' & \equiv & \{(s, \sigma) \mid s \in S, \sigma \in \mathsf{S}, \text{ and } M, s \models \mathsf{pre}(\sigma)\} \\ R'_a((s, \sigma), (t, \rho)) & \text{iff} & R_a(s, t) \text{ and } \mathsf{R}_a(\sigma, \rho) \\ (s, \sigma) \in V'_p & \text{iff} & s \in V_p \end{array}$$

This restricted product construction has become known in the **DEL** literature as "Product Update". Here, we simply call it *action execution*. The intuition is that *indistinguishable actions performed on indistinguishable input-states yield indistinguishable output-states*: if when the real state is s, agent a thinks it is possible that the state might be s', and if when action σ is happening, agent a thinks it is possible that action σ' might be happening, then after this, when the real state is (s, σ), agent a thinks it is possible that the state might be (s', σ').

Example 13 At this point we can go back and justify all our previous state transformers in a uniform manner. The model in Figure **??** can be obtained by calculating $\mathit{Hexa} \otimes ! \neg \mathit{Hearts}_a$, where *Hexa* is the model shown in Figure **??**, and the action model $\mathit{Pub}\neg\mathit{Hearts}_a$ was described above. The model (6) from Section 2.3 can be computed by calculating the restricted modal product of the model (19) from Section 5.3 and the action model $\mathit{Pub}_b\mathsf{H}$ above. The pointed model shown in Figure **??** is obtained by calculating

$$(\mathit{Hexa} \otimes \mathsf{U}, (\clubsuit\heartsuit\spadesuit, \mu)),$$

where U is from (21) above.

Finally, we justify the pointed model in in Figure **??**, by calculating

$$(\mathit{Hexa} \otimes \mathsf{U}', (\clubsuit\heartsuit\spadesuit, \mu)),$$

where the action model U' is as above. Let us look at this last calculation in more detail: the restricted product itself contains the twelve pairs

$$\begin{array}{c} (\clubsuit\heartsuit\spadesuit, \rho), (\clubsuit\heartsuit\spadesuit, \mu), (\clubsuit\spadesuit\heartsuit, \rho), (\clubsuit\spadesuit\heartsuit, \mu), (\heartsuit\clubsuit\spadesuit, \sigma), (\heartsuit\clubsuit\spadesuit, \mu), \\ (\heartsuit\spadesuit\clubsuit, \sigma), (\heartsuit\spadesuit\clubsuit, \mu), (\spadesuit\clubsuit\heartsuit, \sigma), (\spadesuit\clubsuit\heartsuit, \rho), (\spadesuit\heartsuit\clubsuit, \sigma), (\spadesuit\heartsuit\clubsuit, \rho) \end{array}$$

The valuation only looks at the first components. For example $(\heartsuit\spadesuit\clubsuit, \sigma) \models \mathit{Hearts}_a \wedge \mathit{Spades}_b \wedge \mathit{Clubs}_c$. The epistemic relations are determined in the usual way of products. For example, $R'_c((\spadesuit\heartsuit\clubsuit, \rho), (\heartsuit\spadesuit\clubsuit, \sigma))$ because Chandra cannot tell the difference between $\spadesuit\heartsuit\clubsuit$ and $\heartsuit\spadesuit\clubsuit$ in *Hexa*, and she also cannot tell the difference between ρ and σ in U.

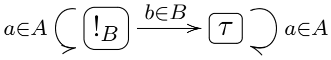
5.5 Logics for epistemic actions

There is only one more step to make: to give a logical language with an inductive construct for action models. The task of finding a natural general syntax for epistemic actions is not an easy problem. A number of different such languages have been proposed, see e.g., [36, 8, 11, 102, 105, 106, 111]. We follow [11], presenting here only one type of syntax, based on the notion of *signature.*

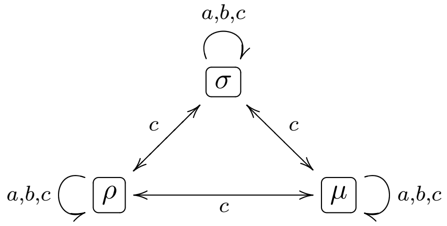
Definition [Signature] An *action signature* is a *finite* Kripke frame Σ, together with a *list* $(\sigma_1, \ldots, \sigma_n)$ enumerating some of the elements of Σ without repetitions. The elements of Σ are called *action types.*

Example 14 The *public announcement signature Pub* is a singleton frame, consisting of an action type !, accessible to all agents, and the list (!).

The signature Pri_B of *fully private announcements to a subgroup B* is a two-point Kripke frame, consisting of an action type $!_B$ (corresponding to fully private announcements) and an action type τ (for the case in which no announcement is made). The list is $(!_B)$, and the structure is given by:



The signature of *fair-game announcements* (to a given group of insiders, and with common knowledge of a given finite set of alternatives) can be similarly formalized. For instance, the signature $Show_{a,b}$ for the logic of the actions showclubs and whispernospades (with a, b as insiders and c as outsider) is a frame with three action types listed as (σ, ρ, μ). The structure is:



(22)

Definition [Language] For a given action signature Σ with a list $(\sigma_1, \ldots, \sigma_n)$, the language $\mathcal{L}_\Sigma$ of the logic of Σ-actions is the union of the *formulas* φ and the *epistemic actions*[8] α defined by

$$\begin{array}{lll} \varphi & ::= & p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid C_B\varphi \mid [\alpha]\varphi \\ \alpha & ::= & \sigma\varphi_1 \ldots \varphi_n \mid \alpha \cup \beta \end{array}$$

where $p \in P$, $a \in A$, $B \subseteq A$, $\sigma \in \Sigma$, and $\sigma\varphi_1 \ldots \varphi_n$ above is an expression consisting of a basic action σ followed by a string of n formulas, where n is taken from the listing in Σ.

The expressions of the form $\sigma\vec{\varphi}$ are called *basic epistemic actions.* In addition, we have included in the language $\mathcal{L}_\Sigma$ an operation of *non-deterministic choice* on the actions, mainly to show the reader familiar with dynamic logic, process algebra, and the like that it is possible to add such operations. One can also add *sequential composition, iteration* (Kleene star $*$), etc.

Definition [Action model induced by a signature] Given an action signature Σ with its list $(\sigma_1, \ldots, \sigma_n)$ of action types, and given a list $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ of n formulas in $\mathcal{L}_\Sigma$, the action model $\Sigma\vec{\varphi}$ is obtained by endowing the Kripke frame Σ with the following precondition map:

[8] We are using the letter α here for both action points and also for syntactic expressions denoting them. This ambiguity should not cause problems, but we wish to alert the careful reader of it.

if $\sigma = \sigma_i$ is in the given list $\vec{\sigma}$, we take $\mathsf{pre}(\sigma_i) := \varphi_i$; while, if σ is not in the given list $(\sigma_1, \ldots, \sigma_n)$, $\mathsf{pre}(\sigma)$ is taken to be some *tautology* $p \vee \neg p$. When seen as an action point in the action model $\Sigma\vec{\varphi}$, the point $\sigma \in \Sigma$ is denoted by $\sigma\vec{\varphi}$. Since the frame is the same as Σ, having the relation $\sigma\vec{\varphi} \xrightarrow{a} \sigma'\vec{\varphi}$ in the action model $\Sigma\vec{\varphi}$ is the same as having the relation $\sigma \xrightarrow{a} \sigma'$ in the frame Σ.

Example 15 The action model $Pub\ \varphi$ from the previous section is induced by the signature *Pri* above, in the obvious way. The action model $Pri_B\varphi$ from the previous section is induced by the signature *Pri* above. The action model U for the $\mathsf{showclubs}$ action in the previous section is induced by the signature $Show_{a,b}$, since it coincides with the model $Show_{a,b}Spades_a Clubs_a Hearts_a$. (This is an action signature followed by three propositions.) The model U' for the $\mathsf{whispernospades}$ action is induced by the same signature, since it can be written as $Show_{a,b}\neg Clubs_a \neg Hearts_a \neg Spades_a$.

Definition [Semantics]

$$\begin{array}{lll} M, s \models [\sigma\vec{\varphi}]\psi & \text{iff} & M, s \models \mathsf{pre}(\sigma\vec{\varphi}) \text{ implies } (M \otimes \Sigma\vec{\varphi}), (s, \sigma\vec{\varphi}) \models \psi \\ M, s \models [\alpha \cup \beta]\varphi & \text{iff} & M, s \models [\alpha]\varphi \text{ and } M, s \models [\beta]\varphi \end{array}$$

Note that the preconditions in an action model are arbitrary sentences in the language, since we want to talk about announcements concerning announcements and similar things. In fact, to avoid vicious circles, the definition of the semantics of $\mathcal{L}_\Sigma$ and the definition of action execution (as in the previous section) for action model over $\mathcal{L}_\Sigma$ should be taken to form one single definition (by simultaneous double induction) of both concepts. As usual, $\langle\alpha\rangle\varphi$ is defined by duality as $\neg[\alpha]\neg\varphi$.

It is easy to see that the logic of public announcements (**PAL**) from Section 5.1, the logic of fully private announcements and the logic of fair-game announcements are examples of signature-based logics. The only syntactic difference is the presence of modalities $[\tau\varphi]\psi$ in the signature-based language for the signature *Pri*; but it is easy to see that $[\tau\varphi]\psi$ is logically equivalent to ψ, and so this language reduces to the logic of fully private announcements.

The logical system for this language is a generalization of what we have seen for the public announcement logic earlier. A statement of it may be found in Baltag and Moss [11], and the completeness in the final version of Baltag, Moss, and Solecki [12]. For each of the operators of basic epistemic logic, one has a *Reduction Axiom* which allows one to push the dynamic (action) modalities past that operators, given a certain context. But the main difficulty comes in the combination of the action modality with common knowledge statements. An alternative system which uses the relativized common knowledge operators may be found in van Benthem, van Eijck and Kooi [102]. Since we are not going to need any of these systems or any of their interesting fragments, we leave matters at that. The only exception is the generalization of the Announcement-Knowledge Axiom, which we deem worth explaining in some detail.

The Action-Knowledge Axiom The Reduction Axiom for the K operator will be a generalization of of the Announcement-Knowledge Axiom, which we call the *Action-Knowledge*

Axiom: for every basic action α, we have

$$[\alpha]K_a\varphi \leftrightarrow \left(\mathsf{pre}(\alpha) \to \bigwedge_{\alpha \xrightarrow{a} \alpha'} K_a[\alpha']\varphi\right).$$

To state it in a more transparent form, we need the notion of *appearance of an action to an agent*: for each basic action α of our language and for each agent a, the *appearance of* α *to* a is the action

$$\alpha_a := \bigcup_{\alpha \xrightarrow{a} \alpha'} \alpha',$$

where $\bigcup$ is the non-deterministic choice of a (finite) set of actions. The action α_a describes the way action α *appears* to agent a: when α is happening, agent a thinks that (one of the deterministic actions subsumed by) α_a is happening. With this notation, the Action-Knowledge Axiom says that, for every basic action α, we have:

$$[\alpha]K_a\varphi \leftrightarrow (\mathsf{pre}(\alpha) \to K_a[\alpha_a]\varphi).$$

In other words: knowledge commutes with action modalities, modulo the satisfaction of the action's precondition and modulo the replacement of the real action with its appearance. One can regard this as a fundamental law governing the dynamics of knowledge, a law that may be used to compute or predict future knowledge states from past ones, given the actions that *appear* to happen in the meantime. The law embodies one of the important insights that dynamic-epistemic logic brings to the philosophical understanding of information change.

Notes The action model framework has been developed by Baltag, Solecki, and Moss, and has appeared in various forms [12, 13, 8, 11]. The signature-based languages are introduced in Baltag and Moss [11]. A final publication on the completeness and expressivity results is still in preparation. A different but also rather expressive way to model epistemic actions was suggested by Gerbrandy in [36]; this generalizes the results by Gerbrandy and Groeneveld in [38]. Gerbrandy's action language can be seen as defined by relational composition, interpreted on non-wellfounded set theoretical structures corresponding to bisimilarity classes of pointed Kripke models. Van Ditmarsch explored another relational action language—but based on standard Kripke semantics—[105, 106] and was influenced by both Gerbrandy and Baltag et al. His semantics is restricted to $S5$ model transformations. Van Ditmarsch et al. later proposed *concurrent epistemic actions* in [111]. How the expressivity of these different action logics compares is unclear. Recent developments include a proposal by Economou in [29]. *Algebraic axiomatizations* of a logic of epistemic actions may be found in [9] and [10], while a *coalgebraic* approach is in [24]. A logic that extends the logic of epistemic actions by allowing for *factual change* and by closing epistemic modalities under regular operations is axiomatized in [102]. A *probabilistic version of the action model framework* is presented by van Benthem, Gerbrandy and Kooi in [99]. For a more extensive and up-to-date presentation of dynamic epistemic logic (apart from the present contribution), see the textbook '*Dynamic Epistemic Logic*' by van Ditmarsch, van der Hoek, and Kooi [113].

6 Temporal Reasoning and Dynamic Epistemic Logic

It is very natural in a conversation about knowledge to refer to the past knowledge of oneself or others: *I didn't know that, but now I do.* We have already mentioned briefly the "Mr. Sum and

Mr. Product" puzzle, illustrating that agents' comments on the past ignorance and knowledge of others can lead to further knowledge. In addition, all treatments of the Hangman paradox mentioned in Section 3.7 must also revolve around the issue of temporal reasoning concerning the future.

We begin with a scenario in which agent's knowledge and ignorance reverses itself more than once. We present an example, due to Sack [80], because the natural summation of it involves statements about past knowledge.

Our three players Amina, Bao, and Chandra are joined by a fourth, Diego. They have a deck with two indistinguishable ♠ cards, one $\diamondsuit$ and one ♣. The cards are dealt, and in the obvious notation, the deal is (♠, ♠, $\diamondsuit$, ♣). We assume that the following are common knowledge: the distribution of cards in the deck, the fact that each player knows which card was dealt to them, and that they do not initially know any other player's card. Then the following conversation takes place:

i. Amina: "I do not have $\diamondsuit$."

ii. Diego: "I do not have ♠."

iii. Chandra: "Before (i), I knew φ: Bao doesn't know Amina's card. After (i), I did not know φ. And then after (ii), I again knew φ."

All three statements are intuitively correct. After Amina's statement, Chandra considers it possible that the world is w = (♠, ♣, $\diamondsuit$, ♠). In w after the announcement, Bao does know that Amina holds ♠, so φ is false. But Amina no longer reckons this world w to be possible after Diego's announcement. Indeed, she only considers possible v = (♠, ♠, $\diamondsuit$, ♣). And in v after both announcements, Bao thinks that (♣, ♠, ♠, $\diamondsuit$) is possible. Hence φ holds, and Chandra knows that it does.

Our first order of business is to extend the kind of modeling we have been doing to be able to say the sentence in (iii), and also to prove it in a logical system.

6.1 Adding a 'yesterday' operator to the logic of public announcements

To get started, we present here the simplest temporal extension of the simplest dynamic epistemic logic, the logic of public announcements from Section 5.1. We think of a multi-agent epistemic model M_0 subject to a sequence of public announcements of sentence φ_1, φ_2, $\ldots$, φ_n. These determine models M_i: M_0 is given, and for $i < n$, M_{i+1} is given by taking submodels via $M_{i+1} = M_i|\varphi_{i+1}$. We add a single operation Y to the language, with the intended semantics that $Y\varphi$ means that φ was true before the last announcement. Formally, we would set

$$M_i, w \models Y\varphi \quad \text{iff} \quad M_{i-1}, w \models \varphi \tag{23}$$

There are two problems here, one minor and one more significant. The slight problem: what to do about sentence $Y\varphi$ in the original model M_0? The choice is not critical, and to keep our operators $\Box$-like, we'll say that all sentences $Y\varphi$ are automatically true in M_0, w.

The larger problem has to do with the semantics of public announcement sentences $[!\psi]\chi$. We know how to deal with announcement of one of the φ sentences with which we started, since these figure into the definition of the models M_i. But for announcement of other sentences, those models are of no help. One solution is to think in terms of *histories*

$$H \quad = \quad (M_0, \varphi_1, M_1, \varphi_2, \ldots, M_{n-1}, \varphi_n, M_n) \tag{24}$$

Again, we require that the models and sentences be related by $M_{i+1} = M_i|\varphi_{i+1}$. We recast (23) as a relation involving a history H as in (24) and a world $w \in M_n$:

$$\begin{array}{lll} H, w \models Y\varphi & \text{iff} & i = 0, \text{ or } w \in M_{n-1} \text{ implies } (M_0, \varphi_1, \ldots, M_{n-1}), w \models \varphi \\ H, w \models [!\psi]\chi & \text{iff} & H, w \models \psi \text{ implies } (M_0, \varphi_1, \ldots, M_n, \psi, M_n|\psi), w \models \chi \end{array}$$

So the effect of public announcements is to extend histories.

We turn to the logical principles that are reflected in the semantics. The decision to have Y be $\Box$-like means that the distribution axiom and the rule of necessitation formulated with Y are going to be sound for the logic. Here are the additional logical principles that are sound for this semantics (true in all worlds in all models in all histories):

$$\begin{array}{ll} (p \to Yp) \wedge (\neg p \to Y\neg p) & \text{atomic permanence} \\ \neg Y\bot \to (Y\neg\varphi \to \neg Y\varphi) & \text{determinacy} \\ (Y\bot \to K_a Y\bot) \wedge (\neg Y\bot \to K_a\neg Y\bot) & \text{initial time} \\ (\varphi \to \psi) \leftrightarrow [!\varphi]Y\psi & \text{action-yesterday} \\ YK_a\varphi \to K_a Y\varphi & \text{memory} \end{array}$$

These are due to Yap [123] and Sack [80]. In these, p must be atomic. And $\bot$ is a contradiction, so $Y\bot$ is only true at the runs of length 1. Most of the axioms are similar to what we have seen in other systems, except that one must be careful to consider those runs of length 1. The initial time axiom implies that it is common knowledge whether the current history is of length 1 or not The memory axiom is named for obvious reasons. Notice that the converse is false.

Sack's dissertation [80] also contains the completeness proof for this logic, with common knowledge operators added. In fact, his work also includes operators for the complete past (not just the one step 'yesterday'), the future, and also arbitrary epistemic actions formulated in the same language. This means that one can model private announcements concerning the past knowledge of other agents, to name just one example. His language also contains *nominals* to allow reference to particular states (we do not discuss these here) and also allows agents to, in effect, know what epistemic action they think just took place.

Returning to the flip-flop of knowledge In the previous section we presented a scenario that involved statements of previous knowledge and ignorance. Here is how this is formalized. Let φ be $\neg(K_b \mathit{Spades}_a \vee K_b \mathit{Diamonds}_a \vee K_b \mathit{Clubs}_a)$. Then a formalized statement of the entire conversation would be

$$\langle !\neg \mathit{Diamonds}_a\rangle\langle !\neg \mathit{Spades}_d\rangle(YYK_a\varphi \wedge Y\neg K_a\varphi \wedge K_a\varphi).$$

All of the background information about the scenario and the initial deal can be written as a sentence ψ in the language, assuming that we have common knowledge operators. Then the fact that we have a completeness result means that $\psi \vdash \varphi$ in the proof system. The logic is moreover decidable. As a result, it would be possible to have a computer program find a formal proof for us.

6.2 The future

Adding temporal operators for the future is more challenging, both conceptually and technically. To see this, let us return to the modeling of *private announcements* which we developed in Example 12 in Section 5.4. The way we modeled things, private announcements to groups seem to come from nowhere, or from outside the system as a whole. Let us enrich this notion just a bit, to see a simple setting in which temporal reasoning might be profitably modeled.

Consider a setting where each individual agent a might send a message m to some set B of agents, with the following extra assumptions: (0) m is a sentence in whatever language we are describing; (1) the names of the recipient agents B are written into m; (2) sending m take arbitrarily long, but eventually each agent in B will receive m; (3) all agents in B receive m at the same time; (4) the sending and receipt of messages is completely private; (5) at each moment, at most one message is sent or received; (6) messages are delivered in the order sent. We make these assumptions only to clarify our discussion, not because they are the most realistic or useful.

One might like to have temporal operators in the language so that agents can "say" sentences like *at some future point, all agents in B will receive the message I just sent, resulting in the common knowledge for this group of φ*, or *I sent m_1 and then m_2 to b, and the message I received just now from b shows me that it was sent between the time b received m_1 and the time he received m_2*.

At the time of this writing, there are no formalized systems which include knowledge and temporal operators, epistemic actions as we have been presenting them in this chapter, and also have temporally extended events such as the *asynchronous message passing* we have just mentioned. There is a separate tradition from the computer science literature which incorporates temporally extended events, knowledge, and temporal assertions along different lines than this chapter. We are going to present the ideas behind one of those approaches, that of *interpreted systems.*

Before that, we want to mention a different approach, the *history-based semantics* of messages due to Parikh and Ramanujam [76]. This work has a different flavor than interpreted systems. (But the two are equivalent in the sense the semantic objects in them may be translated back and forth preserving truth in the most natural formal language used to talk about them. See Pacuit [75] for this comparison.) The only reason we present the interpreted systems work instead is that it has a larger literature.

6.3 Interpreted systems and temporal epistemic logic

A general framework involving information change as a feature of *interpreted systems* was developed by Halpern and collaborators in the 1990s [30]. There are a few basic notions.

We start with a collection of agents or processors, each of which has a *local state* (such as 'holding clubs' for agent Amina), a *global state* is a list of all the local states of the agents involved in the system, plus a state of the environment. The last represents actions, observations, and communications, possibly outside the sphere of influence of the agents. An example global state is $(\clubsuit\heartsuit\spadesuit, \emptyset)$ wherein Amina has local state $\clubsuit$, i.e., she holds clubs, Bao local state $\heartsuit$, and Chandra local state $\spadesuit$, and where 'nothing happened so far in the environment,' represented by a value $\emptyset$. It is assumed that agents know their local state but cannot distinguish global states from one another when those states have the same local state. This induces an equivalence relation among global states that the reader will will play the role of an accessibility relation. Another crucial concept in interpreted systems is that of a *run*: a run is a (typically infinite) sequence of global states. For example, when Amina says that she does not have hearts, this corresponds to a transition from global state $(\clubsuit\heartsuit\spadesuit, \emptyset)$ to global state $(\clubsuit\heartsuit\spadesuit, \mathsf{nohearts})$. Atomic propositions may also be introduced to describe facts.

For example, not surprisingly, one may require an atom Hearts_a to be true in both global state $(\clubsuit\heartsuit\spadesuit, \emptyset)$ and in global state $(\clubsuit\heartsuit\spadesuit, \mathsf{nohearts})$.

Formally, a *global state* $g \in \mathcal{G}$ is a tuple consisting of local states g_a for each agent and a state g_ϵ of the environment. A *run* $r \in \mathcal{R}$ is a sequence of global states. The m-th global state occurring in a run r is referred to as $r(m)$, and the local state for agent a in a global state $r(m)$ is written as $r_a(m)$. An *interpreted system* $\mathcal{I}$ is a pair $(\mathcal{G}, \mathcal{R})$ consisting of a set of global states $\mathcal{G}$ and a set of runs $\mathcal{R}$ relating those states.

A *point* (r, m) is a pair consisting of a run and a point in time m—this is the proper abstract domain object when defining epistemic models for interpreted systems. In an interpreted system, agents cannot distinguish global states from one another iff they have the same local state in both, which induces the relation shown below:

$$(r, m) \overset{a}{\sim} (r', m') \text{ iff } r(m) \overset{a}{\sim} r'(m') \text{ iff } r_a(m) = r'_a(m')$$

(For an indistinguishability relation that is an equivalence, we usually write $\sim$ instead of R.) With the obvious valuation for local and environmental state values, that defines an epistemic model. For convenience we keep writing $\mathcal{I}$ for that. Given a choice of a *real* (or *actual*) point (r', m'), we thus get an epistemic state $(\mathcal{I}, (r', m'))$. Epistemic and (LTL) temporal (next) operators have the interpretation

$$\begin{array}{lll} \mathcal{I}, (r, m) \models X\varphi & \text{iff} & \mathcal{I}, (r, m+1) \models \varphi \\ \mathcal{I}, (r, m) \models K_a\varphi & \text{iff} & \text{for all } (r', m') : (r, m) \overset{a}{\sim} (r', m') \text{ implies } \mathcal{I}, (r', m') \models \varphi \end{array}$$

It will be clear that subject to some proper translation (see e.g. [62]) interpreted systems correspond to some subclass of the $S5$ models: all relations are equivalence relations, but the interaction between agents is even more than that. The relation between Kripke models and interpreted systems is not entirely trivial, partly because worlds or states in Kripke models are abstract entities that may represent the same set of local states. The main difference between the treatment of dynamics in interpreted systems and that in dynamic epistemics is that in the former this is encoded in the state of the environment, whereas in the latter it emerges from the relation of a state (i.e., an abstract state in a Kripke model) to other states.

Example For a simple example, consider the single agent the case of our three players as usual. Suppose that Amina holds clubs, and the hearts card is on top of the spades card (both facedown) on the table. She may now be informed about the card on top of the stack. This is represented by the interpreted system depicted in Figure **??**. It consists of four global states. The card Amina holds represents her local state. The other cards are (in this case, unlike in the three-agent card deal) part of the environment. The state of the environment is represented by which of the two cards is on top, and by an 'observation' state variable *obs* that can have three values $u\heartsuit$, $y\heartsuit$, and $n\heartsuit$, corresponding to the state before the announcement which card is on top, the state resulting from the announcement that hearts is on top, and the other state resulting from the announcement that it is at the bottom. The valuation V is now such that $V(\mathit{Clubs}_a) = \{(\clubsuit\heartsuit\spadesuit, u\heartsuit), (\clubsuit\spadesuit\heartsuit, u\heartsuit), (\clubsuit\heartsuit\spadesuit, y\heartsuit), (\clubsuit\spadesuit\heartsuit, n\heartsuit)\}$, and $V(\mathit{Hearts}_t) = \{(\clubsuit\heartsuit\spadesuit, u\heartsuit), (\clubsuit\heartsuit\spadesuit, y\heartsuit)\}$. The system consists of two runs, one from $(\clubsuit\heartsuit\spadesuit, u\heartsuit)$ to $(\clubsuit\heartsuit\spadesuit, y\heartsuit)$ (optionally extended with an infinite number of idle transitions), and the other run from $(\clubsuit\spadesuit\heartsuit, u\heartsuit)$ to $(\clubsuit\spadesuit\heartsuit, n\heartsuit)$. One can now compute that in the actual state $(\clubsuit\heartsuit\spadesuit, u\heartsuit)$ it is true that $\neg K_a \mathit{Hearts}_t$, but in state $(\clubsuit\heartsuit\spadesuit, y\heartsuit)$ she has learned that hearts is on top: $K_a \neg \mathit{Hearts}_t$ is now true. For another example: in the actual state $X K_a \mathit{Hearts}_t$. How the treatment of announcements in interpreted systems relates to public announcement logic will be made precise at the end of the following section.

Interpreted systems have been highly successful as an abstract architecture for multi-agent systems, where agents are either human operators or computer processors, and where the assumption that an agent 'knows its own state' is a realistic simplification. For that reason they can be said to model interaction between *ideal agents*. This assumption is also implicitly applied when modeling perfectly rational agents as in game theory and economics. Also, given that all the dynamics is *explicitly* specified in the runs through the system, it combines well with temporal epistemic logics (LTL, CTL) wherein dynamics is *implicitly* specified by referring to an underlying structure wherein such a change makes information sense. Temporal epistemic logics have been fairly successful. Their computational properties are well-known and proof tools have been developed. See, for example, [104, 27, 42]. The work of Fagin et al. [30] also generated lots of complexity results on knowledge and time, we also mention the work of van der Meyden in this respect, e.g. [103, 104].

There are two rather pointed formal differences between the temporal epistemic approach and the dynamic epistemic approach.

Closed versus open systems First, the temporal epistemic description takes as models systems together with their whole (deterministic) history and future development, in the shape of 'runs'. As such, it can be easily applied to 'closed' systems, in which all the possible developments are fixed in advance, where there are no accidents, surprises or new interactions with the outside world, and thus the future is fully determined. Moreover, in practice the approach is more applicable to closed systems having a *small* number of possible moves: that's the only ones for which it is feasible to work explicitly with the transition graph of the full history.

The dynamic epistemic approach is better suited to 'open' systems. This is for example the case with epistemic protocols which can be modified or adapted at any future time according to new needs, or which can interact with an unpredictable environment. But it is also applicable to closed systems in which the number of possible different changes is large or indefinite.

There are two analogies here. The first is with open-versus-closed-system paradigms in programming. People in concurrency are usually interested in open systems. The program might be run in many different contexts, in the presence of many other programs, etc. More recently (in the context of mobile computation), people have looked at approaches that allow programs to be changed at any time inside the same logical frame. The temporal logic approach is not fit for this, since it assumes the full current program to be fixed and given as 'the background model'. That is why people in this area have used totally different kinds of formalisms, mainly process algebraic, such as the π-calculus. In contrast to that, dynamic epistemic logics are interesting in that, although based on a modal logic, which is not an algebraic kind of formalism, they are able to express changes in an open system through the semantic trick of changing the models themselves, via 'epistemic updates'.

The second analogy is with *game theory*. The temporal approach is like the description of a game through explicitly giving its full extensive form: the graph of all possible plays. For instance, chess (in this approach) is defined as the set of all possible chess plays. But there is another way to describe a game: by giving only the 'rules of the game' (which type of actions are allowed in which type of situations), together maybe with an 'initial state' (or set of states) and some 'winning rules'. This is a much more economical and way to describe a game, and it is more common as well. Of course, once this description is given, one could draw the game in extensive form as the set of all plays, if one is given enough computational power… If we neglect the aspects of the game that deal with who wins (and what), the dynamic epistemic approach can naturally describe epistemic games in precisely this way: one gives an epistemic Kripke model of 'initial states' and also an epistemic Kripke model or other semantically precise description of possible 'epistemic actions', including preconditions that tell us on which type of states a given action may be applied. Then one can play the game by repeatedly updating the state model with the action model. A 'full play' or 'run' of the game is obtained when we reach a state (at the end of many updates) on which no action (in our given action model) can be applied.

Information change description The second difference between the interpreted systems and the dynamic epistemics approach simply concerns the ability to model and classify various 'types' or 'patterns' of information change, or information exchange, such as public announcements, private announcements, game announcements etc. The dynamic epistemic approach obviously has this in-built ability, while the temporal approach doesn't have it, at least not in a direct, usable manner. There is nothing like an "announcement". All of the structure is encoded in the set of runs that serves as a model. Even the semantics of knowledge uses this set of runs, and so if one wants to use this as a model of real knowledge, it means that the agents must have implicit access to the overall model. To put it differently, in the temporal approach, one can only say what is true 'before' and 'after' a given action, and thus only implicitly get some information about the type of the action itself, through its input-output behavior. Moreover, this information is *not* enough to isolate the type of the action, since it only gives us the *local* input-output behavior of a given action; and different actions may behave identically in one local context, but differ in general. For instance in the two players and two cards case, in an epistemic state in which the fact that the card deal is ♣♡♠ is common knowledge, a public announcement of that fact will have the same input-output description as a 'skip' action corresponding to 'nothing happens'. But in the epistemic state where the cards were dealt but not seen, or the subsequent one where all players only know their own card, this fact was not common knowledge and its public announcement will in that case induce an informative (i.e. non-skip) transition. For the same reason, actions like private announcements, announcements with suspicion, etc., are harder to model in the interpreted systems approach.

A number of people are investigating the relation between dynamic epistemic logic and either interpreted systems or history-based models. One should see, for example, van Benthem and Pacuit [101] for hints in this direction and also for related work on temporal epistemic reasoning.

7 Belief Change and Dynamic Epistemic Logic

Our final section is concerned with the interaction of **DEL** with the topic of belief revision. The material of this section is very new and still in a state of flux, so our discussion here cannot claim in any way to represent the definitive word on the matter.

Here is our plan for the section: First, we briefly present the classical **AGM** theory of belief revision. We then briefly mention some dynamic (but non-epistemic) versions of **AGM**. Finally, we present some of the recent work that incorporates belief revision into the **DEL** framework, in an attempt to overcome the above-mentioned classical problems: the work of van Benthem on the dynamic logic of belief upgrades, the action plausibility models of Aucher and van Ditmarsch, and the action-priority update of Baltag and Smets. As before, we follow a "logical" rather than a historical order, leaving the history for the Notes at the end of each section.

Classical AGM theory A *belief set* (or *theory*) is a set K of sentences in some language. We at first take the language to be propositional logic, but we are keen to extend this to various modal languages, where the modalities are either one of the knowledge or belief modalities which we have already seen, or an operation coming from this subject itself.

The notion of a belief set is intended to model the set of sentences believed by some agent. So to incorporate the reasoning of the agent, one usually works on top of some logical system or other and then requires belief sets to be closed under deduction in the system; they need not be consistent, however. They certainly need not be complete either: we might have $\varphi \notin K$ and $\neg\varphi \notin K$ as well. The **AGM** theory of belief revision deals with changes to an agent's belief set when presented with a new sentence φ. The main point is that φ might conflict with what the agent believes, and so the theory is exactly about this issue. The theory is named for its founding paper, the celebrated Alchourrón, Gärdenfors, and Makinson [1]. Overview publications include Gärdenfors [35] and most notably for us, Chapter 4c by Hans Rott.

The **AGM** theory employs three basic operations and presents postulates concerning them. Since belief sets are *sets*, the overall theory is second-order. Moreover, it is an interesting issue to then construct and study semantic models of the **AGM** postulates, or of related ones.

The first operation is called *expansion*. Intuitively, this is what happens when the agent takes K as a given and simply adds φ as a new belief. We write $K + \varphi$ for the result. The postulates for expansion as follows:

(1)	Closure	$K + \varphi$ is a belief set.
(2)	Success	$\varphi \in K + \varphi$
(3)	Inclusion	$K \subseteq K + \varphi$
(4)	Vacuity	If $\varphi \in K$, then $K = K + \varphi$.
(5)	Monotonicity	If $J \subseteq K$, then $J + \varphi \subseteq K + \varphi$.
(6)	Minimality	$K + \varphi$ is the minimal set with (1) – (5).

It is easy to check that these postulates exactly capture the operation of taking the consequences of $K \cup \{\varphi\}$ in the underlying logical system.

More interesting are the other two operations, *contraction* and *revision*. Intuitively, the contraction of K by φ models the result of the agent's giving up the belief in φ and doing this without giving up too much. Revision models changing K to definitely include φ. There are postulates for both operations, and we are only going to spell out those for revision. The reason is that on top of the postulates for expansion, those of revision determine the contraction operation (and vice-versa). We write the revision operation as $K * \varphi$. The postulates are:

(1)	Closure	$K * \varphi$ is a belief set.
(2)	Success	$\varphi \in K * \varphi$
(3)	Inclusion	$K * \varphi \subseteq K + \varphi$
(4)	Preservation	If $\neg\varphi \notin K$, then $K + \varphi \subseteq K * \varphi$.
(5)	Vacuity	$K * \varphi$ is inconsistent iff $\neg\varphi$ is provable.
(6)	Extensionality	If φ and ψ are equivalent, then $K * \varphi = K * \psi$
(7)	Subexpansion	$K * (\varphi \wedge \psi) \subseteq (K * \varphi) + \psi$
(8)	Superexpansion	If $\neg\psi \notin K * \varphi$, then $K * (\varphi \wedge \psi) \supseteq (K * \varphi) + \psi$.

The result of a contraction operation $K - \varphi$ satisfying some postulates which we did not list turns out to be the same as $K \cap (K * \neg\varphi)$; this is called the *Harper identity.* And given a contraction operation satisfying the postulates, one can define revision by the *Levi identity* $K * \varphi = (K - \neg\varphi) + \varphi$; this operation will then satisfy the eight postulates above.

One important issue in the area is the relation between belief revision and the older topic of conditional logics which began with Lewis' book [59]. To see what this is about, assume that we are working over a logical system with a symbol $\Rightarrow$ that we want to use in the modeling of some natural language conditional, say the subjunctive one. Then a belief set K might well contain sentences $\varphi \Rightarrow \psi$ and $\neg\varphi$. So in this context, we would like or even expect to have $\psi \in K * \varphi$. In other words, we ask about the condition

$$\varphi \Rightarrow \psi \in K \quad \text{iff} \quad \psi \in K * \varphi.$$

This is called the *Ramsey test.* A key result in the subject is Gärdenfors' *Impossibility Theorem*: there is no operation of revision on belief sets which both satisfies the postulates of $*$ and also the Ramsey test. (The result itself depends on some non-triviality condition which we ignore here.)

Although the literature on belief revision may be read as a discussion of changes in belief, it may also be read as an extended discussion about the correspondence between various axiom systems and types of semantic structures. These include structures akin to what we have seen. In particular, the *sphere systems* of Grove (based on earlier work of Lewis) come from belief revision theory.

7.1 Dynamic versions of revision theory

Moving now to a more *semantical* setting, we show how some of the operations which we have already seen can be interpreted as belief change operators. We then present some dynamic versions of **AGM**: the Katsuno-Mendelzon theory **KM** of belief update, de Rijke's dynamic modal logic **DML** and Segerberg's dynamic doxastic logic **DDL**. These are all dynamic in some sense, but not "epistemic" in that knowledge is not modeled via the Kripke (relational) semantics, or any other semantics for that matter. We mention some of the difficulties and problems encountered by classical belief revision theory.

Examples of belief change via dynamic epistemic logic Consider expressing and changing uncertainty about the truth of a single fact p, and assume an information state where the agent (whose beliefs are interpreted by the unlabeled accessibility relation depicted) may be uncertain about p and where p is actually false (indicated by 'designating' the actual state by underlining it). Figure **??** lists all conceivable sorts of belief change.

In the top structure, uncertainty about the fact p (i.e., absence of belief in p and absence of belief in $\neg p$) is changed into belief in $\neg p$. On the left, $\neg Bp$ is true, and on the right $B\neg p$. In the second from above, belief in p is weakened to uncertainty about p, and in the third from above we change from Bp to $B\neg p$. Note that also in this semantic setting of Kripke-structure transformation, belief revision can again be seen as a contraction followed by an expansion, so we may in principle consider semantic alternatives for the Levi-identity. The last information state transition in Figure **??** depicts factual change. The state with changed valuation has suggestively been renamed from 1 to 00, although formally, of course, it is only the valuation of a named state that changes. The 'assignment' or substitution $p := \bot$ indicates that the valuation of atom p is revised into the valuation of the assigned formula. As this is $\bot$, the new valuation of p (seen as a subset of the domain) is now the empty set of states.

Updates and the KM theory A topic in traditional belief revision comes under the name of 'update'. An *update*—unfortunately a clash cannot be avoided with the more general meaning of that term in dynamic epistemic logic, where it incorporates belief revision as well—is a *factual* change, as opposed to a belief change in the three previously distinguished notions. The latter merely express a different agent stance towards a non-changing world, but in an 'update' the world itself changes. The standard reference for updates in belief revision is Katsuno and Mendelzon [48]. Recent investigations on updates (factual change) in a dynamic epistemic setting are [104, 95, 49]. These ideas also deserve to be properly applied to the belief revision arena.

We mention one of the motivating examples, mainly to contrast with the **AGM** postulates. It is taken directly from [48].

Consider a belief set with two atomic propositions, b and m, standing for "there is a book on the floor" and "there is a magazine on the floor". Suppose that

$$K \quad = \quad \uparrow \{b \wedge \neg m, m \wedge \neg b\},$$

where the arrow denotes the deductive closure in propositional logic. This models a situation in which an agent believes that exactly one item is on the floor, but does not have a specific belief of which it is. Suppose we wish to change the world by instructing a robot (as they have it) to put the book on the floor. So we wish to consider $K * b$ or $K + b$. Now $K + b = \uparrow \{b \wedge \neg m\}$. As for $K * b$, since $\neg b \notin K$, we see from the Inclusion and (especially the) Preservation postulates that $K * b = K + b$ anyways. In particular, $\neg m \in K * b$. This seems like an unintuitive result: given that we want to model a change in the world resulting from putting the book on the floor, why should we believe afterwards that the magazine is not on the floor? The **KM** theory addresses this by proposing **AGM**-like postulates on the matter of *update*, the phenomenon illustrated in this example.

Belief change with dynamic non-epistemic logic The three 'theory change operators' $\oplus$, $\ominus$, and $\circledast$ can be reinterpreted as dynamic modal operators. A straightforward way to model these operators would be a logic in which $[\circledast\varphi]\psi$ expresses that after revision with φ, (the agent believes) ψ. This approach was suggested by van Benthem in [94][9] and further developed by de Rijke in [26]. They propose a semantical counterpart of a total order on theories, in the form of 'updating' and 'downdating' relations between states or worlds, standing

[9] It is only one of many topics covered in that publication, namely Section 6, pages 714–715, 'Cognitive procedures over information patterns'. Note this work is similar to a 1991 technical report.

for theories, and interpret the modal operator as a transition in such a structure according to these relations. 'Updating' models expansion: it relates the current state to states that result from expansion. 'Downdating' models contraction. It relates states that result from contraction to the current state. Revision is indeed downdating followed by updating. In this overview we focus on approaches that extend epistemic logics, therefore we do not give more details on this non-epistemic approach.

Dynamic Doxastic Logic (DDL) In the approach by Segerberg and collaborators [60, 83, 82, 61], beliefs are represented explicitly. We now identify a theory K with the believed formulas (or some subset of the believed formulas) in an epistemic state:

$$K = \{\psi \mid M, s \models B\psi\}.$$

As in [26] , **DDL** expresses belief change with dynamic modal operators $[\oplus\varphi]$, $[\ominus\varphi]$, and $[\circledast\varphi]$. In a typical revision where we have that $\neg\varphi \in K$, $\varphi \in K \circledast \varphi$, and $\neg\varphi \notin K \circledast \varphi$, we now get

- $M, s \models B\neg\varphi$
- $M, s \models [\circledast\varphi]B\varphi$
- $M, s \models [\circledast\varphi]\neg B\neg\varphi$

For contraction, we want that in case $M, s \models B\varphi$, after contraction φ is no longer believed, i.e., $M, s \models [\ominus\varphi]\neg B\varphi$. Similarly, for expansion we aim to achieve $M, s \models [\oplus\varphi]B\varphi$.

This approach is known as *dynamic doxastic logic* or **DDL**. Similar to [26] it presumes a transition relation between states representing theories, but this is now differently realized, namely using what is known as a Segerberg-style semantics wherein factual and epistemic information— called the *world component* and *doxastic component* —are strictly separated. A dynamic operator is interpreted as a transition along the lines of minimal theory change set out by this given structure, with the additional restriction that the transitions describe epistemic (doxastic) change only, and not factual change. This restriction is enforced by not allowing the 'world component' to change in the transition relation but only the doxastic component [60, p.18].

There are now two options: either we restrict ourselves to beliefs in *objective* (boolean, non-epistemic) formulas, and we get what is known as basic **DDL**, as in [60, 83]. Or we allow higher-order beliefs, as in the dynamic epistemics described in previous sections of our chapter. We thus get 'full' or 'unlimited' **DDL**, also discussed in [60] but mainly in [61].

Incidentally, the semantic models of **DDL** are rather different from those in this chapter, at least on the surface. They are more similar to neighborhood models, or topological models for modal logics. These are too different from what we have seen to allow us to present them in this chapter. Getting back to **DDL** and the systems we have presented, we know of no publications offering detailed comparisons; surely this is because the work of this section is so new. For this, and for discussion of recent work on **DDL**, see Leitgeb and Segerberg [55].

Problems of the classical theory Classical belief revision, and its dynamic versions presented in the previous section, encounter a number of problems: the difficulties in extending them to *iterated belief revision*; the *multiplicity of belief revision policies*; difficulties in dealing with *multi-agent beliefs* and even more with *higher-order beliefs*, that is beliefs about other beliefs. We refer for details to Chapter 4c on the subject in this handbook. We are mainly interested in higher-order beliefs and iteration. Our discussion amounts to a suggestion that the work of this chapter can be useful in work on these matters.

If one drops the restriction to belief in objective formulas and allows higher-order beliefs, then the standard **AGM** postulates lead to paradoxes. In particular, *the Success postulate for revision is problematic for sentences that involve doxastic modalities*: we have already noted in Section 5.2 that Moore sentences $p \wedge \neg K_a p$ are not successful. Similar examples apply to formalisms which have the syntactic means to specify semantic properties of evident interest. We now continue with the in-depth treatment of recent dynamic epistemic approaches to belief revision. The hallmark of the approach is that the transition that interprets the dynamic operators is *constructed* (as a state transformer) from the (specific action of announcing the) revision formula, instead of *assuming* as given such a transition relation.

7.2 The dynamic logic of belief change

This section is mainly based on the work of J. van Benthem [98, 100] on the dynamic logic of belief change and "preference upgrade" (with some additional input from Baltag and Smets [15]). Essentially, this work uses the **DEL** paradigm to develop a logic for belief change that completely solves the problems posed to belief revision by multi-agent beliefs and higher-order beliefs, iterated revision, as well as partially addressing the problem of the multiplicity of belief revision policies.

Static versus dynamic belief revision The first fundamental distinction underlying this work is the one between *static* and *dynamic* belief revision (in the terminology of [15, 19]): the first has to do with *conditional beliefs*, while the second has to do with the *beliefs acquired after a belief-changing action.* The distinction is only significant when dealing with higher-order beliefs: in the case of factual beliefs, the two types of revision coincide. Static belief revision captures the agent's changing beliefs about an unchanging world. But, if we take the "world" as incorporating all the agents' higher-order beliefs, then *the world is in fact always changed by our changes of belief* (as shown above, using examples involving Moore sentences). As a consequence, the best way to understand static belief revision with a proposition P is as expressing *the agent's revised beliefs, after learning* P, about what *was the case, before the learning.* In contrast, dynamic belief revision captures the agent's revised beliefs about the world *as it is after revision.*

Static revision as conditional belief Classical **AGM** theory deals with changing beliefs about an unchanging world. In our terminology, it is static belief revision. In a modal logic setting, it is natural to formalize static revision as hypothetical belief change, using *conditional belief* operators[10] $B_a^\alpha \varphi$, as in Section 4.7: if K is agent a's current belief set at state s and $*_a$ is her belief revision operator, then writing $s \models B_a^\varphi \psi$ is just a way of saying that $\psi \in K *_a \varphi$. Based on the above discussion, we can thus read a doxastic conditional $B_a^\varphi \psi$ as follows: *if learning* φ*, agent* a *would come to believe that* ψ *was the case (before the*

[10] It may seem that the failure of Ramsey's test for **AGM** revision would conflict with a conditional belief interpretation of **AGM**. But this is not the case. In the conditional belief setting, Gardenfors' impossibility result simply shows that "a conditional belief" is not the same as "a belief in a conditional"; more precisely, there doesn't exists any non-epistemic, non-doxastic notion of conditional that would validate this equivalence. For more on this, cf. Leitgeb [54].

learning). The semantics is given by plausibility models (or systems of Grove spheres, see Section 2.4), as in Section 4.7, with a conditional belief $B_a^\varphi \psi$ defined via the the most plausible states (satisfying φ) and being epistemically indistinguishable from the current state). If we translate the **AGM** postulates into the language of conditional beliefs, while taking into account the concept of "(fully introspective) knowledge" and the limitations that it poses to belief revision, we obtain the axioms of conditional doxastic logic **CDL** from Section 4.7.[11] In particular, observe that the **AGM** Success postulate *is* valid for static belief revision, even if we allow doxastic modalities (and thus higher-order beliefs) in our language: it is always true (even for Moore sentences φ) that, after learning that φ is the case, agents come to believe that φ *was* the case (before the learning).

In contrast, a statement $[!\varphi]B_a\psi$ involving a dynamic modality says that *after learning φ, agent a would come to believe that ψ is the case (in the world after the learning).* Due to Moore-type sentences, dynamic belief revision will *not* satisfy the **AGM** postulates (and in particular, the Success postulate will fail).

Triggers of information change As van Benthem [98] observes, in order to understand and formalize dynamic belief revision, it is essential to take into account the actual "learning event" that "triggered" the belief change. For example, our beliefs about *the current situation after* hearing a *public* announcement are different from our beliefs after receiving a *fully private* announcement. In the public case, we may come to believe that the content of the announcement is now *common knowledge* (or at least common belief); in the private case, we may come to believe the opposite: that the content of the announcement forms now our *secret knowledge*. In contrast, our beliefs about the triggering action are irrelevant as far our static revision is concerned: our conditional beliefs about the current situation given some hypothetical information do not depend on the way this information might be acquired. This explains a fact observed in [98], namely that by and large, the standard literature on belief revision (or belief update) does not mention in any way the explicit triggers (the actual doxastic events) that cause the belief changes (dealing instead only with types of abstract operations on beliefs, such as update, revision and contraction etc). The reason for this lies in the static character of **AGM** revision, as well as its restriction (shared with the **KM** updates and basic **DDL**) to one-agent, first-level, factual beliefs.

This is where the **DEL** paradigm can help: as already seen in this chapter, **DEL** explicitly analyzes the triggers for information change, from simple announcements of facts to individual agents to complex information-carrying events, involving many agents and their different perspectives on the learning event.

Revision as relation change If we model static beliefs (including conditional beliefs) using plausibility relations, then dynamic belief revision corresponds to *relation change*: the model is modified by changing the plausibility arrows. Different types of dynamic belief revision (induced by different triggering events) will correspond to different such changes. In other words, we can *model the triggers of belief revision as relation transformers*, similarly to the way we previously modeled knowledge updates as epistemic state transformers.

Hard versus soft information The second fundamental distinction made in [98] is between learning *'hard facts'* and acquiring *'soft information'*. Unlike in the classical belief-

[11] The translation is carried out in detail in [15].

revision setting, in an epistemic-logic setting we need to distinguish between the announcements that lead to *knowledge* (in the absolute, un-revisable sense) of some hard fact and the ones that only affect our *beliefs*. The first type is exemplified by the "truthful public announcement" actions $!\ P$, that we have already seen. The second type will correspond to "soft" informational actions, of the kind that is more standard in belief revision. One can also have more complex *mixtures of hard and soft information*, giving rise to more complex belief-revision policies.

Defining, classifying, and axiomatizing belief-revision policies We mentioned **PDL** in Section 4.9, and one reason for doing so is the work here. The natural language to define *relation changes* is the set of programs of **PDL**: one can then redefine the plausibility relations R_a (corresponding to the "at least as plausible as" relations $\leq_a$ in Section 4.7), via a clause of the form

$$R_a := \pi(R_a),$$

where $\pi(R_a)$ is any **PDL** program built using tests $?\varphi$, the universal relation $\top$, the old plausibility relations R_a and regular operations on relations: union $\cup$, composition ; and iteration *. In other words, one can use a *relation transformer* $\pi(r)$ as in Section 4.9, and redefine the plausibility relations via $R_a := [\![\pi(r)]\!](R_a)$. In their analysis of revision policies, van Benthem and Liu [100] propose as a natural class of relation transformers the ones that are definable in **PDL** *without iteration*, showing that these are particularly well-behaved. In particular, one can read off the relational definition a set of *reduction laws* for each such relation transformer, automatically providing a *complete axiomatization* of the corresponding dynamic logic. It is important to note that the reduction laws that are immediately obtainable through this method are for the *safe belief*[12] and *knowledge* modalities, not for conditional belief. But one *can* derive reduction laws for conditional belief in many instances, using the observation made in Section 4.8 concerning the definability of conditional belief in terms of knowledge and safe belief.

Examples of definable revision policies and their reduction laws: We give here only three examples of multi-agent belief-revision policies: truthful public announcements of hard facts, lexicographic update and conservative upgrade. They were all introduced by van Benthem in [98] as dynamic multi-agent versions of revision operators previously considered by Rott [79] and other authors. In each case, we give here only one example of a reduction law, namely the analogue for belief of the **DEL** Action-Knowledge Axiom which we mentioned in Section 5.5.

1. Belief change under hard information Truthful public announcements $!\varphi$ of hard facts can be considered as a (limit-)case of belief revision policy. Instead of defining it by world elimination as before, we can equivalently define it using the relation transformer

$$\pi(r) \quad = \quad ?\varphi; r; ?\varphi.$$

So the new accessibility relations are $R_a := [\![\pi(r)]\!](R_a)$, where R_a are the old relations. (See Example 9 in Section 5.1.) The corresponding Reduction Axiom for belief is

$$[!\varphi]B_a\theta \leftrightarrow (\varphi \rightarrow B_a^{\varphi}[!\varphi]\theta)\,,$$

[12] This is called the "preference modality" by van Benthem and Liu [100].

which generalizes the Announcement-Knowledge Axiom from Section 5.1 to the case of beliefs. There also exists a more general reduction law for *conditional* belief.

2. Public Announcements of Soft Facts: The "Lexicographic Upgrade" To allow for soft belief revision, an operation $\Uparrow \varphi$ was introduced in [98], essentially adapting to public announcements the 'lexicographic' policy for belief revision described in [79]. This operation, called *lexicographic update* consists of changing the current plausibility order on any given state model as follows: all φ-worlds become more plausible than all $\neg\varphi$-worlds, and within the two zones, the old ordering remains. Following what we did at the end of Section 4.9 and in Example 9 in Section 5.1, we are using the **PDL** relation transformer

$$\pi(r) \quad = \quad (?\varphi; \top; ?\neg\varphi) \cup (?\neg\varphi; r; ?\neg\varphi) \cup (?\varphi; r; ?\varphi)$$

where $\top$ is the universal relation. As before, the accessibility relations R_a change to $[\![\pi(r)]\!](R_a)$. The important new step in verifying that this does what we want is

$$\begin{array}{rcl} [\![?\varphi; \top; ?\neg\varphi]\!] & = & \{(w,w) : w \in [\![\varphi]\!]\}; \{(u,v) : u, v \in W\}; \{(w,w) : w \in [\![\neg\varphi]\!]\} \\ & = & \{(u,v) : u \in [\![\varphi]\!] \text{ and } v \in [\![\neg\varphi]\!]\}. \end{array}$$

The corresponding Reduction Axiom for belief is

$$[\Uparrow \varphi]B_a\theta \leftrightarrow (\hat{K}_a\varphi \wedge B_a^{\varphi}[\Uparrow \varphi]\theta) \vee (\neg\hat{K}_a\varphi \wedge B_a[\Uparrow \varphi]\theta)$$

where again $\hat{K}$ is the "epistemic possibility" operator (the $\Diamond$-like dual of the K operator). As in the case of hard announcements, there also exists a more general reduction law for conditional belief.

3. Public Announcements of Soft Facts: The Conservative Upgrade. The operation $\uparrow \varphi$ of *conservative upgrade*, as defined in [98], changes any model as follows: *the best φ-worlds come on top* (i.e., the most plausible φ-states become the most plausible overall), *and apart from that, the old order remains.* The reduction law for belief is the same as in the previous case. The difference can be only be seen by looking at the reduction law for *conditional* belief. See [98] for details.

7.3 Logics for doxastic actions: the action-priority update

The work of Aucher [4, 5], Baltag and Smets [16, 17, 19], and van Ditmarsch and Labuschagne [109, 107, 110] can be considered as an attempt to extend to dynamic belief revision the unified **DEL** setting based on *action models.* We focus on the approach by Baltag and Smets and specifically on their 'action-priority update'. This gives currently the most convincing picture given the relational approach. It also goes some way towards addressing the problem of the multiplicity of belief revision policies: as we will see below, the action-priority update unifies and subsumes many different policies and types of revision, which come to be seen as the result of applying the same update operation to different triggers given by specific learning events. Indeed, in this interpretation, the triggering events for belief revision are *doxastic actions*, modeled using *action plausibility models*, in a similar way to the way epistemic actions were modeled using epistemic action models. The actions' "preconditions" encode the *information* carried by each action. The plausibility relations between actions are meant to represent *the agent's (conditional) beliefs about the learning event at the moment of its happening.*

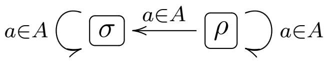
We assume here the setting of plausibility models and the conditional doxastic logic **CDL** from Section 4.7. The following definition gives the natural plausibility analogue of the action models from Section 5.4, by incorporating the main intuition underlying the **AGM** belief revision: that *new information has priority over old beliefs* .

Definition [Action plausibility model] (Aucher) Let $\mathcal{L}$ be a logical language which extends the language of **CDL**. An *action plausibility model over* $\mathcal{L}$ is a structure $\mathsf{U} = \langle \mathit{Actions}, \leq, \mathsf{pre} \rangle$ such that $\langle \mathsf{S}, \leq \rangle$ is a plausibility frame and $\mathsf{pre} : \mathsf{S} \to \mathcal{L}$ is a precondition function that assigns a *precondition* $\mathsf{pre}(\sigma) \in \mathcal{L}$ to each $\sigma \in \mathsf{S}$. As in Section 5.4, the elements of S are called *action points*, and for each $a \in A$, $\leq_a$ is a plausibility relation on S. For $\sigma, \rho \in \mathsf{S}$, we read $\sigma \leq_a \rho$ as follows: agent a considers action σ as being at least as plausible as action ρ. A *doxastic action* is a pointed action plausibility model (U, σ), with $\sigma \in \mathsf{S}$.

Example 16 The *truthful public announcement of a hard fact* φ is modeled by a singleton action model consisting of an action point σ, with identity as the plausibility relation for every agent, and with precondition $\mathsf{pre}(\sigma) = \varphi$. As in Section 5.4, we call this action model *Pub* φ, and denote by $!\ \varphi$ the action corresponding to the point σ.[13]

Fully private announcements and *fair-game announcements* can be similarly modeled, essentially by reading the arrows in their epistemic action models from Section 5.4 as plausibility arrows.

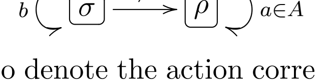
Announcements of Soft Information. We can simulate public announcements of soft facts, as described in the previous section, using action plausibility models. For instance, the lexicographic update $\Uparrow \varphi$ has the following action model:



with $\mathsf{pre}(\sigma) = \varphi$ and $\mathsf{pre}(\rho) = \neg\varphi$. The action point on the left corresponds to the case that the announcement happens *true*; the action point on the right corresponds to the case that the announcement is *false*.

The conservative upgrade $\uparrow \varphi$ can be similarly encoded, using a more complicated action model.

Successful Lying. The action of an agent b's "lying in a publically successful manner" by can be modeled as follows: given a sentence φ, the model consists of two action points σ and ρ, the first being the action in which agent b publicly lies that (he knows that) φ (while in fact he doesn't know it), and the second being the action in which b makes a truthful public announcement that (he knows that) φ. The preconditions are $\mathsf{pre}(\sigma) = \neg K_b\varphi$ and $\mathsf{pre}(\rho) = K_b\varphi$. Agent b's plausibility relation is simply the identity: she knows whether she's lying or not. The relation for any hearer $c \neq b$ should make it more plausible to him that b is telling the truth rather than lying: $\sigma <_c \rho$. This reflects the fact that we are modeling "typically successful lying": by default, in such an action, the hearer trusts the speaker, so he is inclined to believe the lie.



We call this model $\mathit{Lie}_b\ \varphi$, and also denote the action corresponding to the point σ by $\mathit{Lie}_b\ \varphi$, and the action corresponding to the point ρ by $\mathit{True}_b\ \varphi$.

[13] Technically, we should distinguish between this plausibility model and the corresponding action model in Section 5.4, but we choose to use the same notation, relying on the context for deciding when to interpret it as a plausibility model. The same applies to all the other plausibility action models in this section having the same notation as an epistemic model.

Definition [Execution, Action-Priority Update] Given a doxastic state (M, s) with $M = (S, \leq, V)$, and a doxastic action (U, σ) with $\mathsf{U} = (\mathsf{S}, \leq, \mathsf{pre})$, the result of executing (U, σ) in (M, s) is only defined when $M, s \models \mathsf{pre}(\sigma)$. In this case, it is the doxastic state $(M \otimes_{\leq} \mathsf{U}), (s, \sigma))$ where $M \otimes_{\leq} \mathsf{U} = \langle S', \leq', V' \rangle$ is a *restricted anti-lexicographic product* of the structures M and U, defined by

$$\begin{array}{lcl} S' & \equiv & \{(s, \sigma) \mid s \in S, \sigma \in \mathsf{S}, \text{ and } M, s \models \mathsf{pre}(\sigma)\} \\ (s, \sigma) \leq'_a (t, \rho) & \text{iff} & \sigma <_a \rho \text{ and } s \overset{a}{\sim} t\,; \text{ or else } \sigma \cong_a \rho \text{ and } s \leq_a t \\ (s, \sigma) \in V'_p & \text{iff} & s \in V_p \end{array}$$

where $\overset{a}{\sim}$ is the *epistemic indifference (uncertainty)* relation[14] on states, and $\cong_a$ is the *equi-plausibility relation* on actions (defined by $\sigma \cong_a \rho$ iff both $\sigma \leq_a \rho$ and $\rho \leq_a \sigma$).

This is a generalization of one of the belief-revision policies encountered in the literature (essentially incorporating the so-called "*maximal-Spohn revision*" into plausibility action models), as well as being a natural plausibility analogue of the product update from Section 5.4. The new order is simply the *anti-lexicographic order* on (epistemically indistinguishable) pairs. The name comes from [16, 17, 19]. Van Benthem calls it "Action Priority Update": indeed, this construction gives priority to the *action* plausibility relation. This is not an arbitrary choice, but is motivated by a specific interpretation of action models as encoding *belief changes*. In other words, the (strict) order on actions encodes *changes of order* on states. The definition of execution is a consequence of this interpretation: it just says that a strong plausibility order $\sigma <_a \rho$ on actions corresponds indeed to a change of ordering, (from whatever the ordering was) between the original (indistinguishable) input-states $s \overset{a}{\sim} t$, to the order $(s, \sigma) <_a (t, \rho)$ between output-states; while equally plausible actions $\sigma \cong_a \rho$ will leave the initial ordering unchanged: $(s, \sigma) \leq_a (t, \rho)$ iff $s \leq_a t$. Giving priority to action plausibility does not in any way mean that the agent's belief in actions is stronger than her belief in states; it just captures the fact that, at the time of updating with a given action, *the belief about the action is what is actual, it is the current belief about what is going on, while the beliefs about the input-states are in the past.*[15]

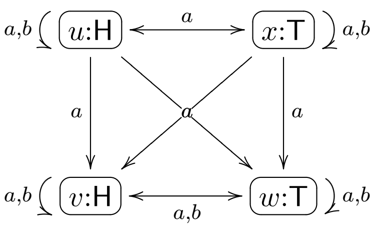
In a nutshell: *the doxastic action is the one that changes the initial doxastic state, and not vice-versa.* The belief update induced by a given action is nothing but an update with the (presently) believed action. If the believed action α requires the agent to revise some past beliefs, then so be it: this is the whole point of believing α, namely to use it to revise or update one's past beliefs. For example, in a successful lying, the action plausibility relation makes the hearer believe that the speaker is telling the truth; so she'll accept this message (unless contradicted by her knowledge), and change her past beliefs appropriately: this is what makes the lying successful.

Example 17 Consider the situation in Section 2.3, in which Bao was told the face of the coin, without Amina suspecting this. Assume moreover the coin lies heads up. This was

[14] Recall that, in a plausibility model, the epistemic uncertainty relation is defined by: $s \overset{a}{\sim} t$ iff either $s \leq_a t$ or $t \leq_a s$.

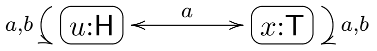
[15] Of course, *at a later moment*, the above-mentioned belief about action (*now* belonging to the past) might be itself revised. But this is another, *future update.*

represented in Section 2.4 by the plausibility model (10):



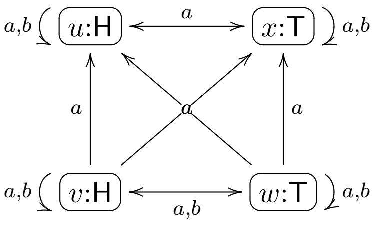
where the real state is the upper-left one.

Next, suppose Bao tells Amina: "I know the face of the coin". Let us first assume this is an evidently truthful statement, coming with a warranty of veracity of some sort or other. Then Amina takes Bao's statement as an announcement of a hard fact. So this action is represented by *Pub* $(K_b\mathsf{H} \vee K_b\mathsf{T})$, with the one-point action model described above (for truthful public announcements of hard information); the action point will be called σ. Execute now this doxastic action on the doxastic state given by (upper-left point in) the model (10) above. We identify the old states u and x with the pairs (u, σ) and (x, σ), respectively, and then we picture the result as



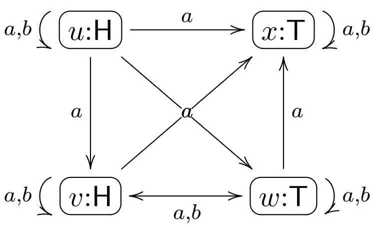
which fits our intuition about the agent's beliefs: it is now common knowledge that Bao knows the face of the coin.

What if Bao's announcement was not evidently truthful? Amina may still *believe* it, but she *doesn't know* that it is true. We model using an announcement of a soft fact rather than a hard one, corresponding to the lexicographic upgrade $\Uparrow (K_b\mathsf{H} \vee K_b\mathsf{T})$. Using the two-point action plausibility model for lexicographic update which we saw above, and computing the execution of this doxastic action on the original doxastic state given by model (10) above, we obtain:



Here and just below, we are identifying the old states with certain pairs to simplify the representation.

What if, instead of making a truthful announcement, Bao chooses to *lie*? For instance (in the initial situation, after he was secretely told that the coin lies heads up), suppose he tells Amina: "I know the coin is lying tails up". This is a lie. Let us assume it is a successful one, so that Amina trusts Bao completely and therefore believes he is telling the truth. We can represent this action using, as described above, the two-point action model $\mathit{Lie}_b\ (K_bH)$ for successful lying. If we now we execute this doxastic action on the original doxastic state from the model (10) above, we obtain



in which the upper left-hand state is the real one. Again, this fits our doxastic intuitions: Amina is deceived and believes the upper right-hand state to be the real one. However, this false belief is *revisable*: a new public announcement *Pub* K_bH (in effect, saying that Bao has lied and that in fact he knows the coin lies heads up) would correct Amina's wrong belief, making her know that the real state is the left-hand one.

Action-Priority Update Generalizes Product Update Recall the definition of the epistemic indistinguishability relation $\overset{a}{\sim}$ in a plausibility model: $s \overset{a}{\sim} t$ iff either $s \leq_a t$ or $t \leq_a s$. It follows that Action Priority Update implies the Product Update rule from Section 5.4:

$$(s,\sigma) \overset{a}{\sim} (t,\rho) \text{ iff } s \overset{a}{\sim} t \text{ and } \sigma \overset{a}{\sim} \rho.$$

The logic of doxastic actions As in Section 5.5, we can consider a signature-based language, where a *doxastic signature* is a *finite (fixed) plausibility frame* $\mathbf{\Sigma}$, together with an ordered list without repetitions $(\sigma_1, \ldots, \sigma_n)$ of some of the elements of $\mathbf{\Sigma}$. As in Section 5.5, each signature induces a syntactic action model, and it gives rise to a dynamic-doxastic logic $L(\mathbf{\Sigma})$. The language is obtained by augmenting either the language of conditional doxastic logic **CDL** from Section 4.7, or the language of the logic of knowledge and safe belief from Section 4.8, with dynamic modalities for (signature-based) doxastic actions. The semantics can be given in a similar way to the one in Section 5.5. We skip here the details, referring to [16, 17, 19]. Just as in **DEL**, and similarly to the approach in the previous subsection, one can automatically read off a set of *Reduction Axioms for knowledge and safe belief*, thus obtaining a complete proof system. But Baltag and Smets also derive in [17] general (though very complex) Reduction laws for conditional belief.

The Action-Safe-Belief Axiom As for **DEL**, we only present here the most important reduction axiom, namely the appropriate generalization of the Action-Knowledge Axiom to the logic of doxastic actions. In fact, there are two such laws: one for knowledge, the other for safe belief. But the reduction law for knowledge is essentially the same as the Action-Knowledge Axiom in Section 5.5. So we only state here an "Action-Safe-Belief Axiom", saying that, for every basic action α, we have:

$$[\alpha]\Box_a\varphi \leftrightarrow \left(\mathsf{pre}(\alpha) \to \bigwedge_{\alpha' <_a \alpha} K_a[\alpha']\varphi \wedge \bigwedge_{\alpha'' \cong_a \alpha} \Box_a[\alpha'']\varphi\right)$$

where $<_a$ is the strict plausibility order on (syntactic) actions (in the action model induced by the signature) and similarly $\cong_a$ is the *equi-plausibility relation* on (syntactic) actions.

This axiom could be thought of as the "fundamental law of dynamic belief revision": it allows us to compute or predict safe beliefs after a learning event in terms of knowledge and safe beliefs before the event. In plain words, it says that: a sentence φ will be safely believed after a doxastic event iff, whenever the action can take place, it is known that φ will become true after all more plausible events and at the same time it is safely believed that φ will become true after all equi-plausible events.

Unifying Diverse Belief-Revision Policies As seen in the examples above, the Action-Priority Update can simulate the various belief revision policies considered in the previous section. More generally, the power of the action model approach is reflected in the fact that many different revision policies can be recovered, in a uniform manner, as instances of the same type of update operation. In this sense, the **DEL** approach can be seen as a change of perspective: the multiplicity of possible revision policies considered in the Belief Revision literature is replaced by the multiplicity of possible action models; the differences are now viewed as *differences in input, rather than having different programs.* For a computer scientist, this resembles *currying* in the lambda-calculus: if every "operation" is encoded as an input-term, then *one operation* (functional application) *can simulate all operations.*[16] In a sense, this is nothing but the idea of Turing's universal machine, from the theory of computation. Note that, by incorporating the Product Update from Section 5.4, the Action-Priority Update gains all its dynamic features and its advantages: in addition to simulating a range of individual belief-revision policies, it can deal with an even wider range of complex types of multi-agent learning and communication actions. It may thus be realistic to expect that, *within its own natural limits*, Action Priority Update could play the role of a "universal qualitative machine" for dynamic interactive belief-revision. The problem of finding these natural limits remains open.

Notes The action plausibility models were first introduced by Aucher [4, 5], as an adaptation of the **DEL** framework of Baltag, Solecki and Moss to the case of dynamic belief revision. Aucher used an equivalent definition, inspired from the work of Spohn [85], describing plausibility models in terms of ordinal plausibility functions, interpreted as "*degrees of belief*". This lead Aucher, and then van Ditmarsch and Labuschagne [109, 107, 110], to propose and study various types of product update, of a different, more "quantitative" flavor than the Action-Priority update presented above; these proposals are based on using various binary operations on ordinals to compute the degree of belief of an updated state in terms of the corresponding degrees of belief of the input-state and of the action. None of these specific proposals seem to correspond to the Action-Priority update (although it is easy to see that this type of update *can* be computed via a special ordinal function, so in a sense it is *subsumed* by the general "quantitative" approach). Aucher introduced a doxastic logic, with operators $B_a^n\varphi$ for each ordinal degree of belief n, and completely axiomatized the dynamic logic corresponding to his proposal of product update. This work was generalized by van Ditmarsch [107], who also gave a good presentation of the various proposals in the literature, as well as of the various problems encountered. A recent breakthrough in the field was the work of van Benthem [98] on the relational approach to belief "upgrades", partially based

[16] Note that, as in untyped lambda-calculus, the input-term encoding the operation (i.e, the action model) and the static input-term to be operated upon (the state mode) are essentially of the same type: epistemic plausibility models for the same language (and for the same set of agents).

on previous work by van Benthem and Liu [100] on preference upgrades. At the same time, Baltag and Smets [15, 16, 17, 19] developed their own relational approach to dynamic belief revision, introducing the Action-Priority Update and the Action-Safe-Belief Axiom. Both van Benthem, and Baltag and Smets, used a qualitative logical language (either based on conditional belief operators, or on knowledge and safe belief operators) rather than one based on degrees of belief. Baltag and Sadrzadeh [14] gave an *algebraic axiomatization* of a type of dynamic belief revision. In more recent work (still to appear), Baltag and Smets [18] develop a *probabilistic version of dynamic belief revision*, based on combining their previous work on safe belief and the Action-Priority update with the work of van Fraasen [31], Boutilier [22] and Parikh [2] on using Popper's counterfactual probability functions to deal with belief revision.

8 Conclusion

As we end this chapter, we step back to try to understand what makes this particular subject of epistemic logic and information update what it is. We especially want to compare what is going on here to what is discussed in other chapters, especially Chapters 4c and 3b.

In a sense, our treatment of epistemic phenomena is *ultra-semantic*. Beginning in Section 2, we depicted representations and treated them as abstract semantic objects. Even before this, we stated openly that our modeling was slanted towards justifiable belief. This stance implicitly allowed us to ignore *reasons to believe* and instead focus on models of the phenomena of interest. All throughout our section on examples, we emphasized that one must test models and semantic definitions against intuitions, that the proof of the pudding is in the eating. Indeed, our subject is not a single pudding at all but rather a whole buffet of delectable semantic desserts. We also made it clear that the chefs used an artificial sweetener, relational models, and so those allergic to logical omniscience might prefer the fresh fruit. But except for this, the models work extremely well: the predictions of the logical languages match the intuitions. And one can use the formal tools as a real aid in building representations.

At the same time, our work is *unexpectedly syntactic*. We saw a series of logical languages crafted to exploit the key semantic features of the models. Whenever one hears about "encoding" in this subject, it is this: the semantic objects quickly become the sites for semantic evaluation in languages which are richer than one might have at first expected. The easiest example is the relational (Kripke) semantics itself. Having a set of Carnapian state descriptions living alone is at this time fairly mundane. Even adding one or more accessibility relation and calling things "possible worlds" does not go far in relating the worlds to one another. But once one has languages with modal operators, statements evaluated at one world in general must refer to other worlds. Thus the worlds *really* are related: since the logical language has iterated modal sentences, what is true here is in general influenced by what is true far away.

The models in this chapter also incorporate dynamics, social features such as common knowledge, and conditional operators. In each case, the languages are taken to be immediately higher-order: we have knowledge about knowledge, belief about beliefs, announcements about announcements, etc. What makes the subject work is that the formal semantics of the languages refer to the structure of the models, and at the same time the intuitive concepts of interest correspond most closely to statements in the formal languages. One aspect of our work which might be unexpected is the emphasis on particular logical systems for specialized phenomena. We presented a logic of public announcements in Section 5, but this is just the tip of the iceberg. One can formulate specialized logics for other epistemic actions. The point again is that we have semantic objects corresponding to these actions (this seems to be an innovation coming from this subject) and then the resulting logical systems take on an interest of their own, qua syntactic systems. And on the opposite pole from the specialized logics are the very general ones which incorporate arbitrary actions in some sense: these logical languages are unexpectedly syntactic in the sense that their very formulation is trickier than usual, as is their semantics. But the arrows inside of relational models are the same kind of thing as the arrows between the models, and this is *why* dynamic epistemic logic works.

One should compare the situation with the belief revision literature surveyed in Chapter 4c. The AGM postulates deal with several operations, most notably revision. These came first, and then later people were concerned with concrete models of them, with representations of theory-change operations, and the like. There is much less of an emphasis on matching the predictions of models to intuitions, mainly because the intuitions are often not as clear, and also because notions like a *theory change operation* are more abstract than a *completely private announcement* in our sense. It also took longer for the matter of iterated revision to become central. So the subject developed in a different way from ours. At the same time, there are interesting similarities: as Table 1 in Chapter 4c shows, the history of work in belief revision might be organized according to the particular kinds of prior and posterior belief states discussed. In our subject, the parallel is the extension of the ideas from "hard" semantic updates to "soft" ones (in the terminology of Chapter 3b). Belief revision theory is a much more active field than dynamic epistemic logic, and so one would expect to see a further push towards varied semantic models. But overall, these parallels could be taken to indicate hidden traces of functionalism in what we are doing, though clearly the emphasis on models and languages here is the most prominent difference.

All of this could be said about other closely related topics, especially work on history-based epistemic systems, interpreted systems, and related models which we surveyed in our temporal reasoning Section 6.

Another difference between the main thrust of belief revision work and recent trends in epistemic logic is the "social" aspect of the latter area. This is not true of the earliest work in the subject, partly because philosophers have tended to look only at public information. But as one can see from our chapter, the subject is now about *public and private* types of information change: how they compare and contrast, and how they are integrated in larger theories. This is clearly of interest in mathematical areas of the social sciences, but we feel it is also of interest to philosophy. To be a person is to relate to others, and so to understand knowledge we should pay special attention to multi-agent phenomena.

What connects the ultra-semantic and unexpectedly syntactic are the results on the logical systems themselves. Details of representations often conceal significant conceptual decisions, and results on logical languages and systems can help in the evaluation of different representations. By formulating sound principles, one uncovers (or highlights) hidden assumptions. The matching completeness theorems indicate the right kind of "harmony" (see Chapter 3b). Even more indicative is the fact that those logical systems typically have axiomatic presentations that make intuitive sense. There is no mathematical reason why the axioms behind logical systems should in any way be "nice." Frequently they are not. But we would like to regard the happy coincidences of axioms and intuitions in our subject as signposts which indicate that we are on the right track and point the way ahead.

Acknowledgments

We thank Anthony Gillies and Joshua Sack for their very useful comments on this chapter at various stages.

References

[1] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] H. Arlo-Costa and R. Parikh. Conditional probability and defeasible inference. *Journal of Philosophical Logic*, 34:97–119, 2005.

[3] S. Artemov. Evidence-based common knowledge. Technical report, CUNY, 2004. Ph.D. Program in Computer Science Technical Report TR-2004018.

[4] G. Aucher. A combined system for update logic and belief revision. In M.W. Barley and N. Kasabov, editors, *Intelligent Agents and Multi-Agent Systems – 7th Pacific Rim International Workshop on Multi-Agents (PRIMA 2004)*, pages 1–17. Springer, 2005. LNAI 3371.

[5] G. Aucher. How our beliefs contribute to interpret actions. In M. Pechoucek, P. Petta, and L.Z. Varga, editors, *Proceedings of CEEMAS 2005: Multi-Agent Systems and Applications IV*, volume 3690 of *Lecture Notes in Computer Science*, pages 276–285. Springer, 2005.

[6] R.J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[7] P. Balbiani, A. Baltag, H.P. van Ditmarsch, A. Herzig, T. Hoshi, and T. De Lima. What can we achieve by arbitrary announcements? A dynamic take on Fitch's knowability. To appear in the proceedings of TARK XI, 2007.

[8] A. Baltag. A logic for suspicious players: epistemic actions and belief updates in games. *Bulletin Of Economic Research*, 54(1):1–46, 2002.

[9] A. Baltag, B. Coecke, and M. Sadrzadeh. Algebra and sequent calculus for epistemic actions. *Electronic Notes in Theoretical Computer Science*, 126:27–52, 2005.

[10] A. Baltag, B. Cooke, and M. Sadrzadeh. Epistemic actions as resources. *Journal of Logic and Computation*, 2007, to appear. Also in LiCS 2004 Proceedings of Logics for Resources, Programs, Processes (LRPP).

[11] A. Baltag and L.S. Moss. Logics for epistemic programs. *Synthese*, 139:165–224, 2004. Knowledge, Rationality & Action 1–60.

[12] A. Baltag, L.S. Moss, and S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pages 43–56, 1998.

[13] A. Baltag, L.S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. Technical report, Centrum voor Wiskunde en Informatica, Amsterdam, 1999. CWI Report SEN-R9922.

[14] A. Baltag and M. Sadrzadeh. The algebra of multi-agent dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, 157(4):37–56, 2006.

[15] A. Baltag and S. Smets. Conditional doxastic models: a qualitative approach to dynamic belief revision. In *Proceedings of WOLLIC'06*, Electronic Notes in Theoretical Computer Science, 2006.

[16] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. Proceedings of LOFT 2006 (7th Conference on Logic and the Foundations of Game and Decision Theory), 2006.

[17] A. Baltag and S. Smets. The logic of conditional doxastic actions: a theory of dynamic multi-agent belief revision. Proceedings of ESSLLI Workshop on Rationality and Knowledge, 2006.

[18] A. Baltag and S. Smets. From conditional probability to the logic of belief-revising actions. To appear in the proceedings of TARK IX, 2007.

[19] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. To appear in G. Bonanno, W. van der Hoek, M. Wooldridge (eds). Selected Papers from LOFT'06, *Texts In Logic and Games*, Amsterdam University Press, 2007.

[20] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001. Cambridge Tracts in Theoretical Computer Science 53.

[21] O. Board. Dynamic interactive epistemology. *Games and Economic Behaviour*, 49:49–80, 2004.

[22] C. Boutilier. On the revision of probabilistic belief states. *Notre Dame Journal of Formal Logic*, 36(1):158–183, 1995.

[23] B. Brogaard and J. Salerno. Fitch's paradox of knowability. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2004. `http://plato.stanford.edu/archives/sum2004/entries/fitch-paradox/`.

[24] C. Cîrstea and M. Sadrzadeh. Coalgebraic epistemic update without change of model. In T. Mossakowski, editor, *Proceedings of 2nd Conference on Algebra and Coalgebra in Computer Science*, needed. needed, 2007.

[25] M. D'Agostino, D.M. Gabbay, and A. Russo. Grafting modalities onto substructural implication systems. *Studia Logica*, VII:1–40, 1997.

[26] M. de Rijke. Meeting some neighbours. In J. van Eijck and A. Visser, editors, *Logic and information flow*, pages 170–195, Cambridge MA, 1994. MIT Press.

[27] C. Dixon, M. Fisher, and M. Wooldridge. Resolution for temporal logics of knowledge. *Journal of Logic and Computation*, 8(3):345–372, 1998.

[28] Ho Ngoc Duc. *Resource-Bounded Reasoning About Knowledge.* PhD thesis, University of Leipzig, 2001.

[29] P. Economou. Sharing beliefs about actions: A parallel composition operator for epistemic programs. In *Proceedings of the ESSLLI 2005 workshop Belief Revision and Dynamic Logic*, 2005. Available on `http://www.irit.fr/~Andreas.Herzig/Esslli05/`.

[30] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge.* MIT Press, Cambridge MA, 1995.

[31] B. C. Van Fraassen. Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic*, 24:349–377, 1995.

[32] H. Freudenthal. (formulation of the sum-and-product problem). *Nieuw Archief voor Wiskunde*, 3(17):152, 1969.

[33] M. F. Friedell. On the structure of shared awareness. *Behavioral Science*, 14(1):28–39, 1969.

[34] G. Gamow and M. Stern. *Puzzle-Math.* Macmillan, London, 1958.

[35] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* Bradford Books, MIT Press, Cambridge, MA, 1988.

[36] J.D. Gerbrandy. *Bisimulations on Planet Kripke.* PhD thesis, University of Amsterdam, 1999. ILLC Dissertation Series DS-1999-01.

[37] J.D. Gerbrandy. The surprise examination in dynamic epistemic logic. *Synthese*, 155(1):21–33, 2007.

[38] J.D. Gerbrandy and W. Groeneveld. Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–169, 1997.

[39] E. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[40] P. Gochet and P. Gribomont. Epistemic logic. In D. M. Gabbay and J. Woods, editors, *Handbook of the History of Logic*, volume 7, pages 99–195. Elsevier, 2006.

[41] W. Groeneveld. *Logical investigations into dynamic semantics.* PhD thesis, University of Amsterdam, 1995. ILLC Dissertation Series DS-1995-18.

[42] J.Y. Halpern, R. van der Meyden, and M.Y. Vardi. Complete axiomatizations for reasoning about knowledge and time. *SIAM Journal on Computing*, 33(3):674–703, 2004.

[43] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic.* MIT Press, Cambridge MA, 2000. Foundations of Computing Series.

[44] J. Heal. Common knowledge. *Philosophical Quarterly*, 28:116–131, 1978.

[45] J. Hintikka. *Knowledge and Belief.* Cornell University Press, Ithaca, NY, 1962.

[46] J. Hintikka. Reasoning about knowledge in philosophy. In J.Y. Halpern, editor, *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 63–80, San Francisco, 1986. Morgan Kaufmann Publishers.

[47] J. Jaspars. *Calculi for Constructive Communication.* PhD thesis, University of Tilburg, 1994. ILLC Dissertation Series DS-1994-4, ITK Dissertation Series 1994-1.

[48] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 387–394, 1991.

[49] B.P. Kooi. Expressivity and completeness for public update logics via reduction axioms. *Journal of Applied Non-Classical Logics*, 2007. To appear.

[50] S. Kripke. A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24:1–14, 1959.

[51] J. L. Kvanvig. Paradoxes, epistemic. In E. Craig, editor, *Routledge Encyclopedia of Philosophy*, volume 7, pages 211–214. Routledge, London, 1998.

[52] F. Landman. *Towards a Theory of Information.* PhD thesis, University of Amsterdam, 1986.

[53] K. Lehrer and T. Paxson. Knowledge: undefeated justified true belief. *The Journal of Philosophy*, 66:225–237, 1968.

[54] H. Leitgeb. Beliefs in conditionals vs. conditional beliefs. *Topoi*, 2007.

[55] H. Leitgeb and K. Segerberg. Dynamic doxastic logic: Why, how, and where to? *Synthese*, 155(2):167–190, 2007.

[56] W. Lenzen. Recent work in epistemic logic. *Acta Philosophica Fennica*, 30:1–219, 1978.

[57] W. Lenzen. Knowledge, belief, and subjective probability: outlines of a unified system of epistemic/doxastic logic. In V.F. Hendricks, K.F. Jorgensen, and S.A. Pedersen, editors, *Knowledge Contributors*, pages 17–31, Dordrecht, 2003. Kluwer Academic Publishers. Synthese Library Volume 322.

[58] H. J. Levesque. A logic of implicit and explicit beliefs. In *Proceedings of the National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas, 1984.

[59] D.K. Lewis. *Convention, a Philosophical Study.* Harvard University Press, Cambridge (MA), 1969.

[60] S. Lindström and W. Rabinowicz. Belief change for introspective agents, 1999. `http://www.lucs.lu.se/spinning/`.

[61] S. Lindström and W. Rabinowicz. DDL unlimited: dynamic doxastic logic for introspective agents. *Erkenntnis*, 50:353–385, 1999.

[62] A.R. Lomuscio. *Knowledge Sharing among Ideal Agents.* PhD thesis, University of Birmingham, Birmingham, UK, 1999.

[63] A.R. Lomuscio and M.D. Ryan. An algorithmic approach to knowledge evolution. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing (AIEDAM)*, 13(2), 1999. Special issue on Temporal Logic in Engineering.

[64] C. Lutz. Complexity and succinctness of public announcement logic. To appear in the proceedings of the Fifth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 06), 2006.

[65] J. McCarthy. Formalization of two puzzles involving knowledge. In Vladimir Lifschitz, editor, *Formalizing Common Sense: Papers by John McCarthy*, Ablex Series in Artificial Intelligence. Ablex Publishing Corporation, Norwood, N.J., 1990. original manuscript dated 1978–1981.

[66] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science.* Cambridge Tracts in Theoretical Computer Science 41. Cambridge University Press, Cambridge, 1995.

[67] J.S. Miller and L.S. Moss. The undecidability of iterated modal relativization. *Studia Logica*, 79(3):373–407, 2005.

[68] G.E. Moore. *Ethics.* Oxford University Press, 1912. Consulted edition: The Home University Library of Modern Knowledge, volume 54, OUP, 1947.

[69] G.E. Moore. A reply to my critics. In P.A. Schilpp, editor, *The Philosophy of G.E. Moore*, pages 535–677. Northwestern University, Evanston IL, 1942. The Library of Living Philosophers (volume 4).

[70] G.E. Moore. Russell's "theory of descriptions". In P.A. Schilpp, editor, *The Philosophy of Bertrand Russell*, pages 175–225. Northwestern University, Evanston IL, 1944. The Library of Living Philosophers (volume 5).

[71] R.C. Moore. Reasoning about knowledge and action. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence (IJCAI-77)*, Cambridge, Massachusetts, 1977.

[72] Y.O. Moses, D. Dolev, and J.Y. Halpern. Cheating husbands and other stories: a case study in knowledge, action, and communication. *Distributed Computing*, 1(3):167–176, 1986.

[73] L.S. Moss. From hypersets to Kripke models in logics of announcements. In J. Gerbrandy, M. Marx, M. de Rijke, and Y. Venema, editors, *JFAK. Essays Dedicated to Johan van Benthem on the Occasion of his 50th Birthday*, Vossiuspers. Amsterdam University Press, 1999.

[74] D.J. O'Connor. Pragmatic paradoxes. *Mind*, 57:358–359, 1948.

[75] E. Pacuit. Some comments on history based structures. *Journal of Applied Logic*, to appear.

[76] R. Parikh and R. Ramanujam. A knowledge-based semantics of messages. *Journal of Logic, Language, and Information*, 12:453–467, 2003.

[77] J.A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[78] L. Qian. Sentences true after being announced. In *Proceedings of 'Student Conference of the 2002 North American Summer School in Logic, Language, and Information'*. Stanford University, 2002.

[79] H. Rott. Adjusting priorities: Simple representations for 27 iterated theory change operators. Manuscript, 2004.

[80] J. Sack. *Adding Temporal Logic to Dynamic Epistemic Logic.* PhD thesis, Indiana University, Bloomington, 2007.

[81] K. Segerberg. Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, 39(3):287–306, 1998.

[82] K. Segerberg. Default logic as dynamic doxastic logic. *Erkenntnis*, 50:333–352, 1999.

[83] K. Segerberg. Two traditions in the logic of belief: bringing them together. In H.J. Ohlbach and U. Reyle, editors, *Logic, Language, and Reasoning*, pages 135–147, Dordrecht, 1999. Kluwer Academic Publishers.

[84] R.A. Sorensen. *Blindspots.* Clarendon Press, Oxford, 1988.

[85] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134, 1988.

[86] R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in Logical Theory*, APQ Monograph No2. Blackwell, 1968.

[87] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[88] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.

[89] D. Steiner. A system for consistency preserving belief change. In *Proceedings of the ESSLLI Workshop on Rationality and Knowledge*, pages 133–144, 2006.

[90] M. Steup. The analysis of knowledge. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Spring 2006. `http://plato.stanford.edu/archives/spr2006/entries/knowledge-analysis/`.

[91] M. Swain. Epistemic defeasibility. *The American Philosophical Quarterly*, 11:15–25, 1974.

[92] N. Tennant. Victor vanguished. *Analysis*, 62:135–142, 2002.

[93] J.F.A.K. van Benthem. Semantic parallels in natural language and computation. In *Logic Colloquium '87*, Amsterdam, 1989. North-Holland.

[94] J.F.A.K. van Benthem. Logic and the flow of information. In *Proceedings of the 9th International Congress of Logic, Methodology and Philosophy of Science (1991)*. Elsevier Science B.V., 1994. Also available as: Report LP-91-10, ILLC, University of Amsterdam.

[95] J.F.A.K. van Benthem. *Exploring logical dynamics*. CSLI Publications, 1996.

[96] J.F.A.K. van Benthem. One is a lonely number: on the logic of communication. Technical report, University of Amsterdam, 2002. ILLC Research Report PP-2002-27 (material presented at the Logic Colloquium 2002).

[97] J.F.A.K. van Benthem. What one may come to know. *Analysis*, 64(2):95–105, 2004.

[98] J.F.A.K. van Benthem. Dynamic logic for belief change. *Journal of Applied Non-Classical Logics*, 2006. To appear.

[99] J.F.A.K. van Benthem, J. Gerbrandy, and B.P. Kooi. Dynamic update with probabilities. In W. van der Hoek and M. Wooldridge, editors, *Proceedings of LOFT'06*. Liverpool, 2006.

[100] J.F.A.K. van Benthem and F. Liu. Diversity of logical agents in games. *Philosophia Scientiae*, 8(2):163–178, 2004.

[101] J.F.A.K. van Benthem and E. Pacuit. The tree of knowledge in action: Towards a common perspective. Technical Report, ILLC, University of Amsterdam, 2006.

[102] J.F.A.K. van Benthem, J. van Eijck, and B.P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[103] R. van der Meyden. Axioms for knowledge and time in distributed systems with perfect recall. In *Proceedings of the Ninth Annual IEEE Symposium on Logic in Computer Science (LICS-94)*, pages 448–457, Paris, July 1994.

[104] R. van der Meyden. Common knowledge and update in finite environments. *Information and Computation*, 140(2):115–157, 1998.

[105] H.P. van Ditmarsch. *Knowledge Games*. PhD thesis, University of Groningen, 2000. ILLC Dissertation Series DS-2000-06.

[106] H.P. van Ditmarsch. Descriptions of game actions. *Journal of Logic, Language and Information*, 11:349–365, 2002.

[107] H.P. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275, 2005.

[108] H.P. van Ditmarsch and B.P. Kooi. The secret of my success. *Synthese*, 151:201–232, 2006.

[109] H.P. van Ditmarsch and W.A. Labuschagne. A multimodal language for revising defeasible beliefs. In E. Álvarez, R. Bosch, and L. Villamil, editors, *Proceedings of the 12th International Congress of Logic, Methodology, and Philosophy of Science (LMPS)*, pages 140–141. Oviedo University Press, 2003.

[110] H.P. van Ditmarsch and W.A. Labuschagne. My preferences about your preferences – a case study in theory of mind and epistemic logic. *Knowledge, Rationality & Action (Synthese)*, 155:191–209, 2007.

[111] H.P. van Ditmarsch, W. van der Hoek, and B.P. Kooi. Concurrent dynamic epistemic logic. In V.F. Hendricks, K.F. Jørgensen, and S.A. Pedersen, editors, *Knowledge Contributors*, pages 45–82, Dordrecht, 2003. Kluwer Academic Publishers. Synthese Library Volume 322.

[112] H.P. van Ditmarsch, W. van der Hoek, and B.P. Kooi. Playing cards with Hintikka: An introduction to dynamic epistemic logic. *The Australasian Journal of Logic*, 3:108–134, 2005.

[113] H.P. van Ditmarsch, W. van der Hoek, and B.P. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2007.

[114] P. van Emde Boas, J. Groenendijk, and M. Stokhof. The conway paradox: Its solution in an epistemic framework. In J. Groenendijk, T. M. V. Janssen, and M. Stokhof, editors, *Truth, Interpretation and Information: Selected Papers from the Third Amsterdam Colloquium*, pages 159–182. Foris Publications, Dordrecht, 1984.

[115] B. van Linder, W. van der Hoek, and J.-J.Ch. Meyer. Actions that make you change your mind. In A. Laux and H. Wansing, editors, *Knowledge and Belief in Philosophy and Artificial Intelligence*, pages 103–146, Berlin, 1995. Akademie Verlag.

[116] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[117] G.H. von Wright. *An Essay in Modal Logic.* North Holland, Amsterdam, 1951.

[118] H. Wansing. Diamond's are a philosopher's best friends. the knowability paradox and modal epistemic relevance logic. *Journal of Philosophical Logic*, 31:591–612, 2002.

[119] R. Wassermann. Resource bounded belief revision. *Erkenntnis*, 50(2–3):429–446, 1999.

[120] R. Wassermann. *Resource Bounded Belief Revision.* PhD thesis, University of Utrecht, 2000.

[121] T. Williamson. *Knowledge and Its Limits.* Oxford University Press, Oxford, 2000.

[122] A. Wisniewski. Two logics of occurrent belief. *Acta Universitatis Wratislavensis*, 2023(18):115–121, 1998.

[123] A. Yap. Product update and looking backward. Technical report, University of Amsterdam, 2006. ILLC Research Report PP-2006-39.

Index