14

Beyond Belief: Probability, Dynamics, and Intention

In this chapter we go beyond the model of knowledge and belief introduced in the previous chapter. Here we look at how one might represent statements such as “Mary believes that it will rain tomorrow with probability $> .7$,” and even “Bill knows that John believes with probability .9 that Mary believes with probability $> .7$ that it will rain tomorrow.” We will also look at rules that determine how these knowledge and belief statements can change over time, more broadly at the connection between logic and games, and consider how to formalize the notion of intention.

14.1 Knowledge and probability

In a Kripke structure, each possible world is either possible or not possible for a given agent, and an agent knows (or believes) a sentence when the sentence is true in all of the worlds that are accessible for that agent. As a consequence, in this framework both knowledge and belief are binary notions in that agents can only believe or not believe a sentence (and similarly for knowledge). We would now like to add a quantitative component to the picture. In our quantitative setting we will keep the notion of knowledge as is, but will be able to make statements about the degree of an agent’s belief in a particular proposition. This will allow us to express not only statements of the form “the agent believes with probability .3 that it will rain” but also statements of the form “agent i believes with probability .3 that agent j believes with probability .9 that it will rain.” These sort of statements can become tricky if we are not careful—for example, what happens if $i = j$ in the last sentence?[1]

There are several ways of formalizing a probabilistic model of belief in a multiagent setting, which vary in their generality. We will define a relatively restrictive class of models, but will then mention a few ways in which these can be generalized.

[1] Indeed, some years back the *New Yorker* magazine published a cartoon with the following caption: *There is now 60% chance of rain tomorrow, but there is 70% chance that later this evening the chance of rain tomorrow will be 80%.*

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

common prior

Our technical device will be to simply take our partition model and overlay a commonly known probability distribution (called the *common prior*) over the possible worlds.

common prior

Definition 14.1.1 (Multiagent probability structure) *Given a set X, let $\Pi(X)$ be the class of all probability distributions over X. Then we define a* (common-prior) multiagent probability structure

M over a nonempty set Φ of primitive propositions as the tuple $(W, \pi, I_1, \ldots, I_n, \mathcal{P})$, where:

interpretation

- *W is a nonempty set of* possible worlds*;*
- *$\pi : \Phi \mapsto 2^W$ is an* interpretation *that associates with each primitive proposition $p \in \Phi$ the set of possible worlds $w \in W$ in which p is true;*
- *each I_i is a partition relation, just as in the original partition model (see Definition 13.1.1); and*
- *$\mathcal{P} \in \Pi(W)$ is the* common prior probability.

Adding the probability distribution does not change the set of worlds that an agent considers possible from a world w, but it does allow us to quantify how likely an agent considers each of these possible worlds. In world w, agent i can condition on the fact that it is in the partition $I(w)$ to determine the probability that it is in each $w' \in I(w)$. For all $w' \notin I_i(w)$ it is the case that $P_i(w' \mid w) = 0$, and for all $w' \in I_i(w)$ we have:

$$P_i(w'|w) = \frac{\mathcal{P}(w')}{\sum_{v|v \in I(w)} \mathcal{P}(v)}. \tag{14.1}$$

Note that this means that if w' and w'' lie in the same partition for agent i, then $P_i(w|w') = P_i(w|w'')$. This is one of the restrictions of our formulation to which we return later.

We will often drop the designations "common prior" and "multiagent" when they are understood from the context and simply use the term "probability structure." Next we discuss the syntax of a language to reason about probability structures and its semantics. We define the syntax of this new language formally as follows.

Definition 14.1.2 (Well-formed probabilistic sentences) *Given a set Φ of primitive propositions, the set of well-formed probabilistic sentences $\mathcal{L}_P$ is defined by:*

- *$\Phi \subseteq \mathcal{L}_P$.*
- *If $\varphi, \psi \in \mathcal{L}_P$, then $\varphi \wedge \psi \in \mathcal{L}_P$ and $\neg\varphi \in \mathcal{L}_P$.*
- *If $\varphi \in \mathcal{L}_P$, then $P_i(\varphi) \geq a \in \mathcal{L}_P$, for $i = 1, \ldots, n$ and $a \in [0, 1]$.*

The intuitive meaning of these statements is clear; in particular, $P_i(\varphi) \geq a$ states that the probability of φ is at least a. For convenience, we can use several other comparison operators on probabilities, including $P_i(\varphi) \leq a \equiv P_i(\neg\varphi) \geq a$, $P_i(\varphi) = a \equiv P_i(\varphi) \geq a \wedge P_i(\varphi) \leq a$, and $P_i(\varphi) > a \equiv P_i(\varphi) \geq a \wedge \neg P_i(\varphi) = a$.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

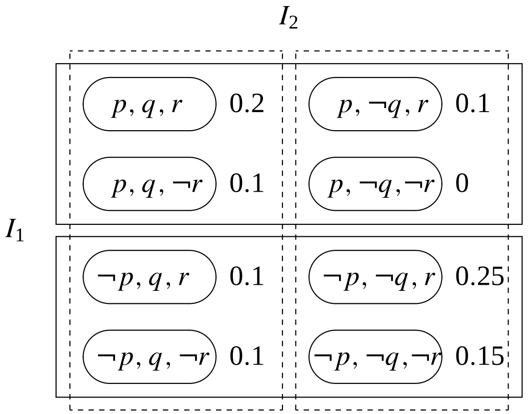

Figure 14.1 A KP structure with a common prior.

Now note that we can use this language and its abbreviations to express formally our sample higher-order sentences given earlier. To express the sentence "agent 1 believes with probability 0.3 that agent 2 believes with probability 0.7 that q," we would write $P_1(P_2(q) = 0.7) = 0.3$. Similarly, to express the sentence "agent 1 believes with probability 1 that she herself believes with probability at least 0.5 that p," we would write $P_1(P_1(p) \geq 0.5) = 1$.

Now we define $\models$, the satisfaction relation, which links our syntax and semantics.

Definition 14.1.3 ($\models$ relation) *Let $p \in \Phi$ be a primitive proposition and φ and ψ be sentences of modal logic. We define the $\models$ relation as:*

- *$M, w \models p$ if and only if $w \in \pi(p)$.*
- *$M, w \models \neg\varphi$ if and only if $M, w \not\models \varphi$.*
- *$M, w \models \varphi \wedge \psi$ if and only if $M, w \models \varphi$ and $M, w \models \psi$.*
- *$M, w \models P_i(\varphi) \geq a$ if and only if $\frac{\sum_{v \mid (v \in I(w)) \wedge (M, v \models \varphi)} \mathcal{P}(v)}{\sum_{v \mid v \in I(w)} \mathcal{P}(v)} \geq a$.*

The following example illustrates the definitions. Consider Figure 14.1. The interpretation of this structure is that each agent knows the truth of exactly one proposition: p for agent 1, and q for agent 2. If the real world is $w = (p, q, r)$, then we have $M, w \models (P_1(q) = 0.75)$, because conditioning on the partition $I_1(w)$ yields: $\frac{0.2+0.1}{0.2+0.1+0.1+0} = 0.75$. We now go a level deeper to consider agent 2 modeling agent 1. In the partition for agent 2 containing the true world, $I_2(w)$, the probability that agent 2 assigns to being in the partition $I_1(w)$ is $\frac{0.2+0.1}{0.2+0.1+0.1+0.1} = 0.6$. Since the other partition for agent 1 yields a different $P_1(q)$, we have the sentence $M, w \models P_2(P_1(q) = 0.75) = 0.6$.

This is a good point at which to discuss to some of the constraints of the theory as we have presented it. To begin with, we have implicitly assumed that the partition structure of each agent is common knowledge among all agents. Second, we assumed that the beliefs of the agents are based on a common prior and are obtained by conditioning on the worlds in the partition. Both are substantive assumptions and have ramifications. For example, we have the fact that the beliefs

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

of an agent are the same within all worlds of any given partition. Also note that we have a strong property of discreteness in the higher-order beliefs. Specifically, the number of statements of the form "$M, w \models P_i(P_j(\varphi) = a) = b$" in which $b > 0$ is equal to at most the number of partitions for agent j, and the sum of all b's from these statements is equal to 1. Thus, you do not need intervals to account for all of the probability mass, as you would with a continuous distribution. In fact, it is the case that for any depth of recursive modeling, you can always decompose a probability over a range into a finite number of probabilities of individual points. One could imagine an alternative formulation in which agent i did not know the prior of agent j, but instead had a continuous distribution over agent j's possible priors. In that case, you could have $M, w \models P_i(P_j(\varphi) \geq b) > 0$ without there being any specific $a \geq b$ such that $M, w \models P_i(P_j(\varphi) = a) > 0$.

We could in fact endow agents with probabilistic beliefs without assuming a common prior and with a much more limited use of the partition structure. For example, we could replace, in each world and for each agent, the set of accessible worlds by a probability distribution over a set of worlds $\mathcal{P}_{i,w}$. In this case the semantic condition for belief would change to:

- $M, w \models P_i(\varphi) \geq a$ if and only if $\sum_{v;M,v\models\varphi} \mathcal{P}_{i,w}(v) \geq a$.

If we want to retain the partition structure so we can speak about knowledge as well as (probabilistic) belief, we could add the requirement that $\mathcal{P}_{i,w}(w') = 0$ for all $w' \notin I(w)$.

However, such interesting extensions bring up a variety of complexities, including, in particular, the axiomatization of such a system, and throughout this book we stay within the confines of the theory as presented.

As we have discussed, every probability space gives rise to an infinite hierarchy of beliefs for each agent: beliefs about the values of the primitive propositions, about other agents' beliefs, beliefs about other agents' beliefs about other agents' beliefs, and so on. This collection of beliefs is called the "epistemic type space" in game theory and plays an important role in the study of games of incomplete information, or Bayesian games, discussed in Section 6.3. (There, each possible world is a different game, and the partition of each agent represents the set of games that the agent cannot distinguish between, given the signal it receives from nature.) One can ask whether the converse holds: Is it the case that every type space is given rise to by some multiagent probability structure? The perhaps surprising answer is yes so long as the type space is *self coherent*. Self coherence, which plays the role analogous to a combination of the positive and negative introspection properties of knowledge and (qualitative) belief, is defined as follows.

self-coherent belief

Definition 14.1.4 (Self-coherent belief) *A collection of higher-order beliefs is* self-coherent *iff, for any agent i, any proposition φ, and any $a \in [0, 1]$, if $P_i(\varphi) \geq a$ then $P_i(P_i(\varphi) \geq a) = 1$.*

common probabilistic belief

Recall that in modal logics of knowledge and belief we augmented the single-agent modalities with group ones. In particular, we defined the notion of *common knowledge*. We can now do the same for probabilistic beliefs and define the notion of *common (probabilistic) belief.*

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

Definition 14.1.5 (Common belief) *A sentence φ is* commonly believed *among two agents a and b, written $C^p_{a,b}(\varphi)$, if $P_a(\varphi) = 1$, $P_b(\varphi) = 1$, $P_a(P_b(\varphi) = 1) = 1$, $P_b(P_a(\varphi) = 1) = 1$, $P_a(P_b(P_a(\varphi) = 1) = 1) = 1$, and so on. The definition extends naturally to common (probabilistic) belief among an arbitrary set G of agents, denoted C^p_G.*

Now that we have the syntax and semantics of the language, we may ask whether there exists an axiomatic system for this language that is sound and complete for the class of common-prior probability spaces. The answer is that one exists, and it is the following.

Definition 14.1.6 (Axiom system $\mathbf{AX}_P$) *The axiom system $\mathbf{AX}_P$ consists of the following axioms and inference rules (in the schemas that follow, φ and ψ range over all sentences, and $n_i, n_j \in [0, 1]$).*

Axiom 14.1.7 (**A1*****)*** *All of the tautological schema of propositional logic*

Axiom 14.1.8 (**P1:** ***Nonnegativity)*** $P_i(\varphi) \geq 0$

Axiom 14.1.9 (**P2:** ***Additivity)*** $P_i(\varphi \wedge \psi) + P_i(\varphi \wedge \neg\psi) = P_i(\varphi)$

Axiom 14.1.10 (**P3:** ***Syntax independence)*** *$P_i(\varphi) = P_i(\psi)$ if $\varphi \Leftrightarrow \psi$ is a propositional tautology*

Axiom 14.1.11 (**P4:** ***Positive introspection)*** $(P_i(\varphi) \geq a) \rightarrow P_i(P_i(\varphi) \geq a) = 1$

Axiom 14.1.12 (**P5:** ***Negative introspection)*** $(\neg P_i(\varphi) \geq a) \rightarrow P_i(\neg P_i(\varphi) \geq a) = 1$

Axiom 14.1.13 (**P6:** ***Common-prior assumption)*** $C^p_{i,j}(P_i(\varphi) = n_i \wedge P_j(\varphi) = n_j) \rightarrow n_i = n_j$

Axiom 14.1.14 (**R1*****)*** *From φ and $\varphi \rightarrow \psi$ infer ψ*

Axiom 14.1.15 (**R2*****)*** *From φ infer $P_i(\varphi) = 1$*

Note that axioms P4 and P5 are the direct analogs of the positive and negative introspection axioms in the modal logics of knowledge and belief. Axiom P6 is novel and captures the common-prior assumption.

Now we are ready to give the soundness and completeness result.

Theorem 14.1.16 *The axiom system $\mathbf{AX}_P$ is sound and complete with respect to the class of all common-prior multiagent probability structures.*

14.2 Dynamics of knowledge and belief

We have so far discussed how to represent "snapshots" of knowledge and belief. We did speak a little about how, for example, knowledge changes over time, for example in the context of the Muddy Children problem. But the theories presented were all static ones. For example, in the Muddy Children problem we used the partition model to represent the knowledge state of the children after each time the father asks his question, but did not give a formal account for how the system

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

transitions from one state to the other. This section is devoted to discussing such dynamics.

We will first consider the problem of *belief revision*, which is the process of revising an existing state of belief on the basis of newly learned information. We will consider the revision of both qualitative and quantitative (i.e., probabilistic) beliefs. Then we will briefly look at dynamic operations on beliefs that are different from revision, such as *update* and *fusion*.

14.2.1 *Belief revision*

One way in which the knowledge and belief of an agent change is when the agent learns new facts, whether by observing the world or by being informed by another agent. Whether the beliefs are categorical (as in the logical setting) or quantitative (as in the probabilistic setting), we call this process *belief revision*. belief revision

When the new information is consistent with the old beliefs, the process is straightforward. In the case of categorical beliefs, one simply adds the new beliefs to the old ones and takes the logical closure of the union. That is, consider a knowledge base (or belief base—here it does not matter) K and new information φ such that $K \not\models \neg\varphi$. The result of revising K by φ, written $K * \varphi$, is simply $Cn(K, \varphi)$, where Cn denotes the logical closure operator. Or thought of semantically, the models of $K * \varphi$ consist of the intersection of the models of K and the models of φ.

The situation is equally straightforward in the probabilistic case. Consider a prior belief in the form of a probability distribution $P(\cdot)$ and new information φ such that $P(\varphi) > 0$. $P * \varphi$ is then simply the posterior distribution $P(\cdot \mid \varphi)$.

Note that in both the logical and probabilistic settings, the assumption that the new information is consistent with the prior belief (or knowledge) is critical. Otherwise, in the logical setting the result of revision yields the empty set of models, and in the probabilistic case the result is undefined. Thus the bulk of the work in belief revision lies in trying to capture how beliefs are revised by information that is inconsistent with the prior beliefs (and, given their defeasibility, these are really beliefs, as opposed to knowledge).

One might be tempted to argue that this is a waste of time. If an agent is misguided enough to hold false beliefs, let him suffer the consequences. But this is not a tenable position. First, it is not only the agent who suffers the consequences, but also other agents—and we, the modelers—who must reason about him. But beyond that, there are good reasons why agents might hold firm beliefs and later retract them. In the logical case, if an agent were to wait for foolproof evidence before adopting any belief, the agent would never believe anything but tautologies. Indeed, the notion of belief is intimately tied to that of *default reasoning* and *nonmonotonic reasoning*, which are motivated by just this observation. default reasoning nonmonotonic reasoning

In the probabilistic case too there are times at which it is unnatural to assign an event a probability other than zero. We encounter this, in particular, in the context of noncooperative game theory. There are situations in which it is a *strictly dominant strategy* (see Section 3.4.3) for an agent to take a certain action, as in the following example.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

> Two foes, Lance and Lot, about to enter into a duel. Each of them must choose a weapon and then decide on a fighting tactic. Lance can choose among two swords—an old, blunt sword, and a new, sharp one. The new one is much better than the old, regardless of the weapon selected by Lot and the tactics selected by either foe. So in selecting his fighting tactic, Lot is justified in assuming that Lance will have selected the new sword with probability one. But what should Lot do if he sees that Lance selected the old sword after all?

backward induction

If this example seems a bit unnatural, the reader might refer to the discussion of *backward induction* in Chapter 3.

Indeed, a great deal of attention has been paid to belief revision by information inconsistent with the initial beliefs. We first describe the account of belief revision in the logical setting; we then show how the account in the probabilistic setting is essentially the same.

Logical belief revision: The AGM model

We start our discussion of belief revision semantically and with a familiar structure—the KB-models of Section 13.7. In that section, KB-models were used to distinguish knowledge from (a certain kind of) belief. Here we use them for an additional purpose, namely, to reason about conditional beliefs. Technically, in Section 13.7, KB-models were used to give meaning to the two modal operators K_i and B_i. Given a particular world, K_i was defined as truth in all worlds "downward" from it and B_i was defined (loosely speaking) as truth in all worlds in the "bottom-most" cluster. But the KB-model can be mined for further information, and here is the intuition. Imagine a KB-model, and a piece of evidence φ. Now erase from that model all the worlds that do not satisfy φ and all the links to and from those worlds; the result is a new, reduced KB-model. The new beliefs of the agent, after taking φ into account, are the beliefs in this reduced KB-model. The following definition makes this precise.

Definition 14.2.1 (Belief revision in a KB-model) *Given a KB-model* $A = (W, \pi, \leq_1, \dots, \leq_n)$ *over* Σ *and any* $\varphi \in \Sigma$, *let* $W(\varphi) = \{w \in W \mid A, w \models \varphi\}$, *let* $\leq_i (\varphi) = \{(w_1, w_2) \in \leq_i \mid A, w_1 \models \varphi, A, w_2 \models \varphi\}$, *and let* $A(\varphi) = (W(\varphi), \pi, \leq_1 (\varphi), \dots, \leq_n (\varphi))$. *Then the beliefs of agent i after receiving evidence φ, denoted* B_i^{φ}, *are defined as*

$$A, w \models B_i^{\varphi}(\psi) \textit{ iff } A(\varphi), w \models B\psi.$$

Figure 14.2 illustrates this definition. The checkered areas contain all worlds that do no satisfy φ.

Note some nice properties of this definition. First, note that you have that $B_i\psi \leftrightarrow B_i^{true}\psi$, where *true* is any tautology. Second, recall the philosophical slogan regarding knowledge as belief that is stable with respect to the truth. We now see, in a precise sense, why our definitions of knowledge and belief can be viewed as embodying this slogan. Consider ψ such that $A, w \models K_i\psi$. Obviously, $A, w \models B_i\psi$, but it is also the case that $A, w \models B_i^{\varphi}\psi$ for any φ such

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

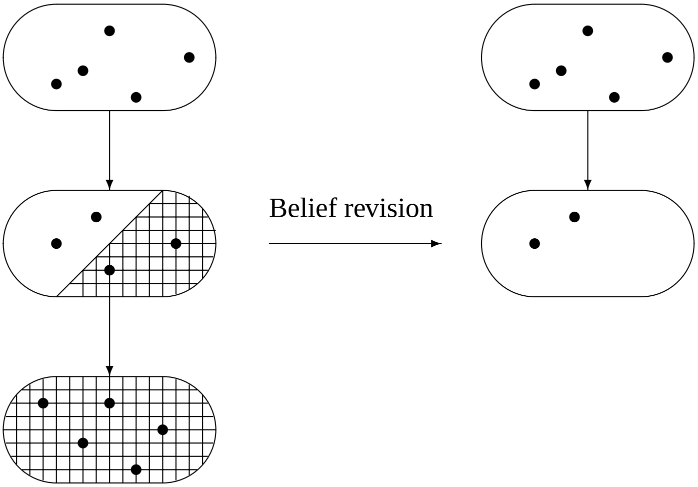

Figure 14.2 Example of belief revision.

that $A, w \models \varphi$. So we get one direction of the slogan—if a proposition is known, then indeed it is a belief that will be retained in the face of any correct evidence. While the converse—that any believed proposition that is not known can be falsified by some evidence—is less straightforward, we point that it would hold if the language Σ were rich enough to capture all propositions.

AGM postulate

prioritization rule

gullibility rule

As usual, we ask whether this semantic definition can be characterized through alternative means. The answer is yes, and in fact we will discuss two ways of doing so. The first is via the so-called *AGM postulates*, so named after Alchurron, Gärdenfors, and Makinson. These set-theoretic postulates take arbitrary belief revision operator $*$ and the initial belief set K and ask, for any evidence φ, what the properties should be of $K * \varphi$, the revision K by φ. For example, one property is the following: $\varphi \in K * \varphi$. This is the *prioritization rule*, or the *gullibility rule*—new information is always accepted and is thus given priority over old one. (As we see below, other rules require that the theory be consistent, which can cause old information to be discarded as a result of the revision.)

The AGM postulates for revising a theory K are as follows. (Recall that $Cn(T)$ denotes the tautological consequence of the theory T.)

($*$1) $K * \varphi = Cn(K * \varphi)$.

($*$2) $\varphi \in K * \varphi$.

($*$3) $K * \varphi \subseteq Cn(K, \varphi)$.

($*$4) If $\neg\varphi \notin K$ then $Cn(K, \varphi) \subseteq K * \varphi$.

($*$5) If φ is consistent then $K * \varphi$ is consistent.

($*$6) If $\models \varphi \leftrightarrow \psi$ then $K * \varphi = K * \psi$.

($*$7) $K * (\varphi \wedge \psi) \subseteq Cn(K * \varphi, \psi)$

($*$8) If $\neg\psi \notin K * \varphi$ then $Cn(K * \varphi, \psi) \subseteq K * (\varphi \wedge \psi)$.

In a sense that we will make precise later, these exactly characterize belief revision with KB-models. But before making this precise, let us discuss an alternative

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

axiomatic characterization of belief revision operators. This is an axiomatic theory of consequence relations. By definition a meta-theory, this axiomatic theory consists of rules of the form “if derivation x is valid then so is derivation y.”

monotonicity rule

It turns out that this meta-theoretic characterization lends particularly deep insight into belief dynamics. For example, in the classical logic setting, the *monotonicity rule* is valid.

(Monotonicity) $\frac{\alpha \vdash \beta}{\alpha, \gamma \vdash \beta}$

nonmonotonic reasoning

The analogous rule for belief revision would be “if $\beta \in K * \alpha$ then $\beta \in K * (\alpha \wedge \gamma)$.” This does not hold for belief revision, and indeed for this reason belief revision is often called *nonmonotonic reasoning*. There are however other meta-rules that hold for belief revision. Consider the following rules; in these we use the symbol $\mid\!\sim$ to represent the fact that we are reasoning about a nonmonotonic consequence relation, and not the classical $\vdash$.

(Left logical equivalence) $\frac{\models \alpha \leftrightarrow \beta, \alpha \mid\!\sim \gamma}{\beta \mid\!\sim \gamma}$

(Right weakening) $\frac{\models \alpha \rightarrow \beta, \gamma \mid\!\sim \alpha}{\gamma \mid\!\sim \beta}$

(Reflexivity) $\alpha \mid\!\sim \alpha$

(And) $\frac{\alpha \mid\!\sim \beta, \alpha \mid\!\sim \gamma}{\alpha \mid\!\sim \beta \wedge \gamma}$

(Or) $\frac{\alpha \mid\!\sim \gamma, \beta \mid\!\sim \gamma}{\alpha \vee \beta \mid\!\sim \gamma}$

(Cautious monotonicity) $\frac{\alpha \mid\!\sim \beta, \alpha \mid\!\sim \gamma}{\alpha \wedge \beta \mid\!\sim \gamma}$

(Rational monotonicity) $\frac{\alpha \wedge \beta \not\mid\!\sim \gamma, \alpha \not\mid\!\sim \neg \beta}{\alpha \not\mid\!\sim \gamma}$

rational consequence relation

A consequence relation that satisfies all of these properties is called a *rational consequence relation*.

The following theorem ties together the semantic notion, the axiomatic one, and the meta-axiomatic one.

Theorem 14.2.2 *Consider propositional language L with a finite alphabet, a revision operator $*$, and a theory $K \subseteq L$. Then the following are equivalent:*

1. *$*$ is defined by a finite total preorder: There is a single-agent KB-model A and a world w such that $A, w \models B\rho$ for each $\rho \in K$, and for each $\varphi, \psi \in L$ it is the case that $\psi \in K * \varphi$ iff $A, w \models B^{\varphi}\psi$;*
2. *$*$ satisfies the AGM postulates;*
3. *$*$ is a rational consequence relation.*

Probabilistic belief revision

As we have discussed, the crux of the problem in probabilistic belief revision is the inability, in the traditional Bayesian view, to condition beliefs on evidence whose prior probability is zero (also known as measure-zero events). Specifically, the definition of conditional probability,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

leaves the conditional belief $P(A \mid B)$ undefined when P(B) = 0.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

Popper function

nonstandard probability

There exist various extensions of traditional probability theory to deal with this problem. In one of them, based on so-called *Popper functions*, one takes the expression $P(A \mid B)$ as primitive, rather than defined. In another, called the theory of *nonstandard probabilities*, one allows as probabilities not only the standard real numbers but also *nonstandard reals*. These two theories, as well as several others, turn out to be essentially the same, except for some rather fine mathematical subtleties. We discuss explicitly one of these theories, the theory of lexicographic probability systems (LPSs).

lexicographic probability system (LPS)

Definition 14.2.3 (Lexicographic probability system) *A (finite)* lexicographic probability system (LPS) *is a sequence $p = p_1, p_2, \ldots, p_n$ of probability distributions. Given such an LPS, we say that event $p(A \mid B) = c$ if there is an index $1 \leq i \leq n$ such that for all $1 \leq j < i$, it is the case that $p_j(B) = 0$, $p_i(B) \neq 0$, and (now using the classical notion of conditional probability) $p_i(A|B) = c$. Similarly, we say that event A has a higher probability than event B ($p(A) > p(B)$) if there is an index $1 \leq i \leq n$ such that for all $1 \leq j < i$, it is the case that $p_j(A) = p_j(B)$, and $p_i(A) > p_i(B)$.*

In other words, to determine the probability of an event in an LPS, one uses the first probability distribution if it is well defined for this event. If it is not, one tries the second distribution, and so on until the probability is well defined. A standard example of this involves throwing a die. It may land on one of its six faces, with equal probability. There is also a possibility that it will land on one of its twelve edges; minuscule as this probability is, it is nonetheless a possible event. Finally, even more improbably, the die may land and stay perched on one of its eight corners. In a classical probabilistic setting one usually accords each of the first six events—corresponding to the die landing on one of the faces—a probability of 1/6, and the other 20 events a probability of zero. In this more general setting, however, one can define an LPS (p_1, p_2, p_3) as follows. p_1 would be the classical distribution just described. p_2 would give a probability of 1/12 to the die landing on each of the edges, and 0 to all other events. Finally, the p_3 would give a probability of 1/8 to the die landing on each of the corners, and 0 to all other events.

LPSs are closely related to AGM-style belief revision.

Theorem 14.2.4 *Consider propositional language L with a finite alphabet, a revision operator $*$, and a theory $K \subseteq L$. Then the following are equivalent.*

1. *$*$ satisfies the AGM postulates for revising K.*
2. *There exists an LPS $p = p_1, \ldots, p_n$ such that $p_1(K) = 1$, and such that for every φ and ψ it is the case that $\psi \in K * \varphi$ iff $p(\psi \mid \varphi) = 1$.*

Thus, we have extended the three-way equivalence of Theorem 14.2.2 to a four-way one.

14.2.2 *Beyond AGM: update, arbitration, fusion, and friends*

In discussing the dynamics of belief we have so far limited ourselves to belief revision, and a specific type of revision at that (namely, AGM-style belief

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

revision). But there are other forms of belief dynamics, and this section discusses some of them. We do so more briefly than with belief revision; at the end of the chapter we provide references to further readings on these topics.

Expansion and contraction

To begin with, there are two simple operations closely linked to revision that are worth noting. *Expansion* is simply the addition of a belief, regardless of whether it leads to a contradiction. The expansion of a theory K by a formula φ, written $K + \varphi$, is defined by

belief expansion

$$K + \varphi = Cn(K \cup \{\varphi\}).$$

belief contraction

Harper identity

Contraction is the operation of removing just enough from a theory to make it consistent with new evidence. The contraction of a theory K by a formula φ, written $K - \varphi$, is reduced to the revision operator via the *Harper identity*,

$$K - \varphi = K \cap (K * \neg\varphi).$$

Levi identity

The *Levi identity* relates the three operations of revision, expansion, and contraction:

$$K * \varphi = (K - \neg\varphi) + \varphi.$$

Update

belief update

Belief *update* is another interesting operation. Similar to revision, it incorporates new evidence into an existing belief state, ensuring that consistency is maintained. But the intuition behind update is subtly different from the intuition underlying revision. In revision, the second argument represents new evidence of facts that were true all along. In update, the second argument represents facts that have possibly become true only after the original beliefs were formed. Thus if an agent believes that it is not raining and suddenly feels raindrops, in the case of revision the assumption is that he was wrong to have those original beliefs, but in the case of update the assumption is that he was right, but that subsequently it started raining.

The different intuitions underlying revision and update translate to different conclusions that they can yield. Consider the following initial belief: "Either the room is white, or else the independence day of Micronesia is November 2 (or both)." Now consider two scenarios. In the first one, you look in the room and see that it is green; the white hypothesis is ruled out and you infer that the independence day of Micronesia is November 2. This is an instance of belief revision. But now consider a different scenario, in which a painter emerges from the room and informs you that he has just painted the room green. You now have no business inferring anything about Micronesia. This is an instance of belief update.[2]

[2] And good thing too, as the Federated States of Micronesia, which have been independent since 1986, celebrate independence day on November 3.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

Since revision and update are different, it is to be expected that the set of postulates governing update will differ from the AGM postulates. The update of a theory K by a formula φ is written $K \diamond \varphi$. The so-called KM postulates for updating a (consistent) theory K (so named after Katsuno and Mendelzon) read as follows.

($\diamond$1) $K \diamond \varphi = Cn(K \diamond \varphi)$.

($\diamond$2) $\varphi \in K \diamond \varphi$.

($\diamond$3) If $\varphi \in K$ then $K \diamond \varphi = K$.

($\diamond$4) $K \diamond \varphi$ is inconsistent iff φ is inconsistent.

($\diamond$5) If $\models \varphi \equiv \psi$ then $K \diamond \varphi = K \diamond \psi$.

($\diamond$6) $K \diamond (\varphi \wedge \psi) \subseteq (K \diamond \varphi) + \psi$.

($\diamond$7) If $\psi \in K \diamond \varphi$ and $\varphi \in K \diamond \psi$ then $K \diamond \varphi = K \diamond \psi$.

($\diamond$8) If K is complete[3] then $K \diamond (\varphi \wedge \psi) \subseteq K \diamond \varphi \cap K \diamond \psi$.

($\diamond$9) $K \diamond \varphi = \cap_{M \in Comp(K)} M \diamond \varphi$, where $Comp(K)$ denotes the set of all complete theories that entail K.

As in the case of revision, the model theory of update is also well understood, and is as follows (this discussion is briefer than that for revision, and we point the reader to further reading at the end of the chapter).

Revision and update can be related by the identity

$$K \diamond \varphi = \cap_{M \in Comp(K)} M * \varphi.$$

Comparing this property to the last postulate for update, one can begin to gain intuition for the model theoretic characterization of the operator. In particular, it is the case that there is no distinction between revision and update when dealing with complete theories. More generally, the following completely characterizes the class of update operators that obey the KM postulates.

Theorem 14.2.5 *The following two statements about an update operator $\diamond$ are equivalent:*

1. *$\diamond$ obeys the KM postulates;*
2. *There exists a function that maps each interpretation M to a partial preorder $\leq_M$ such that $Mod(K \diamond \varphi) = \cup_{M \in Comp(K)} Min(Mod(M), \leq_M)$.*

Note two important contrasts with the model-theoretic characterization of belief revision. First, here the preorders need not be total. Second, and more critically, in revision there is one global preorder on worlds; here each interpretation induces a different preorder.

Arbitration

Returning to AGM-style revision, we note two aspects of asymmetry. The first is blatant—new evidence is given priority over existing beliefs. The second is more subtle—the first argument to any AGM revision operator is a “richer” type

[3] A theory K is complete if for each sentence φ, either $\varphi \in K$ or $\neg\varphi \in K$.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

of object than the second. In this subsection we discuss the first asymmetry, and in the next subsection the second.

There are certainly situations in which new evidence is not necessarily given precedence over existing beliefs. Indeed, from the multiagent perspective, each revision can be seen as involving at least two agents—the believer and the informer. When synthesizing a new belief, the believer may wish to accord himself higher priority, to not favor one of them over the other, or to give priority to one over the other depending on the subject matter. Various theories exist to capture this intuition. One of them is the theory of belief *arbitration*, which takes an egalitarian approach: it does not favor either of the two sides over the other. Technically speaking, it does so by jettisoning the second AGM postulate $\varphi \in K * \varphi$, and replaces it with a "fairness" axiom:

belief arbitration

$$\text{if } K \cup \{\varphi\} \models \bot, \text{ then } K \not\subseteq K * \varphi \text{ and } \varphi \notin K * \varphi.$$

The intuition is that if the new evidence is inconsistent with the initial beliefs, then each must "give up something."

Fusion

The other source of asymmetry in AGM revision is more subtle. The AGM postulates obscure the asymmetry between the two arguments. In $K * \varphi$, K is a *belief set*, or a set of sentences (which happen to be tautologically closed). True, usually we think of φ as a single sentence, but for most purposes it would matter little if we have it represent a set of sentences. However, it must be remembered that the AGM postulates do not define the operator $*$, but only constrain it. As is seen from Theorem 14.2.2, any particular $*$ is defined with respect to a complete *belief state* or total preorder on possible worlds. This belief state defines the initial belief set, but it defines much more, namely, all the beliefs conditional on new evidence. Thus every specific AGM operator takes as its first input a belief state and as its second a mere belief set. We now consider what happens when the second argument is also a belief state.

belief set

belief state

The first question to ask is what does it mean intuitively to take a second belief state as an argument. Here again the multiagent perspective is useful. We think of a belief set as describing "current" beliefs and a belief state as describing "conditional" (i.e., "current" as well as "hypothetical") beliefs. According to AGM revision, then, the believer has access to his full conditional beliefs, but the informer reveals only (some of his) current beliefs. However, one could imagine situations in which the informer engages in "full disclosure," revealing not only all his current beliefs but also all his hypothetical ones. When this is the case, the believer is faced with the task of merging two belief states; this process is called belief *fusion*.

Belief fusion calls for resolving conflicts among the two belief states. From the technical standpoint, the basic unit of conflict here is no longer the inconsistency of two beliefs, but rather the inconsistency of the two orderings on possible worlds. Here is how one theory resolves such conflicts; we present it in some detail since it very explicitly adopts a multiagent perspective.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

To develop intuition for the following definitions, imagine a set of information sources and a set of agents. The sources can be thought of as primitive agents with fixed belief states. Each source informs some of the agents of its belief state; in effect, each source offers the opinion that certain worlds are more likely than others and remains neutral about other pairs.

An agent's belief state is simply the amalgamation of all these opinions, each annotated by its origin (or "pedigree"). Of course, these opinions in general conflict with one another. To resolve these conflicts, the agent places a strict "credibility" ranking on the sources and accepts the highest-ranked opinion offered on every pair of worlds.

We define the *pedigreed belief state* as follows.

pedigreed belief state

Definition 14.2.6 (Pedigreed belief state) *Given a finite set of belief states S (over $\mathcal{W}$), the* pedigreed belief state *(over $\mathcal{W}$) induced by S is a function $\Psi : \mathcal{W} \times \mathcal{W} \mapsto 2^{S \cup \{s_0\}}$ such that $\Psi(w_1, w_2) = \{(\mathcal{W}, \leq) \in S : w_2 \not\leq w_1\} \cup \{s_0\}$.*

In words, $\Psi(w_1, w_2)$ is the set of all agents who do not believe that world w_2 is at least as likely as w_1. The agent s_0 has no beliefs and is thus always present.

We will use $\mathcal{S}$ to denote the set of all of sources over $\mathcal{W}$, and throughout this section we will consider pedigreed belief states that are induced by subsets of $\mathcal{S}$. Note that both $\{\}$ and s_0 induce the same pedigreed belief state; by slight abuse of notation we will denote it too by s_0.

Next we define a particular policy for resolving conflicts within a pedigreed belief state, since for any two worlds w_1 and w_2, we have the competing camps of $\Psi(w_1, w_2)$ and $\Psi(w_2, w_1)$. We assume a strict ranking $\sqsubset$ on $\mathcal{S}$ (and thus also on the sources that induce any particular Ψ). We interpret $s_1 \sqsubset s_2$ as "s_2 is more credible than s_1." As usual, we define $\sqsubseteq$, read "as credible as," as the reflexive closure of $\sqsubset$.

We also assume that s_0 is the least credible source, which may merit some explanation. It might be asked why equate the most agnostic source with the least credible one. In fact we do not have to, but since in the definitions that follow, agnosticism is overridden by any opinion regardless of credibility ranking, we might as well assume that all agnosticism originates from the least credible source, which will permit simpler definitions.

Intuitively, given a pedigreed belief state Ψ, $\Psi_{\sqsubset}$ will retain from Ψ the highest-ranked opinion about the relative likelihood between any two worlds.

dominating belief state

Definition 14.2.7 (Dominating belief state) *Given $\mathcal{W}$, $\mathcal{S}$, Ψ and $\sqsubset$ as defined earlier, the* dominating belief state *of Ψ is the function $\Psi_{\sqsubset} : \mathcal{W} \times \mathcal{W} \mapsto \mathcal{S}$ such that $\forall w_1, w_2 \in \mathcal{W}$ the following holds: If $\max(\Psi(w_2, w_1)) \sqsubset \max(\Psi(w_1, w_2))$ then $\Psi_{\sqsubset}(w_1, w_2) = \max(\Psi(w_1, w_2))$. Otherwise, $\Psi_{\sqsubset}(w_1, w_2) = s_0$.*[4]

Clearly, for any $w_1, w_2 \in \mathcal{W}$ either $\Psi_{\sqsubset}(w_1, w_2) = s_0$ or $\Psi_{\sqsubset}(w_2, w_1) = s_0$ or both.

It is not hard to see that $\Psi_{\sqsubset}$ induces a standard (anonymous) belief state.

[4] Note the use of the restrictions. Finiteness assures that a maximal source exists; we could readily replace it by weaker requirements on the infinite set. The absence of ties in the ranking $\sqsubset$ ensures that the maximal source is unique; removing this restriction is not straightforward.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

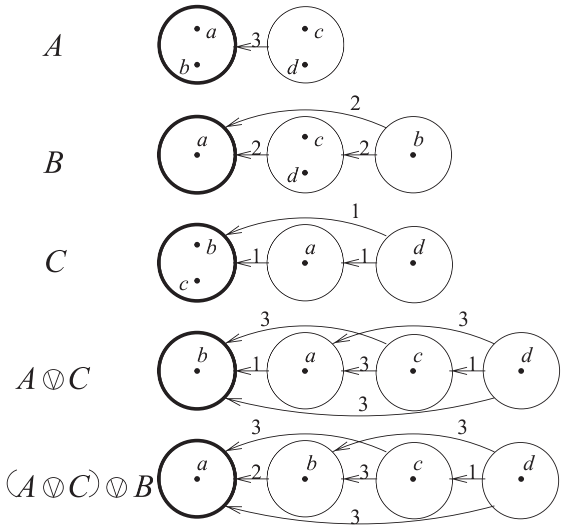

Figure 14.3 Example of a fusion operator. Only the dominating belief states are shown.

Definition 14.2.8 (Ordering induced by $\Psi_\sqsubset$) *The* ordering induced by $\Psi_\sqsubset$ *is the relation $\preceq$, a binary relation on $\mathcal{W}$, such that $w_1 \preceq w_2$ iff $\Psi_\sqsubset(w_2, w_1) = s_0$.*

Clearly, $\preceq$ is a total preorder on $\mathcal{W}$. Thus a dominating belief state is a generalization of the standard notion of belief state.

We are now ready to define the fusion operator.

belief fusion

Definition 14.2.9 (Belief fusion) *Given a set of sources $\mathcal{S}$ and $\sqsubset$ as previously, $S_1, S_2 \subset \mathcal{S}$, the pedigreed belief state Ψ_1 induced by S_1, and pedigreed belief state Ψ_2 induced by S_2, the* fusion *of Ψ_1 and Ψ_2, denoted $\Psi_1 \circledvee \Psi_2$, is the pedigreed belief state induced by $S_1 \cup S_2$.*

Figure 14.3 illustrates the operation of the fusion operator. Of the three agents, A has the highest priority at 3, followed by B and C with priorities of 2 and 1, respectively. The first three lines describe the beliefs of the agents over the four worlds a,b,c, and d in the form of a dominating belief state. An arrow from one circle to another means that all worlds in the second circle are considered at least as likely as all worlds in the first. Each arrow is labeled with the priority of the agent who holds those beliefs. The final two lines show examples of fusion, with the corresponding diagrams showing the resulting dominating belief state. When fusing the beliefs of A and C, we see that all of A's beliefs are retained because it has a higher priority. Since A has no opinion on the pairs (a, b) and (c, d), we take C's beliefs. Of C's two remaining beliefs, $c \leq_C a$ is overruled by $a \leq_A c$, while $b \leq_c d$ is consistent with A's belief and thus does not show up in the output. When we fuse this dominating belief state with that of B, the only possible changes that can be made are on the pairs (a, b) and (c, d), since A has a belief on all other pairs and has a higher priority than B. Agent B has no opinion on (c, d) but disagrees with C on (a, b), causing a reversal of the arrow, which is now labeled with B's priority.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

14.2.3 *Theories of belief change: a summary*

We have discussed a number of ways in which beliefs change over time. Starting with AGM revision and its probabilistic counterpart, we went on to discuss the operations of expansion, contraction, update, arbitration, and fusion.

It is important to emphasize that these examples are not exhaustive. Because it is so important, we devote this short subsection to making just this point. There are other specific theories, for example, theories accounting for the *iteration* of belief revision; most of the theories we discussed do not say what happens if one wishes to conduct, for example, two revisions in sequence. Other theories are more abstract, and provide a general framework for belief change, of which revision, update, and other operators are special instances. Among them, importantly, are theories with *information change operators*. These are couched in propositional dynamic logic, which we will encounter in Section 14.4, but in which the modal action operators are indexed by a logical proposition, denoting, for example, the learning of that proposition. In Section 14.5 we provide some references to this literature.

iterated belief revision

information change operators

14.3 Logic, games, and coalition logic

So far in this chapter we looked at the use of logic to reason about purely “informational” aspects of agents. Now we broaden the discussion to include also “motivational” aspects. In this section we (briefly) look at the interaction between modal logic and game theory.

The connection between logic and games is multifaceted. Even before the advent of modern-day game theory, in the context of classical logic, it was proposed that a logical system be viewed as a game between a prover (the “asserter”) and a disprover (the “respondent”), with a sentence being valid if the prover had a winning strategy. In this case, games are used to reason about logic. But much recent work has been aimed in the opposite direction, with logic being used to reason about games. Here games are meant in the formal sense of modern game theory, and the logics are modal rather than classical.

We have so far seen how modal logic can be used to model and reason about agents’ knowledge and the beliefs and how those change in time. But modal logic can be used to reason also about their actions, preferences, and hence also about games and (certain) solution concepts. Much of the literature here focuses on extensive-form games of both perfect and imperfect (though not yet incomplete) information. There exists a rapidly expanding literature on the topic which is somewhat complex. We will just mention here that these logics allow us to reason about certain solution concepts, particularly those involving only pure strategies. And so one can recapture in logic the notion of (pure strategy) dominant strategy, iterated elimination of dominated strategies, rationalizability, and pure-strategy Nash equilibria.

In lieu of full discussion of this somewhat complex material, to give a feel for what can be expressed in such logics we briefly look at one particular

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

coalition logic

exemplar—so-called *coalition logic* (CL). In the language of CL we do not model the actions of the agents, but rather the capabilities of groups. (In this respect, CL is similar to coalitional game theory.) For any given set of agents C, the modal operator $[C]$ is meant to capture the capability of the group. $[C]\varphi$ means that the group can bring about φ (or, equivalently, can ensure that φ is the case), regardless of the actions of agents outside the set C.

The formal syntax of CL is defined as follows. Given a (finite, nonempty) set of agents N and a set of primitive propositions Φ_0, the set Φ of well-formed formulas is the smallest set satisfying the following:

$\Phi_0 \subset \Phi$;

If $\varphi_1, \varphi_2 \in \Phi$, then $\neg\varphi_1 \in \Phi$ and $\varphi_1 \vee \varphi_2 \in \Phi$;

If $\varphi \in \Phi$, $C \subset N$, and φ is $[C']$-free for all C', then $[C]\varphi \in \Phi$.

$\top$ is shorthand for $\neg\bot$, and $\wedge$, $\rightarrow$, and $\leftrightarrow$ are defined as usual. $\bot$ can be viewed as shorthand for $p \wedge \neg p$, and $[i]\varphi$ is shorthand for $[\{i\}]\varphi$ for any $i \in N$.[5]

The formal semantics of CL are as follows. A CL model is a triple (S, E, V), such that:

S is a set of states or worlds;

$V : \Phi_0 \mapsto 2^S$ is the valuation function, specifying the worlds in which primitive propositions hold;

$E : 2^N \mapsto 2^{2^S}$ such that:

- $E(\{\}) = \{S\}$,
- if $C \subset C'$, then $E(C')$ is a refinement of $E(C)$.

The satisfaction relation is defined as follows:

$(S, E, V) \not\models \bot$;

for $p \in \Phi_0$, $(S, E, V) \models p$ iff $s \in V(p)$;

$(S, E, V) \models \neg\varphi$ iff $(S, E, V) \not\models \varphi$;

$(S, E, V) \models \varphi_1 \vee \varphi_2$ iff $(S, E, V) \models \varphi_1$ or $(S, E, V) \models \varphi_2$;

$(S, E, V) \models [C]\varphi$ iff there exists $S' \in E(C)$ such that for all $s \in S'$ it is the case that $s \models \varphi$ (here $\models$ is used in the classical sense).

What can be said about the sentences that are valid in this logic? For example, clearly, both $\neg[C]\bot$ and $[C]\top$ are valid (no coalition can force a contradiction, and tautologies are true in any model, and thus in any set forced by a coalition). Equally intuitively, $[C](\varphi_1 \wedge \varphi_2) \rightarrow [C]\varphi_2$ is also valid (after all, if a coalition can enforce an outcome to lie within a given set of worlds, it can also enforce it to lie within a superset). Perhaps more insightful is the following valid sentence: $([C_1]\varphi_1 \wedge [C_2]\varphi_2) \rightarrow [C_1 \cup C_2](\varphi_1 \wedge \varphi_2)$, for any $C_1 \cap C_2 = \emptyset$ (if one coalition

[5] This is in fact a simplified version of CL, which might be called *flat CL*. The full definition allows for the nesting of $[C]$ operators, but that requires semantics that are too involved to include here.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

can force some set of worlds, and a disjoint coalition can force another set of worlds, then together they can enforce their intersection). The discussion at the end of the chapter points the reader to a more complete discussion of this and related logics.

14.4 Towards a logic of "intention"

In this section we look, briefly again, at modal logics that have explicit "motivational" modal operators, ones that capture the motivation of agents and their reason for action. The specific notion we will try to capture is that of *intention* first as it is attributed to a single agent and then as it is attributed to set of agents functioning as a group. As we shall see, in service of "intention" we will need to define several auxiliary notions. These extended theories are sometimes called jokingly *BDI (pronounced 'beady eye') theories*, because, beside the notion of belief (B), they include the notions of desire (D) and intention (I), as well as several others.

intention

BDI theories

It should be said at the outset that these extended theories are considerably more complex and messy than the theories of knowledge and belief that were covered in previous sections, and mathematically less well developed than those on games and information dynamics. We will therefore not treat this subject matter with the same amount of detail as we did the earlier topics. Instead we will do the following. We will start with an informal discussion of notions such as intention and some requirements on a formal theory of them. We will then outline only the syntax of one such theory, starting with a single-agent theory and then adding a multiagent component.

14.4.1 *Some preformal intuitions*

Let us start by considering what we are trying to achieve. Recall that our theories of "knowledge" and "belief" only crudely approximated the everyday meanings of those terms, but were nonetheless useful for certain applications. We would like to strike a similar balance in a theory of "intention." Such theories could be use to reason about cooperative or adversarial planning agents, to create intelligent dialog systems, or to create useful personal assistants, to mention a few applications.

What are some of the requirements on such theories? Different applications will give rise to different answers. For motivation behind a particular theory we will outline, consider the following scenario.

> Phil is having trouble with his new household robot, Hector. He says, "Hector, bring me a beer." The robot replies, "OK, boss." Twenty minutes later, Phil yells, "Hector, why didn't you bring that beer?" It answers, "Well, I had intended to get you the beer, but I decided to do something else." Miffed, Phil sends the wise guy back to the manufacturer, complaining about a lack of commitment. After retrofitting, Hector is returned, marked "Model C: The Committed Assistant." Again, Phil asks Hector to bring a beer. Again,

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

> it accedes, replying "Sure thing." Then Phil asks, "What kind do we have?" It answers, "Anchor Steam." Phil says, "Never mind." One minute later, Hector trundles over with an Anchor Steam in its gripper. This time, Phil angrily return Hector for overcommitment. After still more tinkering, the manufacturer sends Hector back, promising no more problems with its commitments. So, being a somewhat trusting consumer, Phil accepts the rascal back into his household, but as a test, he asks Hector to bring him the last beer remaining in the fridge. Hector again accedes, saying, "Yes, sir." The robot gets the beer and starts toward Phil. As it approaches, it lifts its arm, wheels around, deliberately smashes the bottle, and trundles off. Back at the plant, when interrogated by customer service as to why it had abandoned its commitments, the robot replies that according to its specifications, it could not have a commitment that it believed to be unachievable. Once the last remaining bottle was smashed, the commitment became unachievable. Despite the impeccable logic, and the correct implementation, Hector is dismantled.

intention

commitment

capability

desire

goal

agency

This example suggests that in order to capture the notions of *intention* we must also tackle those of *commitment* and *capability*. In addition, we will consider the notions of *desire*, *goal*, and even *agency*. This scenario also suggests some constraints. Here are some of the intuitions that will underlie the formal development.

rational balance

1. Desires are unconstrained; they need not be achievable, and not even consistent (one can desire smoking and good health simultaneously). Goals in contrast must be consistent with each other, and furthermore must be believed to be achievable, or at least not believed to be unachievable. The same is true of intentions, which have the additional property of being persistent in time. Loosely speaking, these three notions form a hierarchy; goals imply desires, and intentions imply goals. These mutual constraints among the difference notions are sometimes called *rational balance*. In the formulation in the next section we will consider goals and intentions, but not desires.
2. Intentions come in two varieties—intentions to achieve a particular state (such as being in San Francisco) and intentions to take a particular action (such as boarding a particular train to San Francisco). In particular, intentions are future directed. The same is true of goals.
3. Plans consist of a set of intentions and goals. Plans are in general *partial*; that is, they have some goals or intentions that are not directly achievable by the agent. For example, an agent may intend to go to San Francisco, even though he may not yet know whether he wants to drive there or take the train.
4. Plans give rise to further goals and intentions. For example, an intention to go to San Francisco requires that the agent further specify some means for getting there (e.g., driving or taking the train).
5. At the same time, since the plan of an agent must be internally consistent, a given plan constrains the addition of new goals and intentions. For example, if the agent's plan already contains the intention of leaving his car at home

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

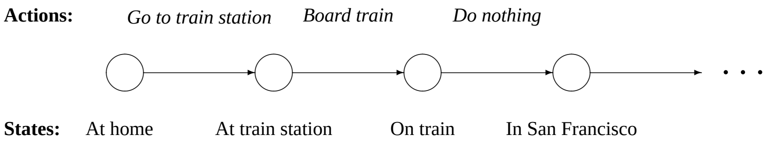

Figure 14.4 An event sequence: actions lead from one state to another.

for his wife to use, then he cannot adopt the intention of driving it to San Francisco as the means of getting there.

6. Intentions are persistent, but not irrevocable. For example, an intention to go to San Francisco may be part of a larger plan serving a higher-level goal, such as landing a new job. Our agent may find another means to achieve the same goal and thus drop his intention to go to San Francisco. Or, he may obtain new information that makes his trip to San Francisco infeasible, in which case too that intention must be dropped.
7. Agents need not intend the anticipated side effects of their intentions. Driving to San Francisco may necessarily increase one's risk of a traffic accident, but an agent intending the former does not generally intend the latter. Similarly, an agent intending to go to the dentist does not typically have the intention of experiencing great pain even if he may fully expect such pain.

Philosophers have written at considerably greater length on these topics, but let us stop here and look at what a formal theory might look like in light of some of these considerations.

14.4.2 *The road to hell: elements of a formal theory of intention*

We will sketch a theory aimed at capturing the notion of intention. Our sketch will be incomplete, and will serve primarily to highlight some of the surprising elements that a complete theory requires. Certainly we will only consider a propositional theory; there is no special difficulty in extending the treatment to the first-order case. More critically, we will discuss only the axiomatic theory, and will have little to say about its formal semantics.

Since intentions are future directed, we must start by representing the passage of time. Although we have just said that we will not discuss semantics, it is useful to keep in mind the picture of a sequence of events, leading from one state to another. Figure 14.4 shows an example.

Such event sequences will form the basis for reasoning about intentions and other notions. The language we use to speak about such events is based on *dynamic logic*. Dynamic logic takes the basic events (such as going to the train station) as the primitive objects. These basic events can be combined in a number of ways to make complex events; see later. By associating an agent with an event (where basic or complex) we get an *action*. Since we will not consider agent-less events here, we will use the term “event” and “action” interchangeably.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

Thus one component of dynamic logic is a language to speak about actions. Another component is a language to speak about what is true and false in states, the end points of actions. In this language we will have two primitive modal operators—B (belief) and G (goal)—and define intention in terms of those. We will not define desire in the language since our notion of intention will not depend on it.

Interestingly, the language of actions and the language of states are mutually dependent. They are defined as follows.

Definition 14.4.1 (Action and state expressions) *Given a set of agents N, a set E of primitive actions, a set V of variables (which will range over actions), and a set P of primitive propositions (describing what is true and false in states), the set of action expressions A and the set of state expressions $\mathcal{L}_{BG}$ are the smallest sets satisfying the following constraints:*

- $E \subseteq A$.
- *If* $a, b \in A$, *then* $a; b \in A$.
- *If* $a, b \in A$, *then* $a|b \in A$.
- *If* $a \in A$, *then* $a^* \in A$.
- *If* $\varphi \in \mathcal{L}_{BG}$, *then* $\varphi? \in A$.

Here is an intuitive explanation of these complex actions. $a; b$ denotes composition: doing a followed by b. $a|b$ denotes nondeterministic choice: doing a or b, nondeterministically. a^* denotes iteration: repeating a zero or more times, nondeterministically. $\varphi?$ is perhaps the most unusual action. It is the test action; it does not change the state, but is successful only in states that satisfy φ. Specifically, if the action $\varphi?$ is taken in a state, then φ must be true in that state.

- $P \subseteq \mathcal{L}_{BG}$.
- *If* $\varphi, \psi \in \mathcal{L}_{BG}$, *then* $\varphi \wedge \psi, \neg\varphi \in \mathcal{L}_{BG}$.
- *If* $a \in A$, *then JustHappened*$(a) \in \mathcal{L}_{BG}$.
- *If* $a \in A$, *then AboutToHappen*$(a) \in \mathcal{L}_{BG}$.
- *If* $i \in N$ *and* $a \in A$, *then Agent*$(a) = i \in \mathcal{L}_{BG}$.
- *If* $v \in V$ *and* $\varphi \in \mathcal{L}_{BG}$, *then* $\forall v\varphi \in \mathcal{L}_{BG}$.
- *If* $a, b \in A \cup V$, *then* $a < b \in \mathcal{L}_{BG}$.
- *If* $i \in N$ *and* $\varphi \in \mathcal{L}_{BG}$, *then* $B_i\varphi \in \mathcal{L}_{BG}$.
- *If* $i \in N$ *and* $\varphi \in \mathcal{L}_{BG}$, *then* $G_i\varphi \in \mathcal{L}_{BG}$.

Again, an explanation is in order. JustHappened(a) captures the fact that action a ended in the current state, and AboutToHappen(a) that it is about to start in the current state. Agent$(a) = i$ identifies i as the agent of action a. $a < b$ means that action a took place before action b. Finally, B_i and G_i of course denote the belief and goal operators, respectively. However, especially for G_i, it is important to be precise about the reading; $G_i\varphi$ means that φ is true in all the states that satisfy the goals of agent i.

While we do not discuss formal semantics here, let us briefly discuss the belief and goal modal operators. They are both intended to be interpreted via possible-worlds semantics. So the question arises as to the individual properties

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

of these operators and the interaction between them. B is a standard (KD45) belief operator. We place no special restriction on the G operator, other than it that must be serial (to ensure that goals are consistent). However, since goals must also be consistent with beliefs, we require that the goal accessibility relation be a subset of the belief accessibility relation. That is, if a possible world is ruled out by the agent's beliefs, it cannot be a goal. This assumption means that you cannot intend that be in San Francisco on Saturday if you believe that you will be in Europe at the same time. (Of course, you can intend to be in San Francisco on Saturday even though you are in Europe today.)[6]

Before proceeding further, it will be useful to define several auxiliary operators.

Definition 14.4.2 $Always(\varphi) \overset{\text{def}}{\equiv} \forall a(AboutToHappen(a) \rightarrow AboutToHappen(a; \varphi?))$.

$Eventually(\varphi) \overset{\text{def}}{\equiv} \neg Always(\neg\varphi)$.

$Later(\varphi) \overset{\text{def}}{\equiv} (\neg\varphi) \wedge (Eventually(\varphi))$.

$Before(\varphi, \psi) \overset{\text{def}}{\equiv} \forall c(AboutToHappen(c; \psi?)) \rightarrow \exists a((a \leq c) \wedge AboutToHappen(a; \varphi?))$.

$JustDid_i(a) \overset{\text{def}}{\equiv} JustHappened(a) \wedge Agent(a) = i$.

$AboutToDo_i(a) \overset{\text{def}}{\equiv} AboutToHappen(a) \wedge Agent(a) = i$.

With these in place, we proceed to define the notion of intention as follows. We first strengthen the notion of having a goal. The G_i operator defines a weak notion; in particular, it includes goals that are already satisfied and thus provides no impetus for action. An *achievement goal*, or $AGoal$, focuses on the goals that are yet to be achieved.

achievement goal

Definition 14.4.3 (Achievement goal)

$$AGoal_i\varphi \overset{\text{def}}{\equiv} G_i(Later(\varphi)) \wedge B_i(\neg\varphi)$$

An achievement goal is useful, but it is still not an intention. This is because an achievement goal has no model of commitment. An agent might form an achievement goal, only to drop it moments later, for no apparent reason. In order to model commitment, we should require that the agent not drop the goal until he reaches it. We call such a goal a persistent goal, defined as follows.

Definition 14.4.4 (Persistent goal)

$$\begin{aligned} PGoal_i\varphi \quad \overset{\text{def}}{\equiv} \quad & AGoal_i\varphi \wedge \\ & Before(B_i(\varphi) \vee B_i(Always(\neg\varphi)), \neg G_i(Later(\varphi))) \end{aligned}$$

In other words, a persistent goal is an achievement goal that the agent will not give up until he believes that it is true or will never be true.

The notion of a persistent goal brings us closer to a reasonable model of intention, but it still misses an essential element of intentionality. In particular, it

[6] Note that this constraint means that the formula $B_i\varphi \rightarrow G_i\varphi$ is valid in our models; this is where the careful reading of the G_i operator is required.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

does not capture the requirement that the agent himself do something to fulfill the goal, let alone do so knowingly. The following definitions attempt to add these ingredients.

Recall that an agent may intend either to do an action or to achieve a state. For this reason, we give two definitions of intention. The following is a definition of intending an action.

Definition 14.4.5 (Intending an action)

$$IntendsA_i a \overset{\text{def}}{\equiv} PGoal_i(JustDid_i(B_i(AboutToDo_i(a))?; a))$$

In other words, if an agent intends an action, he must have a persistent goal to first believe that he is about to take that action and then to actually take it.

The second definition captures the intention to bring about a state with certain properties.

Definition 14.4.6

$$\begin{aligned} IntendS_i(\varphi) \quad \overset{\text{def}}{\equiv} \quad & PGoal_i \exists e JustDid_i(B_i(\exists e' AboutToDo_i(e'; \varphi?)) \wedge \\ & \neg G_i(\neg AboutToDo_i(e; \varphi?))?; e; \varphi?)) \end{aligned}$$

We explain this definition in a number of steps. Notice that to intend a state in which φ holds, an agent is committed to taking a number of actions e himself, after which φ holds. However, in order to avoid allowing him to intend φ by doing something accidentally, we require that he believe he is about to do some series of events e' that brings about φ. In other words, we require that he has a plan e', which he believes that he is executing, which will achieve his goal φ.

As was mentioned at the beginning of this section, the logic of intention is considerably more complex, messy, and controversial than that of knowledge and belief. We will not discuss the pros and cons of this line of definitions further, nor alternative definitions; the notes at the end of the chapter provide references to these. Let us just note that these definitions do have some desirable formal properties. For example, early intentions preempt later potential intentions (namely those that undermine the objectives of the earlier intentions); agents cannot intend to achieve something if they believe that no sequence of actions on their part will bring it about; and agents do not necessarily intend all the side effects of their intentions.

14.4.3 *Group intentions*

The theory of intention outlined in the previous section considers a single agent. We have seen that the extension of the logic of (e.g.,) knowledge to the multiagent setting is interesting; in particular, it gives rise to the notion of common knowledge, which is so central to reasoning about coordination. Is there similar motivation for considering intention in a multiagent setting?

Consider a set of agents acting as a group. Consider specifically the notion of convoy. Consisting of multiple agents, each making local decisions, a convoy nonetheless moves purposefully as a single body. It is tempting to attribute beliefs

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

and intentions to this body, but how do these group notions relate to the beliefs and intentions of the individual agents?

It certainly will not do to equate a group intention simply with all the individual agents having that intention; it is not even sufficient to have those individual intentions be common knowledge among the agents. Consider the following scenario.

> Two ancient Inca kings, Euqsevel and Nehoc, aim to ride their horses in a convoy to the Temple of Otnorot. Specifically, since King Euqsevel knows the way, he intends to simply ride there and King Nehoc intends to follow him. They set out, and since King Euqsevel's horse is faster, he quickly loses King Nehoc, who never makes it to Otnorot.
>
> Centuries later, in old Wales, two noblemen—Sir Ffegroeg and Sir Hgnis—set out on another journey. Here again only one of them, Sir Ffegroeg, knows the way. However, having learned from the Incan mishap, they do not simply adopt their individual intentions. Instead, Sir Ffegroeg agrees to go ahead on his faster horse, but to wait at road junctions in order to show Sir Hgnis the way. They proceed in this fashion until at some point Sir Ffegroeg discovers that snow is covering the mountain pass and that it is impossible to get to their destination. So he correctly jettisons his intention to get there and therefore also the dependent intention to show Sir Hgnis the way. Instead he makes his way to a comfortable hotel he knows in a nearby town; Sir Hgnis is left to wander the unfamiliar land to his last day.

Clearly, the interaction between the intentions of the individual agents and the collective mental state is involved.

One way to approach the problem is to mimic the development in the single-agent case. Specifically, one can first define the notion of *joint persistent goal* and then use it to define the notion of a *joint intention*. We give an informal outline of such definitions.

weak achievement goal

Definition 14.4.7 (Weak achievement goal; informal) *An agent has a* weak achievement goal *with respect to a team of agents iff one of the following is true:*

1. *The agent has a standard achievement goal (AGoal) to bring about φ;*
2. *The agent believes that φ is true or will never be true, but has an achievement goal that the status of φ be common belief within the team.*

joint persistent goal

Definition 14.4.8 (Joint persistent goal; informal) *A team of agents has a* joint persistent goal *to achieve φ iff the following are all true:*

1. *They have common belief that φ is currently false;*
2. *They have common (and true) belief that they each have the goal that φ hold eventually;*
3. *They have common (and true) belief that, until they come to have common belief that φ is either true or never will be true, they will each continue to have φ as a weak achievement goal.*

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

joint intention

Definition 14.4.9 (Joint intention; informal) *A team of agents has a* joint intention *to take a set of actions A (each action by one of the agents) iff the agents have a joint persistent goal of (a) having done A and (b) having common belief throughout the execution of A that they are doing A.*

In addition to avoiding the pitfalls demonstrated in the earlier stories, these definitions have several potentially attractive properties, including the following.

- The joint persistent goals of a team consisting of a single agent coincide with the intentions of that agent.
- If a team has φ as a joint persistent goal then so does every individual agent.
- If a team has a joint intention to take a set of actions A, and action $a \in A$ belongs to agent i, then agent i intends to take action a.

However, these definitions are nothing if not complex, which is why we omit their formal details here. Furthermore, as was said, there is not yet universal agreement on the best definitions. The references point to the reader to further reading on this fascinating, yet incomplete, body of work.

14.5 History and references

The combination of knowledge and probability as discussed in this chapter is based on Fagin and Halpern [1994], and is covered also in Halpern [2005], which remains a good technical introduction to the topic.

Theories of belief dynamics—revision, update, and beyond—are covered in Pappas [2007], as well as in the older but still excellent Gärdenfors and Rott [1995]. Early seminal work includes that of Gärdenfors [1988], Alchourron, Gärdenfors and Makinson [1985] (after whom “AGM revision” is named), and Katsuno and Mendelzon [1991], who introduced the distinction between belief revision and belief update. The material in the chapter on knowledge, certainty, and belief is based on Boutilier [1992] and Lamarre and Shoham [1994]. The material in the section on belief fusion is based on Maynard-Reid and Shoham [2001] and Maynard-Zhang and Lehmann [2003]. A broader introduction to logical theories of belief fusion can be found in Grégoire and Konieczny [2006]. Belief dynamics are closely related to the topic of *nonmonotonic logics*, a rich source of material. An early comprehensive coverage of nonmonotonic logics can be found in Ginsberg [1987], and a more recent survey is provided in Brewka et al. [2007].

nonmonotonic logic

The recent trend toward modal logics of information dynamics is heralded in van Benthem [1997]. The early seminal work with a game-like perspective on logic is due to Peirce [1965]. A recent review of modal logic for games and information change appears in van der Hoek and Pauly [2006]. Our discussion of coalition logic is covered there in detail, and originally appeared in Pauly [2002].

In general, belief dynamics still constitute an active area of research, and the interested reader will undoubtedly want to study the recent literature, primarily in artificial intelligence and philosophical logic.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015

The literature on formal theories of "motivational" attitudes, such as desires, goals, and intentions, is much sparser than the literature on the "informational" attitudes discussed earlier. There are fewer publications on these topics, and the results are still preliminary (which is reflected in the style and length of the book section). The material presented is based largely on work in artificial intelligence by Cohen and Levesque [Cohen and Levesque, 1990, 1991], which in turn was inspired by philosophical work such as that of Bratman [1987]. The Cohen-Levesque formulation has attracted criticism and alternative formulations, some of which can be found in [Rao and Georgeff, 1991, 1998], Singh [1992], Meyer et al. [1999], and van der Hoek and Wooldridge [2003]. Much of this literature is surveyed in Wooldridge [2000]. This area presents many opportunities for further research.

Downloaded from https://www.cambridge.org/core. University of Warwick, on 10 Jan 2019 at 08:28:47, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511811654.015