Chapter 10

Belief Revision

Suppose that an agent believes $\varphi_1, \ldots, \varphi_n$ and then learns or observes ψ. How should she revise her beliefs? If ψ is consistent with $\varphi_1 \wedge \ldots \wedge \varphi_n$, then it seems reasonable for her to just add ψ to her stock of beliefs. This is just the situation considered in Section 4.1. But what if ψ is, say, $\neg\varphi_1$? It does not seem reasonable to just add $\neg\varphi_1$ to her stock of beliefs, for then her beliefs become inconsistent. Nor is it just a simple matter of discarding φ_1 and adding $\neg\varphi_1$. Discarding φ_1 may not be enough, for (at least) two reasons:

1. Suppose that φ_1 is $\varphi_2 \wedge \varphi_3$. If the agent's beliefs are closed under implication (as I will be assuming they are), then both φ_2 and φ_3 must be in her stock of beliefs. Discarding φ_1 and adding $\neg\varphi_1$ still leaves an inconsistent set. At least one of φ_2 or φ_3 will also have to be discarded to regain consistency, but which one?

2. Even if the result of discarding φ_1 and adding $\neg\varphi_1$ is consistent, it may not be an appropriate belief set. For example, suppose that φ_4 is $\varphi_1 \vee p$. Since φ_4 is a logical consequence of φ_1, it seems reasonable to assume that φ_4 is in the agent's belief set (before learning $\neg\varphi_1$). But suppose that the only reason that the agent believed φ_4 originally was that she believed φ_1. By discarding φ_1, we have removed the justification for φ_4. Shouldn't it be removed too?

Intuitively, it seems reasonable to insist that the agent revise her beliefs in light of learning/observing ψ by making the "minimal change" necessary to incorporate ψ. If ψ is consistent with $\varphi_1, \ldots, \varphi_n$ then the "minimal change" is clearly to just add ψ to the stock of beliefs and (perhaps all the logical consequences of $\psi \wedge \varphi_1 \wedge \ldots \wedge \varphi_n$ that were not there already). But in general, "minimal change" is a tricky notion, and seems to involve issues such as justification.

We have seen a similar intuition before, back in Section 4.10. There we were looking for the probability distribution that was "closest" to the original measure that satisfied certain constraints. Here we are essentially looking for the beliefs that are "closest" to the original beliefs but satisfy the constraint of including the new observation. The properties (4.1) and (4.2) can be viewed as trying to characterize the conditional probability measure $\mu|U$ as being the probability measure closest to the prior probability measure μ that gives U probability 1. In fact, in Section 4.10, I said that, for various reasonable notions of closeness, this was indeed the case.

As we saw in Chapter 7, probability is not so useful as a model of belief; however, plausibility can be used (where the agent is said to believe φ if the plausibility of φ is greater than that of $\neg\varphi$; see also Section 8.3). In this chapter, using plausibility to model belief, I show that many conditional plausibility provides a useful model for belief revision.

10.1 The Circuit-Diagnosis Problem

The circuit-diagnosis problem provides a good testbed for understanding the issues involved in belief revision, so I start with that here.

A *circuit* consists of a number of components (AND, OR, NOT, and XOR gates) and lines. For example, the circuit of Figure 10.1 contains 5 components, X_1, X_2, A_1, A_2, O_1 and 8 lines, $l_1, \ldots, l_8$. Inputs (which are either 0 or 1) come in along lines l_1, l_2, and l_3. A_1 and A_2 are AND gates; the output of an AND gate is 1 if both of its inputs are 1, otherwise it is 0. O_1 is an OR gate; its output is 1 if either of its inputs is 1, otherwise it is 0. Finally, X_1 and X_2 are XOR gates; the output of a XOR gate is 1 iff exactly one of its inputs is 1.

The *circuit-diagnosis problem* is that of identifying which components in a circuit are faulty. An agent is given a circuit diagram as in Figure 10.1; she can set the values of input lines of the circuit and observe the output values. By comparing the actual output values with the expected output values, the agent can attempt to locate faulty components.

The agent's knowledge about a circuit can be modeled using a Kripke structure $M_{diag}^K = (W_{diag}, \mathcal{K}_{diag}, \pi_{diag})$. Each possible world $w \in W_{diag}$ is composed of two parts: *fault*(w), the *failure set*, that is, the set of faulty components in w, and *value*(w), the value of all the lines in the circuit. Formally, *value*(w) is a set of pairs of the form (l, i), where l is a line in the circuit and i is either 0 or 1. Components that are not in the failure sets perform as expected. Thus, for the circuit in Figure 10.1, if $w \in W_{diag}$

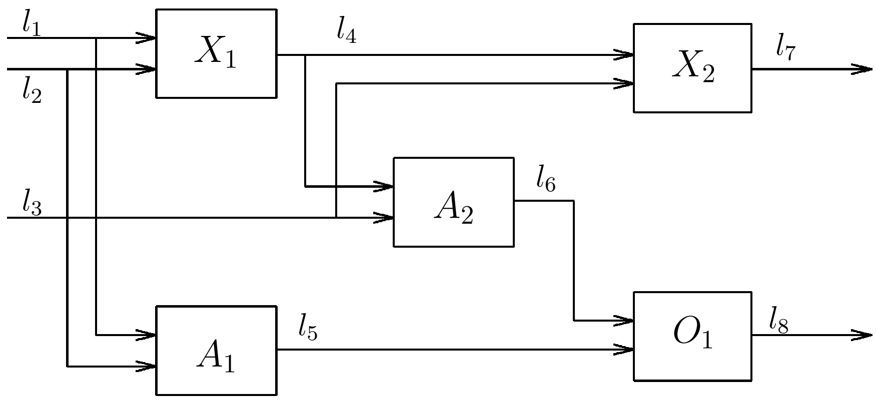

Figure 10.1: A typical circuit.

and $A_1 \notin \mathit{fault}(w)$, then $(l_5, 1)$ is in $\mathit{value}(w)$ if and only if both $(l_1, 1)$ and $(l_2, 1)$ are in $\mathit{value}(w)$.

What language should we use to reason about faults in circuits? Since we need to talk about which components are faulty and the values of various lines, it seems reasonable to take $\Phi_{diag} = \{\mathit{faulty}(c_1), \ldots, \mathit{faulty}(c_n), \mathit{hi}(l_1), \ldots, \mathit{hi}(l_k)\}$, where $\mathit{faulty}(c_i)$ denotes that component i is faulty and $\mathit{hi}(l_i)$ denotes that line i in a "high" state (i.e., has value 1). Define the interpretation π_{diag} in the obvious way: $\pi_{diag}(w)(\mathit{faulty}(c_i)) = \mathbf{true}$ if $c_i \in \mathit{fault}(w)$, and $\pi_{diag}(w)(\mathit{hi}(l_i)) = \mathbf{true}$ if $(l_i, 1) \in \mathit{value}(w)$.

The agent knows which tests she has performed and the results she observed. Let $\mathit{obs}(w) \subseteq \mathit{value}(w)$ consist of the values of those lines that the agent sets or observes. (For the purposes of this discussion, I assume that the agent sets the value of a line only once.) Thus, $(w, w') \in \mathcal{K}_{diag}$ if $\mathit{obs}(w) = \mathit{obs}(w')$. For example, suppose that the agent observes $\mathit{hi}(l_1) \wedge \mathit{hi}(l_2) \wedge \mathit{hi}(l_3) \wedge \mathit{hi}(l_7) \wedge \mathit{hi}(l_8)$. The agent then considers possible all worlds where lines l_1, l_2, l_3, l_7 and l_8 have value 1. Since these observations are consistent with the circuit being correct, one of these worlds has an empty failure set. However, other worlds are possible. For example, it might be that the AND gate A_2 is faulty. This would not affect the outputs in this case, since if A_1 is nonfaulty, then its output is "high", and thus, O_1's output is "high" regardless of A_2's output.

Now suppose that the agent observes $\mathit{hi}(l_1) \wedge \neg\mathit{hi}(l_2) \wedge \mathit{hi}(l_3) \wedge \mathit{hi}(l_7) \wedge$

$\neg hi(l_8)$. These observations imply that the circuit is faulty. (If l_1 and l_3 are "high" and l_2 is "low", then the correct values for l_7 and l_8 should be "low" and "high", respectively.) In this case there are several failure sets consistent with the observations, including $\{X_1\}$, $\{X_2, O_1\}$, and $\{X_2, A_2\}$.

In general, there is more than one explanation for the observed faulty behavior. Thus, the agent cannot *know* exactly which components are faulty, but she may have *beliefs* on that score. To model these beliefs, we must decide on the plausibility measure the agent has at any world. Assume for simplicity that a set's plausibility is determined by the failure sets at the worlds in the set. To construct a plausibility measure with this property, it seems reasonable to start by constructing a plausibility measure over possible failures of the circuit. I actually construct two plausibility measures over failures, each capturing slightly different assumptions. Both plausibility measures embody the assumptions that failures are unlikely and failures of individual components are independent of one another. It follows that the failure of two components is much more unlikely than the failure of any one of them. The plausibility measures differ in what they assume about the relative likelihood of the failure of different components.

The first plausibility measure embodies the assumption that the likelihood of each component failing is the same. This leads to an obvious ordering on failure sets: If f_1 and f_2 are two failure sets, then $f_1 \succ_1 f_2$ if $|f_1| < |f_2|$, that is, if f_1 consists of fewer faulty components than f_2. This leads to an ordering on worlds: $w_1 \succ_1 w_2$ if $\mathit{fault}(w_1) \succ_1 \mathit{fault}(w_2)$. Using the construction of Section 2.3, this gives a total order $\succ_1^s$ on sets of worlds. Moreover, by Proposition 2.3.4, $\succ_1^s$ can be viewed as a qualitative plausibility measure. Call this plausibility measure Pl_1.

Pl_1 can also be constructed by using probability sequences. Let μ_m be the probability measure that takes the probability of a component failing to be $1/m$ and takes component failures to be independent. Then for a circuit with n components,

$$\mu_m(w) = \left(\frac{1}{m}\right)^{|\mathit{fault}(w)|} \left(\frac{m-1}{m}\right)^{n-|\mathit{fault}(w)|}.$$

Then it is easy to check that Pl_1 is just the plausibility measure obtained from the probability sequence $(\mu_1, \mu_2, \mu_3, \ldots)$ using the construction preceding Theorem 7.2.11 (Exercise 10.1(a)). Note that as we go out further in the sequence, the probability of a component being faulty becomes smaller and smaller. However, at each measure in the sequence, each component is equally likely to fail and the failures are independent.

In some situations it might be unreasonable to assume that all components have equal failure probability. Moreover, the relative probability of failure for various components might be unknown. Without assumptions on failure probabilities, it is not possible to compare failure sets unless one is a subset of the other. This intuition leads to a different ordering on failure sets: define $f \succ_2 f'$ if $f \supset f'$. Again this leads to on ordering worlds by taking $w_1 \succ_2 w_2$ if $\mathit{fault}(w_1) \succ_2 \mathit{fault}(w_2)$ and, again, the construction of Section 2.3 gives us a plausibility measure Pl_2 on W_{diag}. It is not hard to find a probability sequence that gives the same plausibility measure (Exercise 10.1(b)).

Pl_1 and Pl_2 determine structures $M_{diag,1}$ and $M_{diag,2}$, respectively, for knowledge and plausibility: $M_{diag,i} = (W_{diag}, \mathcal{K}_{diag}, \mathcal{PL}_{diag,i}, \pi_{diag})$, where $\mathcal{PL}_{diag,i}(w,1) = (\mathcal{K}_{diag}(w), \mathrm{Pl}^i_{w,1})$ and $\mathrm{Pl}^i_{w,1}(U) = \mathrm{Pl}_i(\mathcal{K}_{diag}(w) \cap U)$, for $i = 1, 2$.

Suppose that the agent makes some observations o. In both $M_{diag,1}$ and $M_{diag,2}$, if there is a world w compatible with the observations o and $\mathit{fault}(w) = \emptyset$, then the agent believes that the circuit is fault-free. That is, the agent believes the circuit is fault-free as long as her observations are compatible with this hypothesis. If not, then the agent looks for a *minimal explanation* of her observations, where the notion of minimality differs in the two structures. More precisely, if f is a failure set, let $\mathrm{Diag}(f)$ be the formula that says that precisely the failures in f occur, so that $(M, w) \models \mathrm{Diag}(f)$ if and only if $\mathit{fault}(w) = f$. For example, if $f = \{c_1, c_2\}$, then $\mathrm{Diag}(f) = \mathit{faulty}(c_1) \wedge \mathit{faulty}(c_2) \wedge \neg \mathit{faulty}(c_3) \wedge \ldots \wedge \neg \mathit{faulty}(c_n)$. The agent believes that f is a *possible diagnosis* (i.e., an explanation of her observations) in world w of structure $M_{diag,i}$ if $(M_{diag,i}, w) \models \neg B \neg \mathrm{Diag}(f)$. The set of diagnoses the agent considers possible is $\mathrm{DIAG}(M, w) = \{f : (M, w) \models \neg B \neg \mathrm{Diag}(f)\}$. A failure set f is *consistent* with an observation o if it is possible to observe o when f occurs, i.e., if there is a world w in W such that $\mathit{fault}(w) = f$ and $obs(w) = o$.

Proposition 10.1.1

(a) DIAG$(M_{diag,1}, w)$ contains all failure sets f that are consistent with $obs(w)$ such that there is no failure set f' with $|f'| < |f|$ that is consistent with $obs(w)$.

(b) DIAG$(M_{diag,2}, w)$ contains all failure sets f that are consistent with $obs(w)$ such that there is no failure set f' with $f' \subset f$ that is consistent with $obs(w)$.

Proof See Exercise 10.2. ∎

Thus, both $\mathrm{DIAG}(M_{diag,1}, w)$ and $\mathrm{DIAG}(M_{diag,2}, w)$ consist of minimal sets of failure sets consistent with $obs(w)$, for different notions of minimality.

In the case of $M_{diag,1}$, "minimality" means "of minimal cardinality", while in the case of $M_{diag,2}$, it means "minimal in terms of set containment". More concretely, in the circuit of Figure 10.1, if the agent observes $hi(l_1) \wedge \neg hi(l_2) \wedge hi(l_3) \wedge hi(l_7) \wedge \neg hi(l_8)$, then in $M_{diag,1}$ she would believe that X_1 is faulty, since $\{X_1\}$ is the only diagnosis with cardinality one. On the other hand, in $M_{diag,2}$ she would believe that one of the three minimal diagnoses occurred: $\{X_1\}$, $\{X_2, O_1\}$ or $\{X_2, A_2\}$.

The structures $M_{diag,i}$, $i = 1, 2$, model a static situation. They describe the agent's beliefs given some observations, but do not describe the *process* of belief revision—how those beliefs change in the light of new observations. One way to model the process is to add time to the picture, and model the agent and the circuit as part of an interpreted plausibility system. This can be done by a straightforward modification of what was done in the static case.

The first step is to describe the agent's set of local states and the set of environment states. In the spirit of the static model, I assume that the agent sets the value of some lines in the circuit and observes the value of others. Let $o_{(r,m)}$ be a description of what the agent has set/observed in round m of run r, where $o_{(r,m)}$ is a a conjunction of formulas of the form $hi(l_j)$ and their negations. To model the agent's local states, we need to ask the same questions as in the Listener-Teller protocol of Section 9.6.1. Does the agent remember her observations? If not, what does she remember of them? For simplicity here, I assume that the agent remembers all her observations, and makes an observation at each round. Given these assumptions, it seems reasonable to model the agent's local state at a point (r, m) as the sequence $\langle o_{(r,1)}, \ldots, o_{(r,m)} \rangle$. Thus, the agent's initial state at $(r, 0)$ is $\langle\,\rangle$, since she has not made any observations; after each round in r, a new observation is added.

The environment states play the role of the worlds in the static models; they describe the faulty components of the circuit and the values of all the lines. Thus, I assume that the environment's state at (r, m) is a pair $(\mathit{fault}(r, m), \mathit{value}(r, m))$, where $\mathit{fault}(r, m)$ describes the failure set at the point (r, m) and $\mathit{value}(r, m)$ describes the values of the lines at (r, m). Of course, $o_{(r,m)}$ must be compatible with $\mathit{value}(r, m)$—the values of the lines that the agent observes/sets at (r, m) must be the actual values. (Intuitively, this says that the agents observations are correct and when the agent sets a line's value, it actually has that value.) Moreover, $\mathit{fault}(r, m)$ must be compatible with $\mathit{value}(r, m)$, in the sense discussed earlier: if a component c is not in $\mathit{fault}(r, m)$, then it outputs values according to its specification, while if c is in $\mathit{fault}(r, m)$, then it exhibits its faultiness by not obeying its specification for all inputs. I further assume that the set of faulty components does not change over time; this is captured by assuming

$\mathit{fault}(r, m) = \mathit{fault}(r, 0)$ for all r and m. On the other hand, I do not assume that the values on the lines are constant over time since, by assumption, the agent can set certain values. Let $\mathcal{R}_{diag}$ consist of all runs r satisfying these requirements.

There are obvious analogues to Pl_1 and Pl_2 defined on runs; I abuse notation and continue to call these Pl_1 and Pl_2. For example, to get Pl_1, first define a total order $\succ_1$ on the runs in $\mathcal{R}_{diag}$ by taking $r_1 \succ_1 r_2$ if $\mathit{fault}(r_1(0)) \succ_1 \mathit{fault}(r_2(0))$; the construction of Section 2.3 then gives a total order on sets of runs, which can be viewed as a plausibility measure on runs. Similarly, the plausibility measure Pl_2 on $\mathcal{R}_{diag}$ is the obvious analogue to Pl_2 defined earlier on W_{diag}.

Pl_1 and Pl_2 determine two interpreted plausibility system satisfying PRIOR whose set of runs is $\mathcal{R}_{diag}$; call them $\mathcal{I}_{diag,1}$ and $\mathcal{I}_{diag,2}$. In each of these systems, $\Omega_{(r,m,1)} = \mathcal{K}_i(r, m)$. Thus, at (r, m), the agent considers possible all the points where she performed the same tests up to time m and observed the same results. As before, the agent believes that the failure set is one of the ones that provides a minimal explanation for her observations, where the notion of minimal depends on the plausibility measure. As the agent performs more tests, her knowledge increases and her beliefs might change.

Let $\mathrm{DIAG}(\mathcal{I}, r, m)$ be the set of failure sets (i.e., diagnoses) that the agent considers possible at the point (r, m) in the system $\mathcal{I}$. Belief change in $\mathcal{I}_{diag,1}$ is characterized by the following proposition, similar in spirit to Proposition 10.1.1.

Proposition 10.1.2 *If there is some $f \in DIAG(\mathcal{I}_{diag,1}, r, m)$ that is consistent with the new observation $o_{(r,m+1)}$, then $DIAG(\mathcal{I}_{diag,1}, r, m+1)$ consists of all the failure sets in $DIAG(\mathcal{I}_{diag,1}, r, m)$ that are consistent with $o_{(r,m+1)}$. If all $f \in \mathrm{Bel}(\mathcal{I}_{diag,1}, r, m)$ are inconsistent with $o_{(r,m+1)}$, then $\mathrm{Bel}(\mathcal{I}_{diag,1}, r, m+1)$ consists of all failure sets of cardinality j that are consistent with $o_{(r,m+1)}$, where j is the least cardinality for which there is at least one failure set consistent with $o_{(r,m+1)}$.*

Proof See Exercise 10.3. ∎

Thus, in $\mathcal{I}_{diag,1}$, if an observation is consistent with the pre-observation set of most likely explanations, then the post-observation set of most likely explanations is a subset of the pre-observation set of most likely explanations (the subset consisting of those explanations that are consistent with the new observation). On the other hand, a surprising observation (one inconsistent with the current set of most likely explanations) has a rather drastic effect. It easily follows from Proposition 10.1.2 that if $o_{(r,m+1)}$ is surprising, then $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m) \cap \mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m+1) = \emptyset$, so the agent discards all her pre-observation explanations. Moreover, an easy induction on m shows that if $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m) \cap \mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m+1) = \emptyset$, then the cardinality of the failure sets in $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m+1)$ is greater than the cardinality of the failure sets in $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m)$. Thus, in this case, the explanations in $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m+1)$ are more complicated than those in $\mathrm{Bel}(\mathcal{I}_{diag,1}, r, m)$. Notice that the fact that the agent considers $o_{(r,m+1)}$ surprising can be expressed by the formula $B\neg \bigcirc o_{(r,m+1)}$. That is, $o_{(r,m+1)}$ is a surprising or unexpected observation at (r, m) iff $(\mathcal{I}_{diag,1}, r, m) \models B_i \neg \bigcirc o_{(r,m)}$.

Belief change in $\mathcal{I}_{diag,2}$ is quite different, as the following proposition shows. Roughly speaking, it says that after making an observation, the agent believes possible all minimal extensions of the diagnoses she believed possible before making the observation that are consistent with the observation.

Proposition 10.1.3 *$DIAG(\mathcal{I}_{diag,2}, r, m+1)$ consists of the minimal (according to $\subseteq$) failure sets in $\{f' : f' \supseteq f$ for some $f \in DIAG(\mathcal{I}_{diag,2}, r, m)\}$ that are consistent with $o_{(r,m+1)}$.*

Proof See Exercise 10.4. ▮

As with $\mathcal{I}_{diag,1}$, diagnoses that are consistent with the new observation are retained. However, unlike $\mathcal{I}_{diag,1}$, diagnoses that are discarded are replaced by more complicated diagnoses even if some of the diagnoses considered at (r, m) are consistent with the new observation. Moreover, while new diagnoses in $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m+1)$ can be unrelated to the diagnoses in $\mathrm{DIAG}(\mathcal{I}_{diag,1}, r, m)$, in $\mathcal{I}_{diag,2}$ the new diagnoses must be extensions of some discarded diagnoses. Thus, in $\mathcal{I}_{diag,1}$ the agent does not consider new diagnoses as long as the observation is not surprising. On the other hand, in $\mathcal{I}_{diag,2}$ the agent has to examine new candidates after each test.

This point is perhaps best understood by example. Suppose that in the circuit of Figure 10.1, the agent initially sets $l_1 = 1$ and $l_2 = l_3 = 0$. If there were no failures, then l_4 and l_7 would be 1, while, l_5, l_6, and l_8 would be 0. However, the agent observes that l_8 is 1. In that case, in both systems, the agent would believe that exactly one of X_1, A_1, A_2, or O_1 was faulty—that would be the minimal explanation of the problem, under both notions of minimality. However, suppose that the agent then observes that $l_7 = 0$ while all the other settings remain the same. In that case, the only diagnosis according to Pl_1 is that X_1 is faulty. This is also a minimal explanation according to Pl_2, but there are three other possible diagnoses: X_2 and one of A_1, A_2, or O_1 could be faulty. Thus, even though an explanation considered most likely after the first observation—that X_1 is faulty—is consistent with the second observation, some new diagnoses (all extensions of the diagnoses considered after the first observation) are also considered.

10.2 Belief-Change Systems

I now abstract the discussion of the previous section and consider belief change more generally in the context of interpreted plausibility systems. To do so, I consider a particular class of systems called *belief change systems*. In belief change systems, the agent makes observations about an external environment. For simplicity, as in the analysis of circuit-diagnosis problem, I assume that these observations are described by formulas in some logical language. I then make other assumptions regarding the plausibility measure used by the agent. Among other things, these assumptions make precise that belief change proceeds by conditioning. The assumptions are formalized by conditions BCS1–BCS4, described below. A system $\mathcal{I} = (\mathcal{R}, \pi, \mathcal{P})$ is a *belief-change system (BCS)* if it satisfies these conditions.

Assumption BCS1 formalizes the intuition that the language includes propositions for reasoning about the environment, whose truth depends only on the environment state.

BCS1. There is a subset Φ_e of the set Φ of primitive propositions whose truth depends only on the environment state; that is, for each primitive proposition $p \in \Phi_e$, $\pi(r,m)(p) = \mathbf{true}$ iff $\pi(r',m') = \mathbf{true}$ for all points (r',m') such that $r_e(m) = r'_e(m')$.

BCS1 certainly holds for the interpretations used to capture the circuit-diagnosis problem: $\Phi_e = \Phi_{diag}$.

BCS2 is concerned with the form of the agent's local state. Intuitively, the agent's local state is supposed to encode the information available to the agent. Thus, the agent's local state should be a function of her initial state and her observations. BCS2 makes a stronger assumption. Just as in the structures used for the circuit-diagnosis problem, it asserts that the agent's local state is just the sequence of observations made. This means that the agent remembers all her past observations (so that the agent has perfect recall, in the sense of Section 9.4) and that she has no information at time 0. There is an additional, quite nontrivial, condition imposed by BCS2 (that was also assumed in the analysis of circuit diagnosis): that the agent's observations can be described by formulas in $\mathcal{L}_e$. This says that $\mathcal{L}_e$ may have to be quite an expressive language. In the case of an agent observing a circuit, perhaps all that can be observed is the value of various lines. However, in the case of agents observing people, the observations can include obvious features eye color and skin color (but even getting a language rich enough to describe all the gradations of eye and skin color is nontrivial) as well as more subtle features like facial expressions. $\mathcal{L}_e$ must be expressive enough to describe whatever can be observed.

In BCS2 and for the remainder of this chapter, I use r_a to denote the agent's local state rather than r_1, to stress that we are dealing with a single agent.

BCS2. For all $r \in R$ and for all m, the agent's local state $r_a(m) = \langle o_{(r,1)}, \ldots, o_{(r,m)} \rangle$ where $o_{(r,k)} \in \mathcal{L}_e$ for $1 \leq k \leq m$.

Clearly we want to reason in the language about the observations the agent makes. Thus, we assume that Φ includes propositions that describe the observations made by the agent.

BCS3. The set Φ includes a set Φ_{obs} of primitive propositions disjoint from Φ_e such that $\Phi_{obs} = \{learn(\varphi) : \varphi \in \mathcal{L}_e\}$. Moreover, $\pi(r,m)(learn(\varphi)) =$ **true** if and only if $o_{(r,m)} = \varphi$ for all runs r and times m.

Finally, BCS4 asserts that belief change proceeds by conditioning. While there are certainly other assumptions that can be made, conditioning is a principled approach that captures the intuitions of minimal change, given the observations. For ease of exposition, I make a somewhat stronger assumption; not only does the system satisfy PRIOR, but it is a standard SDP plausibilistic interpreted system. That is, the agent has a single plausibility measure on all of $\mathcal{R}$, and it is the same plausibility measure in all runs.

BCS4. $\mathcal{I}$ is a standard SDP system.

As we observed (Exercise 9.4), in a standard SDP system, the agent's plauibility assignment at each point satisfies the SDP property. It follows from Proposition 8.3.1(b) that the agent's beliefs depend only on the agent's local state. I use the notation $(\mathcal{I}, s_a) \models B\varphi$ as shorthand for $(\mathcal{I}, r, m) \models B\varphi$ for some (and hence for all) (r, m) such that $r_a(m) = s_a$. The agent's *belief set* at s_a is

$$\mathrm{Bel}(\mathcal{I}, s_a) = \{\varphi \in \Phi_e : (\mathcal{I}, s_a) \models B\varphi\}.$$

Since the agent's state is a sequence of observations, the agent's state after observing φ is simply $s_a \cdot \varphi$, where $\cdot$ is the append operation. Thus, $\mathrm{Bel}(\mathcal{I}, s_a \cdot \varphi)$ is the belief set after observing φ. We adopt the convention that if there is no point where the agent has local state s_a in system $\mathcal{I}$, then $\mathrm{Bel}(\mathcal{I}, s_a)$ consists of all the propositional formulas over Φ_e. With these definitions, we can compare the agent's belief set before and after observing φ, that is $\mathrm{Bel}(\mathcal{I}, s_a)$ and $\mathrm{Bel}(\mathcal{I}, s_a \cdot \varphi)$. Thus, in a BCS, we can conveniently talk about belief change. The agent's state encodes observations and there are propositions that allow us to talk about what is observed and how the agents beliefs change over time.

There is one other requirement that is standard in many approaches to belief change considered in the literature: that observations are "accepted", so that after the agent observes φ, she believes φ. This assumption is enforced by the next assumption, BCS5, by assuming that observations are reliable, so that the agent observes φ only if the current state of the environment satisfies φ. This is certainly not the only way of enforcing the assumption that observations are accepted, but it is perhaps the simplest.

BCS5. $(\mathcal{I}, r, m) \models o_{(r,m)}$ for all runs r and times m.

Note that BCS5 implies that the agent never observes *false*. Moreover, it implies that after observing φ, the agent knows that φ is true. A system that satisfies BCS1–5 is said to be a *reliable BCS*.

Example 10.2.1 As they stand, the systems $\mathcal{I}_{diag,1}$ and $\mathcal{I}_{diag,2}$ are not quite BCSs, since π_{diag} is not defined on primitive propositions of the form *learn*(φ). This can easily be rectified. Let Φ^+_{diag} consist of Φ_{diag} together with all the primitive propositions of the form *learn*(φ) for $\varphi \in \mathcal{L}^{Prop}(\Phi_{diag})$. Let π^+_{diag} be the obvious extension of π_{diag} to Φ^+_{diag}, defined so that BCS3 holds. Let $\mathcal{I}^+_{diag,1}$ and $\mathcal{I}^+_{diag,2}$ be the systems that result when π_{diag} is replaced by π^+_{diag}. Clearly, both $\mathcal{I}^+_{diag,1}$ and $\mathcal{I}^+_{diag,2}$ are reliable BCSs. ■

10.3 Belief Revision

The most common approach to studying belief change in the literature has been the axiomatic approach: this has typically involved starting with a collection of postulates, arguing that they are reasonable, and proving some consequences of these postulates. And perhaps the most-studied postulates are the *AGM postulates*, named after the researchers who introduced them, Alchourrón, Gärdenfors, and Makinson. These axioms are intended to characterize a particular type of belief change, called *belief revision*.

The AGM approach assumes that an agent's epistemic state is represented by a belief set, that is, a set K of formulas in a logical language $\mathcal{L}$, describing the subject matter about which the agent holds beliefs. For simplicity here, I assume that $\mathcal{L}$ is propositional (which is consistent with most of the discussions of the postulates). In the background, there are also assumed to be some axioms $\mathrm{AX}_{\mathcal{L}}$ characterizing the situation. For example, for the circuit-diagnosis example of Figure 10.1, $\mathcal{L}$ could be $\mathcal{L}^{Prop}(\Phi_{diag})$.

There would then be an axiom in $\mathrm{AX}_{\mathcal{L}}$ saying that if A_1 is not faulty, then l_5 is 1 if and only iff both l_1 and l_2 are:

$$\neg \mathit{faulty}(A_1) \Rightarrow (\mathit{hi}(l_5) \Leftrightarrow (\mathit{hi}(l_1) \wedge \mathit{hi}(l_2)).$$

Similar axioms would be used to characterize all the other components.

I assume that there is a *consequence relation* $\vdash_{\mathcal{L}}$ such $\Sigma \vdash_{\mathcal{L}} \varphi$ holds iff φ is provable from Σ and the axioms in $\mathrm{AX}_{\mathcal{L}}$, using standard propositional reasoning (Prop and MP). $Cl(\Sigma)$ denotes the logical closure of the set Σ under $\mathrm{AX}_{\mathcal{L}}$; that is, $Cl(\Sigma) = \{\varphi : \Sigma \vdash_{\mathcal{L}} \varphi\}$. I assume for simplicity that belief sets are closed under logical consequence, so that if K is a belief set, then $Cl(K) = K$. This assumption, which is standard in all the belief change literature, essentially says that agents are being treated as perfect reasoners, and can compute all logical consequences of their beliefs.

What the agent learns is assumed to be characterized by some formula φ, also in $\mathcal{L}$; $K * \varphi$ describes the belief set of an agent who starts with belief set K and learns φ. There are two subtle but important assumptions implicit in this notation:

- The functional form of $*$ suggests that all that matters regarding how an agent revises her beliefs is the belief set and what is learnt. In any two situations where the agent has the same beliefs, she will revise her beliefs in the same way.
- The notation also suggests that the second argument of $*$ can be an arbitrary formula in $\mathcal{L}$.

These are nontrivial assumptions. With regard to the first one, it is quite possible for two different plausibility measures to result in the same belief sets and yet behave differently under conditioning, leading to different belief sets after revision. With regard to the second one, at a minimum, it is not clear what it would mean to observe *false*. (It is perfectly reasonable to observe something inconsistent with one's current beliefs, but that is quite different from observing *false*, which is an inconsistent formula.) But even putting this issue aside, it may not be desirable to allow every consistent formula to be observed in every circumstance. For example, in the circuit-diagnosis problem, the agent does not observe the behavior of a component directly; she can only infer it by setting the values of some lines and observing the values of others. While there are some observations that are essentially equivalent to observing that a particular component is faulty (for example, if setting setting l_1 to 1 and l_2 to 1 results in l_5 being 0 in the circuit of Figure 10.1, then A_1 must be faulty), there are no observations that can definitively rule out a component being faulty (the faulty behavior may display itself only sporadically).

Indeed, in general, what is observable may depend on the belief set itself. Consider a situation where an agent can reliably observe colors. After observing that a coat is blue (and thus, having this fact in her belief set), it would not be possible for her to observe that the same coat is red.

The impact of these assumptions will be apparent shortly. For now, I simply state the eight postulates used by Alchourrón, Gärdenfors, and Makinson to characterize belief revision:

R1. $K \circ \varphi$ is a belief set.

R2. $\varphi \in K \circ \varphi$.

R3. $K \circ \varphi \subseteq Cl(K \cup \{\varphi\})$.

R4. If $\neg\varphi \notin K$ then $Cl(K \cup \{\varphi\}) \subseteq K \circ \varphi$.

R5. $K \circ \varphi = Cl(\mathit{false})$ if and only if $\vdash_{\mathcal{L}} \neg\varphi$.

R6. If $\vdash_{\mathcal{L}} \varphi \Leftrightarrow \psi$ then $K \circ \varphi = K \circ \psi$.

R7. $K \circ (\varphi \wedge \psi) \subseteq Cl(K \circ \varphi \cup \{\psi\})$.

R8. If $\neg\psi \notin K \circ \varphi$ then $Cl(K \circ \varphi \cup \{\psi\}) \subseteq K \circ (\varphi \wedge \psi)$.

The essence of these postulates is the following. Revision by φ results in a belief set (postulate R1) that includes φ (R2). If the new belief is consistent with the belief set, then the revision should not remove any of the old beliefs nor add any new beliefs except these implied by the combination of the old beliefs with the new belief (R3 and R4). This condition is called *persistence*, and essentially characterizes conditioning in the simple case discussed in Section 4.1. The next two conditions discuss the coherence of beliefs. Postulate R5 states that φ is consistent with the axioms iff $K \circ \varphi$ is a nontrivial belief set. ($Cl(\mathit{false})$ is the trivial belief set consisting of all formulas in $\mathcal{L}$, since all formulas are provable from *false*.) R6 states that the syntactic form of the new belief does not affect the revision process; it is much in the spirit of the rule LLE in system **P** from Section 7.1. The last two postulates enforce a certain coherency on the outcome of successive revisions. Basically, they state that if ψ is consistent with $A \circ \varphi$ then $K \circ (\varphi \wedge \psi)$ is just $(K \circ \varphi) \circ \psi$. (Recall that, by R3 and R4, if $\neg\psi \notin K \circ \varphi$, then $(K \circ \varphi) \circ \psi = Cl(K \circ \varphi \cup \{\psi\})$.) This is a property that we have seen before in the context of probabilistic conditioning: if $\mu(U_1 \cap U_2) \neq 0$, then $(\mu|U_1)|U_2 = (\mu|U_2)|U_1 = \mu|(U_1 \cap U_2)$.

My goal now is to relate AGM revision to BCSs. More precisely, the plan is to find some additional conditions (REV1–REV3 below) on BCSs that ensure that belief change in a BCS satisfies R1–R8. Doing this will help bring out the assumptions implicit in the AGM approach.

The first assumption is that, although the agent's beliefs may change, the propositions about which the agent has beliefs do not change during the revision process. The original motivation for belief revision came from the study of scientists' beliefs about laws of nature. These laws were taken to be unvarying, although experimental evidence might cause scientists to change their beliefs about the laws.

This assumption underlies R3 and R4. If φ is consistent with K, then according to R3 and R4, observing φ should result in the agent adding φ to her stock of beliefs and the closing off under implication. In particular, this means that all her old beliefs are retained. But if the world can change, then there is no reason for the agent to retain her old beliefs. Consider the systems $\mathcal{I}_{diag,1}$ and $\mathcal{I}_{diag,2}$ used to model the diagnosis problem. In these systems, the values on the line could change at each step. If $l_1 = 1$ before observing $l_2 = 1$, then why should $l_1 = 1$ after the observation, even if it is consistent with the observation that $l_2 = 1$? Perhaps if l_1 is not set to 1, its value goes to 0.

In any case, it is easy to capture the assumption that the propositions observed do not change their truth value—that is the role of REV1.

REV1. $\pi(r,m)(p) = \pi(r,0)(p)$ for all $p \in \Phi_e$ and points (r,m).

Note that REV1 does not say that all propositions are time-invariant, nor that the environment state does not change over time. It simply says that the propositions in Φ_e do not change their truth value over time.

In the BCSs $\mathcal{I}^+_{diag,1}$ and $\mathcal{I}^+_{diag,2}$, propositions of the form *faulty*(c) do not change their truth value over time, by assumption; however, propositions of the form *hi*(l) do. There is a slight modification of these systems that does satisfy REV1. The idea is to take $\mathcal{L}_e$ to consist only of Boolean combinations of formulas of the form *faulty*(c) and then convert the agent's observations to formulas in $\mathcal{L}_e$. Note that to every observation o made by the agent regarding the value of the lines, there corresponds a formula in $\mathcal{L}_e$ that characterizes all the fault sets that are consistent with o. For example, the observation $hi(l_1) \wedge hi(l_2) \wedge hi(l_4)$ corresponds to the conjunction of the formulas characterizing all fault sets that include X_1 (which is equivalent to the formula *faulty*(X_1)). For every observation φ about the value of lines, let $\varphi^\dagger \in \mathcal{L}_e$ be the corresponding observation regarding fault sets. Given a run $r \in \mathcal{I}^+_{diag,i}$, $i = 1, 2$, let $r^\dagger$ be the run where each observation φ is replaced by $\varphi^\dagger$. Let $\mathcal{I}^\dagger_{diag,i}$ be the BCS consisting of all the run $r^\dagger$ corresponding to the runs in $\mathcal{I}^+_{diag,i}$. The plausibility assignments in $\mathcal{I}^\dagger_{diag,i}$ and $\mathcal{I}^+_{diag,i}$ correspond in the obvious way. That means that the agent has the same beliefs about formulas in $\mathcal{L}_e$ at corresponding points in the two systems. More precisely, if $\varphi \in \mathcal{L}_e$, then $(\mathcal{I}^{\dagger}_{diag,i}, r^{\dagger}, m) \models \varphi$ if and only if $(\mathcal{I}^{+}_{diag,i}, r, m) \models \varphi$ for all points (r, m) in $\mathcal{I}_{diag,i}$, and hence $(\mathcal{I}^{\dagger}_{diag,i}, r^{\dagger}, m) \models B\varphi$ if and only if $(\mathcal{I}^{+}_{diag,i}, r, m) \models B\varphi$. By construction, $\mathcal{I}^{\dagger}_{diag,i}$, $i = 1, 2$, are BCSs that satisfy REV1.

Belief change in $\mathcal{I}^{\dagger}_{diag,1}$ can be shown to satisfy all of R1–8 in a precise sense (see Theorem 10.3.2 and Exercise 10.7); however, $\mathcal{I}^{\dagger}_{diag,2}$ does not satisfy R8. Consider the example discussed just after Proposition 10.1.3. Initially (before making any observations) that agent believes that no components are faulty. Let $K = \mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,2}, \langle\,\rangle)$. Then the agent sets $l_1 = 1$ and $l_2 = l_3 = 0$, and observes that l_8 is 1. That is, the agent observes $\varphi^{\dagger} = (hi(l_1) \wedge \neg hi(l_2) \wedge \neg hi(l_3))^{\dagger}$, which is equivalent to observing a fault set that contains at least one of X_1, A_1, A_2, or O_1. Since the agent prefers minimal explanations, $\mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,2}, \langle\varphi^{\dagger}\rangle)$ includes the belief that exactly one of X_1, A_1, A_2, or O_1 is faulty. Think of $\mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,2}, \langle\varphi^{\dagger}\rangle)$ as $K \circ \varphi^{\dagger}$. It is consistent with $K \circ \varphi^{\dagger}$ to observe that $l_7 = 0$ in addition to all the other observations—this is equivalent to observing $\psi^{\dagger}$, the formula that says that the fault set contains X_1 or contains both X_2 and one of A_1, A_2, or O_1. That is, $\mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,2}, \langle\varphi^{\dagger} \wedge \psi^{\dagger}\rangle) = \mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,2}, \langle\psi^{\dagger}\rangle)$ includes the belief that the fault set is exactly one of X_1, $\{X_2, A_1\}$, $\{X_2, A_2\}$, and $\{X_2, O_1\}$. By way of contrast, $\mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,1}, \langle\psi^{\dagger}\rangle)$ includes the belief that X_1 is the only fault. It is also a consequence of $K \circ \varphi^{\dagger} \cup \{\psi^{\dagger}\}$ that X_1 is the only fault. It follows that $\mathrm{Bel}(\mathcal{I}^{\dagger}_{diag,2}, \langle\varphi^{\dagger} \wedge \psi^{\dagger}\rangle) \not\subseteq Cl(K \circ \varphi^{\dagger} \cup \{\psi^{\dagger}\})$.

Why does R8 hold in $\mathcal{I}^{\dagger}_{diag,1}$ and not $\mathcal{I}^{\dagger}_{diag,2}$? It turns out that the key reason is that the plausibility measure in $\mathcal{I}^{\dagger}_{diag,1}$ is totally ordered; in $\mathcal{I}^{\dagger}_{diag,2}$ it is only partially ordered. In fact, as we shall see shortly, R8 turns out to be rational monotonicity in disguise. REV2 strengthens BCS4 to ensure that rational monotonicity holds for $\rightarrow$.

REV2. The prior Pl_a on runs that is guaranteed by BCS4 is totally ordered, that is, for all $U, V \subseteq \mathcal{R}$, either $\mathrm{Pl}_a(U) \le \mathrm{Pl}_a(V)$ or $\mathrm{Pl}_a(U) \le \mathrm{Pl}_a(V)$; moreover, $\mathrm{Pl}_a(U \cup V) = \max(\mathrm{Pl}_a(U), \mathrm{Pl}_a(V))$.

There is yet a third condition on BCSs required to make belief change satisfy R1–R8. It makes precise the intuition that observing φ does not give any information beyond φ. This issue was discussed before, in Example 4.1.2. To see its impact on belief revision, consider the following example.

Example 10.3.1 Suppose that $\mathcal{I}$ is a BCS such that the agent observes p_1 at time 0 only if p_2 and q are also true and she observes $p_1 \wedge p_2$ at time

0 only if q is false. It is easy to construct a BCS satisfying REV1 and REV2 that also satisfies this requirement (Exercise 10.6). In this system, after observing p_1, the agent believes p_2 and q (and does not believe $\neg q$) but after observing $p_1 \wedge p_2$, the agent believes (indeed, knows) $\neg q$. This violates both R7 and R8. To see this, note that the assumptions about $\mathcal{I}$ can be phrased as $p_2 \wedge q \in K \circ p_1$ and $\neg q \in K \circ (p_1 \wedge p_2)$. R7 requires that $K \circ (p_1 \wedge p_2) \subseteq Cl(K \circ p_1 \cup \{p_2\})$. Now suppose that R7 holds. Then $\neg q \in Cl(K \circ p_1 \cup \{p_2\})$. But $Cl(K \circ p_1 \cup \{p_2\}) = K \circ p_1$, since $p_2 \in K \circ p_1$. Thus, $\neg q \in K \circ p_1$. But, by assumption, $p \wedge q \in K \circ p_1$, so $K \circ p_1$ is inconsistent. Belief sets in a BCS are always consistent, so R7 cannot hold in $\mathcal{I}$. Similar arguments show that R8 is violated in $\mathcal{I}$ (Exercise 10.6). The problem here is that observing $p_1 \wedge p_2$ gives a great deal of information beyond just the fact that $p_1 \wedge p_2$ is true—it guarantees that $\neg q$ is also true. ∎

Assumption REV3 ensures that observations do not give such additional information. Given a BCS $\mathcal{I}$ and formulas $\varphi, o_1, \ldots, o_k$ in $\mathcal{L}_e$, let $\mathcal{R}[\varphi]$ consist of all runs r where φ is true initially (if $\mathcal{I}$ satisfies REV1, that means that φ is also true throughout the run); let $\mathcal{R}[\varphi_1; o_1, \ldots, o_{k'}]$ consist of all runs r where φ is true initially and the agent observes $o_1, \ldots, o_k$. That is,

$$\begin{aligned} \mathcal{R}[\varphi] &= \{r \in \mathcal{I} : (\mathcal{I}, r, 0) \models \varphi\} \\ \mathcal{R}[\varphi; o_1, \ldots, o_k] &= \{r \in \mathcal{I} : (\mathcal{I}, r, 0) \models \varphi \text{ and } r_a(k) = \langle o_1, \ldots, o_k \rangle\}. \end{aligned}$$

REV3. If $\mathrm{Pl}_a(\mathcal{R}[\varphi; o_1, \ldots, o_m]) > 0$, then $\mathrm{Pl}_a(\mathcal{R}[\varphi; o_1, \ldots, o_m]) \geq \mathrm{Pl}_a(\mathcal{R}[\psi; o_1, \ldots, o_m])$ if and only if $\mathrm{Pl}_a(\mathcal{R}[\varphi \wedge o_1 \wedge \ldots \wedge o_m]) \geq \mathrm{Pl}_a(\mathcal{R}[\psi \wedge o_1 \wedge \ldots \wedge o_m])$.

This assumption captures the intuition that observing $o_1, \ldots, o_k$ provides no more information than just the fact that $o_1 \wedge \ldots \wedge o_m$ is true. That is, the agent compares the plausibility of φ and ψ in the same way after conditioning by the observations $o_1, \ldots, o_m$ as after conditioning by the fact that $o_1 \wedge \ldots \wedge o_m$ is true. It is not hard to see that REV3 fails in the BCS $\mathcal{I}$ constructed in Example 10.3.1. For suppose that Pl_a is the prior plausibility in $\mathcal{I}$. By assumption, in $\mathcal{I}$, after observing p_1, the agent believes p_2 and q, but after observing $p_1 \wedge p_2$, the agent believes $\neg q$. Thus,

$$\mathrm{Pl}_a(\mathcal{R}[p_2 \wedge q; p_1]) > \mathrm{Pl}_a(\mathcal{R}[\neg(p_2 \wedge q); p_1]) \tag{10.1}$$

and

$$\mathrm{Pl}_a(\mathcal{R}[\neg q; p_1 \wedge p_2]) > \mathrm{Pl}_a(\mathcal{R}[q; p_1 \wedge p_2]). \tag{10.2}$$

If REV3 held, then (10.1) and (10.2) would imply

$$\mathrm{Pl}_a(\mathcal{R}[p_1 \wedge p_2 \wedge q]) > \mathrm{Pl}_a(\mathcal{R}[p_1 \wedge \neg(p_2 \wedge q)]) \tag{10.3}$$

and

$$\mathrm{Pl}_a(\mathcal{R}[p_1 \wedge p_2 \wedge \neg q]) > \mathrm{Pl}_a(\mathcal{R}[p_1 \wedge p_2 \wedge q]). \tag{10.4}$$

Since $\mathcal{R}[p_1 \wedge \neg(p_2 \wedge q)] \supseteq \mathcal{R}[p_1 \wedge p_2 \wedge \neg q]$, from (10.3) it follows that

$$\mathrm{Pl}_a(\mathcal{R}[p_1 \wedge p_2 \wedge q]) > \mathrm{Pl}_a(\mathcal{R}[p_1 \wedge p_2 \wedge \neg q]),$$

contradicting (10.4). Thus, REV3 does not hold in $\mathcal{I}$.

Let $\mathcal{REV}$ consist of all reliable BCSs satisfying REV1–REV3. It is easy to see that $\mathcal{I}^{\dagger}_{diag,1} \in \mathcal{REV}$ (Exercise 10.7). The next result shows that, in a precise sense, every BCS in $\mathcal{REV}$ satisfies R1–R8.

Theorem 10.3.2 *Suppose that $\mathcal{I} \in \mathcal{REV}$ and s_a is a local state of the agent at some point in $\mathcal{I}$. Then there is a belief revision operator $\circ_{s_a}$ satisfying R1–R8 such that for all $\varphi \in \mathcal{L}_e$ such that the observation φ can be made in s_a (i.e., for all φ such that $s_a \cdot \varphi$ is a local state at some point in $\mathcal{I}$),* $\mathrm{Bel}(\mathcal{I}, s_a) \circ_{s_a} \varphi = \mathrm{Bel}(\mathcal{I}, s_a \cdot \varphi)$.

Proof See Exercise 10.8. ∎

Theorem 10.3.2 is interesting not just for what it shows, but for what it does *not* show. Theorem 10.3.2 considers a fixed local state s_a in $\mathcal{I}$ and shows that there is a belief revision operator $\circ_{s_a}$ characterizing belief change from s_a. It does not show that there is a single belief revision operator characterizing belief change in all of $\mathcal{I}$. That is, it does not say that there is a belief revision operator $\circ_{\mathcal{I}}$ such that $\mathrm{Bel}(\mathcal{I}, s_a) \circ_{\mathcal{I}} \varphi = \mathrm{Bel}(\mathcal{I}, s_a \cdot \varphi)$, for all local states s_a in $\mathcal{I}$. This stronger result is, in general false. That is because there is more to a local state than the beliefs that are true at that state. The following example illustrates this point.

Example 10.3.3 Consider a BCS $\mathcal{I} = (\mathcal{R}, \pi, \mathcal{PL})$ such that the following hold:

- $\mathcal{R} = \{r_1, r_1', r_2, r_2', r_3, r_3'\}$.
- π is such that $p_1 \wedge p_2 \wedge p_3$ is true throughout r_1 and r_1', $\neg p_1 \wedge \neg p_2 \wedge p_3$ is true throughout r_2 and r_2', and $p_1 \wedge \neg p_2 \wedge \neg p_3$ is true throughout r_3 and r_3'.
- In runs r_1, r_2, and r_3, the agent observes whether or not p_1 is true in the first round, and then observes whether or not p_2 is true at all subsequent rounds; at runs r_1', r_2', and r_3' observes whether or not p_2 is true at the first round and then observes whether or not p_1 is true at all subsequent rounds. Thus, for example,

  - $o_{(r_1,1)} = p_1$ and $o_{(r_1,2)} = o_{(r_1,3)} = \ldots = p_2$;
  - $o_{(r'_1,1)} = p_2$ and $o_{(r'_1,2)} = o_{(r'_1,3)} \ldots = p_1$;
  - $o_{(r_2,1)} = \neg p_1$ and $o_{(r_2,2)} = o_{(r_2,3)} = \ldots = \neg p_2$.

- $\mathcal{PL}$ is determined by a prior Pl_a on runs, where

$$\mathrm{Pl}_a(r_1) = \mathrm{Pl}_a(r'_1) > \mathrm{Pl}_a(r_2) = \mathrm{Pl}_a(r'_2) > \mathrm{Pl}_a(r_3) = \mathrm{Pl}_a(r'_3).$$

It is easy to check that $\mathcal{I} \in \mathcal{REV}$ (Exercise 10.9). Since p_1 is true in the most plausible runs (r_1 and r'_1), $p_1 \in \mathrm{Bel}(\mathcal{I}, \langle\,\rangle)$. By R3 and R4, the agent's beliefs do not change if she observes p_1. Thus, $\mathrm{Bel}(\mathcal{I}, \langle\,\rangle) = \mathrm{Bel}(\mathcal{I}, \langle p_1\rangle)$. Let $K = \mathrm{Bel}(\mathcal{I}, \langle\,\rangle) = \mathrm{Bel}(\mathcal{I}, \langle p_1\rangle)$. Suppose that there were a revision operator $\circ$ such that $\mathrm{Bel}(\mathcal{I}, s_a) \circ \varphi = \mathrm{Bel}(\mathcal{I}, s_a \cdot \varphi)$ for all local states s_a. It would then follow that $\mathrm{Bel}(\mathcal{I}, \langle\neg p_2\rangle) = \mathrm{Bel}(\mathcal{I}, \langle p_1; \neg p_2\rangle) = K \circ \neg p_2$. However, it is easy to see that $p_3 \in \mathrm{Bel}(\mathcal{I}, \langle\neg p_2\rangle)$ and $\neg p_3 \in \mathrm{Bel}(\mathcal{I}, \langle p_1; \neg p_2\rangle)$ (Exercise 10.9), which leads to a contradiction with R5. ∎

Example 10.3.3 illustrates a problem with the assumption implicit in AGM belief revision, that all that matters regarding how an agent revises her beliefs is her belief set and what is learnt. I return to this problem in the next section.

Theorem 10.3.2 shows that for every BCS $\mathcal{I} \in \mathcal{REV}$ and local state s_a, there is a revision operator characterizing belief change at s_a. The next result is essentially a converse.

Theorem 10.3.4 *Let $\circ$ be a belief revision operator satisfying R1–R8 and let $K \subseteq \mathcal{L}_e$ be a consistent belief state. Then there is a BCS $\mathcal{I}_K$ in $\mathcal{REV}$ such that $\mathrm{Bel}(\mathcal{I}_K, \langle\,\rangle) = K$ and*

$$\mathrm{Bel}(\mathcal{I}_K, \langle\,\rangle) \circ \varphi = \mathrm{Bel}(\mathcal{I}_K, \langle\varphi\rangle)$$

for all $\varphi \in \mathcal{L}_e$.

Proof See Exercise 10.10. ∎

Notice that Theorem 10.3.4 considers only *consistent* belief sets K. The AGM postulates allow the agent to "escape" from an inconsistent belief set, so that $K \circ \varphi$ may be consistent even if K is inconsistent. Indeed, R5 *requires* that it be possible to escape from an inconsistent belief set. The requirement that K be consistent is necessary in Theorem 10.3.4. If *false* $\in \mathrm{Bel}(\mathcal{I}_K, s_a)$ for some state s_a and $r_a(m) = s_a$, then $\mathrm{Pl}_{(r,m)}(W_{(r,m)}) = \perp$. Since updating is done by conditioning, $\mathrm{Pl}_{(r,m+1)}(W_{(r,m+1)}) = \perp$, so the agent's belief set will remain inconsistent no matter what she learns. Thus, BCSs do not allow an agent to escape from an inconsistent belief set. This is a consequence of the use of conditioning to update.

Although it would be possible to modify the definition of BCSs to handle updates of inconsistent belief sets differently (and thus to allow the agent to escape from inconsistent belief set), I believe that it would in fact be more appropriate to reformulate R5 so that it does not require escape from an inconsistent belief set. Consider the following postulate.

R5*. $K \circ \varphi = Cl(\mathit{false})$ if and only if $\vdash_{\mathcal{L}} \neg\varphi$ or $\mathit{false} \in K$.

If R5 is replaced by R5*, then Theorem 10.3.4 holds even if K is inconsistent (for trivial reasons, since in that case $K \circ \varphi = K$ for all φ).

I conclude this section with a result that relates belief revision in systems in $\mathcal{REV}$ to the conditional logic considered in Section 7.4. It is perhaps not surprising that there should be a connection between the two, given that both use plausibility measures as a basis for their semantics.

Theorem 10.3.5 *Suppose that $\mathcal{I}$ is a reliable BCS that satisfies REV1 and REV3. If r is a run in $\mathcal{I}$ such that $o_{(r,m+1)} = \varphi$, then $(\mathcal{I}, r, m) \models \varphi \rightarrow \psi$ iff $(\mathcal{I}, r, m+1) \models B\psi$. Equivalently, if $s_a \cdot \varphi$ is a local state in $\mathcal{I}$, then*

$$(\mathcal{I}, s_a) \models \varphi \rightarrow \psi \text{ iff } (\mathcal{I}, s_a \cdot \varphi) \models B\psi.$$

Proof See Exercise 10.11. ∎

Using Theorems 10.3.4 and 10.3.5, I can make the connection between R8 and rational monotonicity (axiom C5 in Section 7.4) mentioned earlier: If $\mathcal{I}$ is a BCS satisfying REV1 and REV3, then $\mathcal{I}$ satisfies rational monotonicity iff belief change in $\mathcal{I}$ satisfies R8. For suppose that $\mathcal{I}$ is a BCS satisfying REV1 and REV3, and $K = \mathrm{Bel}(\mathcal{I}, s_a)$ for some local state s_a in $\mathcal{I}$. Then $\neg\psi \notin K \circ \varphi$ iff $\neg\psi \notin \mathrm{Bel}(\mathcal{I}, s_a \cdot \varphi)$. By Theorem 10.3.5, this is the case iff $(\mathcal{I}, s_a) \models \neg(\varphi \rightarrow \neg\psi)$. If $\mathcal{I}$ satisfies REV2, then it satisfies rational monotonicity. Thus, $(\mathcal{I}, s_a) \models (\varphi \rightarrow \psi') \Rightarrow ((\varphi \wedge \psi) \rightarrow \psi')$. By Theorem 10.3.5 again, this means that if $(\mathcal{I}, s_a \cdot \varphi) \models B\psi'$ then $(\mathcal{I}, s_a \cdot (\varphi \wedge \psi) \models B\psi'$. That is, $K \circ \varphi \subseteq K \circ (\varphi \wedge \psi)$. Since $K \circ (\varphi \wedge \psi) = \mathrm{Bel}(\mathcal{I}, s_a \cdot (\varphi \wedge \psi))$ contains $K \circ \varphi$ and ψ, and is a closed set, it follows that $Cl(K \circ \varphi \cup \{\psi\}) \subseteq K \circ (\varphi \wedge \psi)$, as required by R8. The argument works equally well in the other direction, showing that R8 implies rational monotonicity.

10.4 Epistemic States and Iterated Revision

Agents do not change their beliefs just once. They do so repeatedly, each time they get new information. The BCS framework models this naturally, showing how the agent's local state changes as a result of each new observation. It would seem at first that revision operators make sense for iterated revision as well. Given a revision operator $\circ$ and an initial belief set K, it seems reasonable, for example, to take $(K \circ \varphi_1) \circ \varphi_2$ to be the result of revising first by φ_1 and then by φ_2. However, Example 10.3.3 indicates that there is a problem with this approach. Even if $(K \circ \varphi_1) = K$, we may not want to have $(K \circ \varphi_1) \circ \varphi_2 = K \circ \varphi_2$. In Example 10.3.3, revising by φ_1 and then φ_2 is not the same as revising by φ_2, even though the agent has the same belief set before and after revising by φ_1.

The culprit here is the assumption that revision depends only on the agent's belief set. In a BCS, there is a clear distinction between the agent's *epistemic state* at a point (r, m) in $\mathcal{I}$, as characterized by her local state $s_a = r_a(m)$, and the agent's belief set at (r, m), $\mathrm{Bel}(\mathcal{I}, s_a)$. As Example 10.3.3 shows, in a system in $\mathcal{REV}$, the agent's belief set does not in general determine how the agent's beliefs will be revised; her epistemic state does.

It is not hard to modify the AGM postulates to deal with revision operators that take as their first argument epistemic states rather than belief sets. Suppose that there is a set of epistemic states (the exact form of the epistemic state is irrelevant for the following discussion) and a function $\mathrm{BS}(\cdot)$ that maps epistemic states to belief sets. There is an analogue to each of the AGM postulates, obtained by replacing each belief set by the beliefs in the corresponding epistemic state. Letting E stand for a generic epistemic state, the modified postulates are

R1$'$. $E \circ \varphi$ is an epistemic state.

R2$'$. $\varphi \in \mathrm{BS}(E \circ \varphi)$.

R3$'$. $\mathrm{BS}(E \circ \varphi) \subseteq Cl(\mathrm{BS}(E) \cup \{\varphi\})$.

and so on, with the obvious transformation. The only problematic postulate is R6. The question is whether R6$'$ should be "If $\vdash_{\mathcal{L}_e} \varphi \Leftrightarrow \psi$ then $\mathrm{BS}(E \circ \varphi) = \mathrm{BS}(E \circ \psi)$" or "If $\vdash_{\mathcal{L}_e} \varphi \Leftrightarrow \psi$ then $E \circ \varphi = E \circ \psi$". Dealing with either version is straightforward. For definiteness, I adopt the first alternative here.

There is an analogue of Theorem 10.3.2 that works at the level of epistemic states. Indeed, working at the level of epistemic states gives a more elegant result. Given a BCS $\mathcal{I} \in \mathcal{REV}$, there is a single revision operator $\circ$ that characterizes belief revision in $\mathcal{I}$; it is not necessary to use a different revision operator for each local state s_a in $\mathcal{I}$.

To make this precise, given a language $\mathcal{L}_e$, let $\mathcal{L}_e^*$ consist of all sequences of formulas in $\mathcal{L}_e$. In a BCS, the local states are elements of $\mathcal{L}_e^*$ (although some elements in $\mathcal{L}_e^*$, such as $\langle p, \neg p\rangle$, cannot arise as local states in a reliable BCS). Define a revision function $\circ$ on $\mathcal{L}_e^*$ in the obvious way: if $E \in \mathcal{L}_e^*$, then $E \circ \varphi = E \cdot \varphi$.

Theorem 10.4.1 *Let $\mathcal{I}$ be a system in $\mathcal{REV}$ whose local states are in $\mathcal{L}_e^*$. There is a function $BS_{\mathcal{I}}$ that maps epistemic states to belief sets such that*

- *if s_a is a local state of the agent in $\mathcal{I}$, then* $\mathrm{Bel}(\mathcal{I}, s_a) = BS_{\mathcal{I}}(s_a)$*, and*
- *$(\circ, BS_{\mathcal{I}})$ satisfies R1′–R8′.*

Proof Note that $\mathrm{BS}_{\mathcal{I}}$ must be defined on all sequences in $\mathcal{L}_e^*$, including ones that are not local states in $\mathcal{I}$. Define $\mathrm{BS}_{\mathcal{I}}(s_a) = \mathrm{Bel}(\mathcal{I}, s_a)$ if s_a is a local state in $\mathcal{I}$. If s_a is not in $\mathcal{I}$, then $\mathrm{BS}_{\mathcal{I}}(s_a) = \mathrm{Bel}(\mathcal{I}, s')$, where s' is the longest suffix of s_a that is a local state in $\mathcal{I}$. The argument that this works is left to the reader (Exercise 10.13). ∎

At first blush, the relationship between Theorem 10.4.1 and Theorem 10.3.2 may not be so clear. However, note that, by definition,

$$\mathrm{BS}_{\mathcal{I}}(\mathcal{I}, \langle s_a\rangle \circ \varphi_1 \circ \ldots \circ \varphi_k) = \mathrm{BS}_{\mathcal{I}}(\mathcal{I}, s_a \cdot \langle \varphi_1, \ldots, \varphi_k\rangle),$$

so, at the level of epistemic states, Theorem 10.4.1 is a generalization of Theorem 10.3.2.

Theorem 10.4.1 shows that any system in $\mathcal{REV}$ corresponds to a revision operator over epistemic states that satisfies the modified AGM postulates. Is there a converse, analogous to Theorem 10.3.4? Not quite. It turns out that R7′ and R8′ are not quite strong enough to capture the behavior of conditioning given a consistent observation. It is not hard to show that R7′ and R8′ (together with R4′ and R5′) imply that

$$\text{if } \neg\psi \notin \mathrm{BS}(E \circ \varphi), \text{ then } \mathrm{BS}(E \circ \varphi \circ \psi) = \mathrm{BS}(E \circ (\varphi \wedge \psi)) \qquad (10.5)$$

(Exercise 10.12(a)). The following postulate strengthens this:

R9′. If $\not\vdash_{\mathcal{L}_e} \neg(\varphi \wedge \psi)$ then $\mathrm{BS}(E \circ \varphi \circ \psi) = \mathrm{BS}(E \circ \varphi \wedge \psi)$.

R9′ says that revising E by φ and then by ψ is the same as revising by $\varphi \wedge \psi$ if $\varphi \wedge \psi$ is consistent. This indeed strengthens (10.5), since (given R1′ and if $\neg\psi \notin \mathrm{BS}(E \circ \varphi)$ then $\not\vdash_{\mathcal{L}_e} \neg(\varphi \wedge \psi)$ (Exercise 10.12(b)). It is not hard to show that it is a nontrivial strengthening; there are systems that satisfy (10.5) and do not satisfy R9′ (Exercise 10.12(c)).

The following generalization of Theorem 10.4.1 shows that R9′ is sound in $\mathcal{REV}$.

Theorem 10.4.2 *Let $\mathcal{I}$ be a system in $\mathcal{REV}$ whose local states are $\mathcal{E}_{\mathcal{L}_e}$. There is a function $BS_{\mathcal{I}}$ that maps epistemic states to belief sets such that*

- *if s_a is a local state of the agent in $\mathcal{I}$, then* $\mathrm{Bel}(\mathcal{I}, s_a) = BS_{\mathcal{I}}(s_a)$*, and*
- $(\circ, BS_{\mathcal{I}})$ *satisfies R1′–R9′.*

Proof See Exercise 10.13. ∎

The converse to Theorem 10.4.2 does hold: a revision system on epistemic states that satisfies the generalized AGM postulates *and R9′* corresponds to a system in $\mathcal{REV}$. Let $\mathcal{L}_e^{\dagger}$ consist of all the sequences $\langle \varphi_1, \ldots, \varphi_k \rangle$ in $\mathcal{L}_e^*$ that are consistent, in that $\not\vdash_{\mathcal{L}_e} \neg(\varphi_1 \wedge \ldots \wedge \varphi_k)$.

Theorem 10.4.3 *Given a function $BS_{\mathcal{L}_e}$ mapping epistemic states in $\mathcal{L}_e^*$ to belief sets over $\mathcal{L}_e$ such that $BS_{\mathcal{L}_e}(\langle\rangle)$ is consistent and $(BS_{\mathcal{L}_e}, \circ)$ satisfies R1′–R9′, there is a system $\mathcal{I} \in \mathcal{REV}$ whose local states consist of all the states in $\mathcal{L}_e^{\dagger}$ such that $BS_{\mathcal{L}_e}(s_a) = BS(s_a)$ for $s_a \in \mathcal{L}_e^{\dagger}$.*

Proof Let $\mathcal{I} = (\mathcal{R}, \mathcal{PL}, \pi)$ be defined as follows. A run in $\mathcal{R}$ is defined by a truth assignment α to the primitive propositions in $\mathcal{L}_e$ and an infinite sequence $\langle o_1, o_2, \ldots \rangle$ of observations each of which is true under truth assignment α. The pair $(\alpha, \langle o_1, o_2, \ldots \rangle)$ define a run r in the obvious way: $r_e(m) = \alpha$ for all m and $r_a(m) = \langle o_1, o_2, \ldots, o_m \rangle$. $\mathcal{R}$ consists of all runs that can be defined in this way. The interpretation is determined by α: $\pi(r, m) = r_e(m)$. All that remains is to define a prior that ensures that $\mathrm{BS}_{\mathcal{L}_e}(s_a) = \mathrm{BS}(s_a)$ for all $s_a \in \mathcal{L}_e^{\dagger}$. This is left to the reader (Exercise 10.14). ∎

So where does this leave us? This discussion shows that, at the level of epistemic states, the AGM postulates are very reasonable (with the possible exception of R5, which perhaps should be modified to R5*) provided that (a) we are interested in reasoning only about static propositions (whose truth values do not change over time), (b) observations are reliable (in that we take what is observed to be true), (c) nothing is learned from observing φ beyond the fact that φ is true, and (d) there is a totally ordered plausibility on truth assignments (which by (a) and (c) determines the plausibility on runs). The generality of plausibility measures is not required for (d); using ranking functions, possibility measures, or total preference orders will do as well.

But what happens if want to consider situations where some of these assumptions are violated? Nothing in the BCS framework requires them; it makes perfect sense to consider BCSs that violate any or all of them. For example, it is easy enough to allow partial orders instead of total orders on runs—the effect of this is just that R8 (or R8′) no longer holds. Still, it would be useful to have a model of belief change that does not make these assumptions, but still has enough structure to be comprehensible. The Markov assumption discussed in Section 9.7 has the required properties. This topic is the subject of the next section.

10.5 Markovian Belief Revision

For the purposes of this section, I restrict attention to BCSs where the plausibility measures are *algebraic*, in the sense defined in Section 4.8; that is, they have operations $\oplus$ and $\otimes$ such that $\mathrm{Pl}(U|V) \otimes \mathrm{Pl}(V|V') = \mathrm{Pl}(U|V')$ if $U \subseteq V \subseteq V'$ and $\mathrm{Pl}(U_1 \cup U_2|V) = \mathrm{Pl}(U_1|V) \oplus \mathrm{Pl}(U_2|V)$ if $U_1 \cap U_2 = \emptyset$.

In BCSs with an algebraic prior plausibility, the notion of a Markovian plausibility measure makes perfect sense. Not surprisingly, such BCSs are called *Markovian BCSs*. To see the power of Markovian BCSs as a modeling tool, consider the following example.

Example 10.5.1 A car is parked with a nonempty fuel tank at time 0. The owner returns at time 2 to find his car still there. Not surprisingly, at this point he believes that the car has been there all along and still has a nonempty tank. He then observes that the fuel tank is empty. He considers two possible explanations: that his wife borrowed the car to do some errands or that the gas leaked. (Suppose that the "times" are sufficiently long and the tank is sufficiently small that it is possible that both doing some errands and a leak can result in an empty tank.)

To model this as a BCS, suppose that Φ_e consists of two primitive propositions: *Parked* (which is true if car is parked where the owner originally left it) and *Empty* (which is true if the tank is empty). The environment state is just a truth assignment to these two primitive propositions. This truth assignment clearly change over time, so REV1 is violated. (It would be possible to instead use propositions of the form Parked_i—the car is parked at time i—which would allow REV1 to be maintained; for simplicity, I consider here only the case where there are two primitive propositions.) There are three environment states: $s_{\mathrm{p}\overline{\mathrm{e}}}$, s_{pe}, and $s_{\overline{\mathrm{pe}}}$. In $s_{\mathrm{p}\overline{\mathrm{e}}}$, $\mathit{Parked} \land \neg \mathit{Empty}$ is true; in s_{pe}, $\mathit{Parked} \land \mathit{Empty}$ is true; and in $s_{\overline{\mathrm{pe}}}$, $\neg \mathit{Parked} \land \neg \mathit{Empty}$ is true. For simplicity, assume that in all runs in the system, $\mathit{Parked} \land \neg \mathit{Empty}$ is true at time 0 and $\mathit{Parked} \land \mathit{Empty}$ is true at times 2 and 3. Further assume that in all runs the agent correctly observes *Parked* at time 2, and *Empty* at time 3, and makes no observations (i.e., observes *true*) at time 1.

I model this system using a Markovian prior. The story suggests that the most likely transitions are the ones where no change occurs, which is why the agent believes at time 2—before he observes that the tank is empty—that the car has not moved and the tank is still not empty. Once he discovers that the tank is empty, the explanation he considers most likely will depend on his ranking of the transitions. This can be captured easily using ranking functions (which are algebraic plausibility measures). For example, the agent's belief that the most likely transitions are ones where no change occurs can be modeled by taking $t_{s,s} = 0$ and $t_{s,s'} > 0$ if $s \neq s'$, for $s, s' \in \{s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\overline{\mathrm{pe}}}, s_{\mathrm{pe}}\}$. This is already enough to make $[s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{p}\overline{\mathrm{e}}}]$ the most plausible 2-prefix. (Since for each time $m \in \{0, \ldots, 3\}$, the agent's local state is the same at time m in all runs, I do not mention it in the global state.) Thus, when the agent returns at time 2 to find his car parked, he believes that it was parked all along and the tank is not empty.

How do the agent's beliefs change when he observes that the tank is empty at time 3? As I said earlier, I restrict attention to two explanations: his wife borrowed the car to do some errands, which corresponds to the runs with 2-prefix is $[s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\overline{\mathrm{pe}}}, s_{\mathrm{pe}}]$, or the gas tanked leaked, which corresponds to the runs with 2-prefix $[s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{pe}}, s_{\mathrm{pe}}]$ and $[s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{pe}}]$ (depending on when the leak started). The relative likelihood of the explanations depends on the relative likelihood of the transitions. He considers it more likely that his wife borrowed the car if the transition from $s_{\mathrm{p}\overline{\mathrm{e}}}$ to s_{pe} less likely than the sum of the transitions from $s_{\mathrm{p}\overline{\mathrm{e}}}$ to $s_{\overline{\mathrm{pe}}}$ and from $s_{\overline{\mathrm{pe}}}$ to s_{pe}, for example, if $t_{s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{pe}}} = 3, t_{s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\overline{\mathrm{pe}}}} = 1$, and $t_{s_{\overline{\mathrm{pe}}}, s_{\mathrm{pe}}} = 1$. Applying the Markovian assumption and the fact that $\otimes$ is $+$ for rankings, these choices make $\kappa([s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\overline{\mathrm{pe}}}, s_{\mathrm{pe}}]) = 2$ and $\kappa([s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{pe}}]) = \kappa([s_{\mathrm{p}\overline{\mathrm{e}}}, s_{\mathrm{pe}}, s_{\mathrm{pe}}]) = 3$. By changing the likelihood of the transitions, it is clearly possible to make the two explanations equally likely or to make the gas leak the more likely explanation. ∎

This example was simple because the agent's local state (i.e., the observations made by the agents) did not affect the likelihood of transition. In general, the observations the agent makes do affect the transitions. Using the Markovian assumption, it is possible to model the fact that an agent's observations are correlated with the state of the world (for example, the agent being more likely to observe p if both p and q are true than if $p \wedge \neg q$ is true) and to model unreliable observations that are still usually correct (for example, the agent being more likely to observe p if p is true than if p is false or p being more likely to be true if the agent observes p than if the agent observes $\neg p$—note that these are two quite different assumptions).

These examples show the flexibility of the Markovian assumption. While it can be difficult to decide how beliefs should change, this approach seems to localize the effort in what appears to be the right place: deciding the relative likelihood of various transitions. An obvious question now is whether making the Markovian assumption puts any constraints on BCSs. As the following result shows, the answer is no, at least as far as belief sets go.

Theorem 10.5.2 *Given a BCS $\mathcal{I}$, there is a Markovian BCS $\mathcal{I}'$ such that the agent's local states are the same in both $\mathcal{I}$ and $\mathcal{I}'$ and for all local states s_a,* $\mathrm{Bel}(\mathcal{I}, s_a) = \mathrm{Bel}(\mathcal{I}', s_a)$.

Proof Suppose that $\mathcal{I} = (\mathcal{R}, \mathcal{PL}, \pi)$. Let Pl be the prior on runs that determines Pl. Although the agent's local state must be the same in $\mathcal{I}$ and $\mathcal{I}'$, there is no such requirement on the environment state. The idea is to define a set $\mathcal{R}'$ of runs where the environment states have the form $\langle g_0, \ldots, g_m\rangle$, for all possible initial sequences $g_0, \ldots, g_m$ of global states that arise in runs of $\mathcal{R}$. Then $\mathcal{I}' = (\mathcal{R}', \mathcal{PL}', \pi')$, where $\pi'(\langle g_0, \ldots, g_m\rangle) = \pi(g_m)$ and $\mathcal{PL}'$ is generated by a Markovian prior Pl' that simulates Pl in that $\mathrm{Pl}'([\langle g_0\rangle, \langle g_0, g_1\rangle, \ldots, \langle g_0, \ldots, g_m\rangle])$ "acts the same as" $\mathrm{Pl}([g_0, \ldots, g_m])$. "Acts the same as" essentially means "is equal to" here; however, since Pl' must be algebraic, equality cannot necessarily be assumed. It suffices that $\mathrm{Pl}'([\langle g_0\rangle, \langle g_0, g_1\rangle, \ldots, \langle g_0, \ldots, g_m\rangle]) > \mathrm{Pl}'([\langle g_0'\rangle, \langle g_0', g_1'\rangle, \ldots, \langle g_0', \ldots, g_{m'}'\rangle])$ iff $\mathrm{Pl}([g_0, \ldots, g_m]) > \mathrm{Pl}([g_0', \ldots, g_{m'}'])$. I leave the technical details to the reader (Exercise 10.15). ∎

Exercises

10.1 (a) Show that Pl_1 is the plausibility measure obtained from the probability sequence $(\mu_1, \mu_2, \mu_3, \ldots)$ defined in Section 10.1, using the construction preceding Theorem 7.2.11.

(b) Define a probability sequence $(\mu_1', \mu_2', \mu_3', \ldots)$ from which Pl_2 is obtained using the construction preceding Theorem 7.2.11.

10.2 Prove Proposition 10.1.1.

10.3 Prove Proposition 10.1.2.

10.4 Prove Proposition 10.1.3.

10.5 Show that a BCS is a synchronous stem where the agent has perfect recall that satisfies CONS.

10.6 Construct a BCS satisfying REV1 and REV2 that satisfies the additional requirement of Example 10.3.1. Show that R8 is violated in this system.

10.7 Show that $\mathcal{I}^{\dagger}_{diag,1} \in \mathcal{REV}$.

* **10.8** Prove Theorem 10.3.2. (The hard part of the proof is coming up with a definition of $\mathrm{Bel}(\mathcal{I}, s_a) \circ_{s_a} \varphi$ for formulas φ that cannot be observed in state s_a. Note also that $K \circ_{s_a} \varphi$ must be defined even for $K \neq \mathrm{Bel}(\mathcal{I}, s_a)$.)

10.9 Show that the BCS $\mathcal{I}$ constructed in Example 10.3.3 is in $\mathcal{REV}$ and that $p_3 \in \mathrm{Bel}(\mathcal{I}, \langle \neg p_2 \rangle)$ and $\neg p_3 \in \mathrm{Bel}(\mathcal{I}, \langle p_1; \neg p_2 \rangle)$.

* **10.10** Prove Theorem 10.3.4. Further show that it

10.11 Prove Theorem 10.3.5.

10.12 (a) Show that (10.5) follows from R4′, R5′, R7′, and R8′.

(b) Show that if BS satisfies R1′ and $\neg\psi \notin \mathrm{BS}(E \circ \varphi)$, then $\not\vdash_{\mathcal{L}_e} \neg(\varphi \wedge \psi)$.

(c) Describe a system $\mathcal{I}$ that satisfies (10.5) and not R9′.

(d) Show that R8′ follows from R9′ and R4′.

* **10.13** Complete the proof of Theorem 10.4.1. Moreover, show that $(\circ, \mathrm{BS}_{\mathcal{I}})$ satisfies R1′–R9′, thus proving Theorem 10.4.2.

* **10.14** Complete the proof of Theorem 10.4.3.

* **10.15** Complete the proof of Theorem 10.5.2. (The difficulty here, as suggested in the text, is making Pl′ algebraic.)

Notes

Belief change has been an active area of study in philosophy and, more recently, artificial intelligence. Probabilistic conditioning can be viewed as one approach to belief change, but the study of the type of belief change considered in this chapter, where an agent must revise her beliefs after learning or observing something inconsistent with them, was mainly initiated by Alchourrón, Gärdenfors, and Makinson, in a sequence of individual and joint papers. A good introduction to the topic, with an extensive bibliography of the earlier work, is Gärdenfors' book *Knowledge in Flux* [1988]. AGM-style belief revision was introduced by Alchourrón, Gärdenfors, and

Makinson [1985]. Interestingly, the topic of belief change was studied independently in the database community—the problem here was how to update a database when the update is inconsistent with information already stored in the database. The original paper on the topic was by Fagin, Ullman, and Vardi [1983]. One of the more influential axiomatic characterizations of belief change—Katsuno and Mendelzon's notion *belief update* [1991a]—was inspired by database concerns.

The presentation in this chapter is largely taken from a sequence of papers Nir Friedman and I wrote. Section 10.1 is largely taken from [Friedman and Halpern 1997]; the discussion of belief change and the AGM axioms as well as iterated belief revision is largely taken from [Friedman and Halpern 1999] (although there are a number of minor differences between the presentation here and that in [Friedman and Halpern 1999]); the discussion of Markovian belief change is from [Friedman and Halpern 1996]. In particular, Propositions 10.1.1, 10.1.2, and 10.1.3 are taken from [Friedman and Halpern 1997], Theorems 10.3.2, 10.3.4, 10.3.5, 10.4.1, 10.4.2, and 10.4.3 are taken (with minor modifications in some cases) from [Friedman and Halpern 1999], and Theorem 10.5.2 is taken from [Friedman and Halpern 1996].

These papers have references to more current research in belief change, which is still an active topic. Here I just give bibliographic references to the specific topics discussed in this chapter.

The circuit diagnosis problem discussed has been well studied in the artificial intelligence literature (see [Davis and Hamscher 1988] for an overview). The discussion here loosely follows the examples of Reiter [1987]. Representation theorems for the AGM postulates are well known. The earliest is due to Grove [1988]; others can be found in [Boutilier 1994; Katsuno and Mendelzon 1991b; Gärdenfors and Makinson 1988]. Iterated belief change has been the subject of much research; see, for example, [Boutilier 1996; Darwiche and Pearl 1997; Freund and Lehmann 1994; Lehmann 1995; Levi 1988; Nayak 1994; Williams 1994]). Markovian belief change is also considered in [**?**; Boutilier, Halpern, and Friedman 1998].