Chapter 7

Defaults and Counterfactuals

Two types of reasoning that arise frequently in everyday life are *default reasoning* and *counterfactual reasoning*. Default reasoning involves leaping to conclusions. For example, if we see a bird, we may conclude that it flies. Now flying is not a logical consequence of being a bird. Not all birds fly. Penguins and ostriches do not fly, nor do newborn birds, injured birds, dead birds, or birds made of clay. Nevertheless we do think of flying as a prototypical property of birds. Concluding that a bird flies seems reasonable, as long as we are willing to retract that conclusion in the face of extra information.

Counterfactual reasoning involves reaching conclusions with assumptions that may be counter to fact. Suppose that we are interested in assigning blame in a car accident. A lawyer might well want to argue as follows: "I admit that my client was drunk and it was raining. Nevertheless, if the car's brakes had functioned properly, the car would not have hit Mrs. McGraw's cow. The car's manufacturer is at fault at least as much as my client."

As the lawyer admits here, his client was drunk and it was raining. He is arguing though that even if the client hadn't been drunk and it weren't raining, the car would have hit the cow. This is a classic case of counterfactual reasoning: reasoning about what might have happened if things had been different from the way they actually were.

Why am I discussing default reasoning and counterfactual reasoning at this point in the book? It should be clear that both involve reasoning about uncertainty. Moreover, it turns out that some of the methods of representing uncertainty that we have been considering—specifically, possibility measures, ranking functions, and plausibility measures—provide good frameworks for capturing both default reasoning and counterfactual reasoning.

7.1 Characterizing Default Reasoning

As a first step to investigating default reasoning, consider a very simple language for representing defaults. Starting with a set Φ of primitive propositions, let the language $\mathcal{L}^{def}(\Phi)$ consist of all formulas of the form $\varphi \rightarrow \psi$, where $\varphi, \psi \in \mathcal{L}^{Prop}(\Phi)$, that is, they are propositional formulas over Φ. Notice that $\mathcal{L}^{def}$ is not closed under negation or disjunction; for example, $\neg(p \rightarrow q)$ is not a formula in $\mathcal{L}^{def}$, nor is $(p \rightarrow q) \Rightarrow (p \rightarrow (q \vee q'))$ (although, of course, $\neg p \rightarrow q$ and $p \rightarrow (q \Rightarrow q')$ are in $\mathcal{L}^{def}$).

We can read $\varphi \rightarrow \psi$ in various ways, depending on the application. For example, it can be read as "if φ (is the case) then typically ψ (is the case)", "if φ, then normally ψ", "if φ, then by default ψ", and "if φ, then ψ is very likely". Thus, the default statement "birds typically fly" is represented as $bird \rightarrow fly$. We can also use $\mathcal{L}^{def}$ for counterfactual reasoning, in which case $\varphi \rightarrow \psi$ is interpreted as "if φ were true, then ψ would be true".

Notice that all these readings are similar in spirit to the reading of the formula $\varphi \Rightarrow \psi$ in propositional logic as "if φ then ψ". How do the properties of $\Rightarrow$ (often called a *material conditional*) and $\rightarrow$ compare? More generally, what properties should $\rightarrow$ have? That depends to some extent on our interpretation of $\rightarrow$. We should not expect default reasoning and counterfactual reasoning to have the same properties (although, as we shall see, there is a lot of overlap). In this section, I focus on default reasoning, with the goal of defining a *logic* of default reasoning.

There has been some disagreement in the literature as to what properties $\rightarrow$ should have. However, there seems to be some consensus on a minimal set of six *core* properties. Whatever other properties default reasoning should have, these six should be among them. These six axioms and inference rules, given below, make up the axiom system **P**.

LLE. If $\varphi \Leftrightarrow \varphi'$ is a propositional tautology, then from $\varphi \rightarrow \psi$ infer $\varphi' \rightarrow \psi$ (left logical equivalence).

RW. If $\psi \Rightarrow \psi'$ is a propositional tautology, then from $\varphi \rightarrow \psi$ infer $\varphi \rightarrow \psi'$ (right weakening).

REF. $\varphi \rightarrow \varphi$ (reflexivity).

AND. From $\varphi \rightarrow \psi_1$ and $\varphi \rightarrow \psi_2$ infer $\varphi \rightarrow \psi_1 \wedge \psi_2$.

OR. From $\varphi_1 \rightarrow \psi$ and $\varphi_2 \rightarrow \psi$ infer $\varphi_1 \vee \varphi_2 \rightarrow \psi$.

CM. From $\varphi \rightarrow \psi_1$ and $\varphi \rightarrow \psi_2$ infer $\varphi \wedge \psi_2 \rightarrow \psi_1$ (cautious monotonicity).

The first three properties of **P** seem noncontroversial. If φ and φ' are logically equivalent, then surely if we can conclude ψ by default from φ, we should also be able to conclude it from φ'. Similarly if we can conclude ψ by default from φ, and ψ logically implies ψ', then we surely should be able to conclude ψ' as well. Finally, reflexivity just says that we should be able to conclude φ from itself.

The latter three properties get more into the heart of default reasoning. The AND rule says that defaults are closed under conjunction. For example, if we see a bird, it seems reasonable to conclude that it flies. It also seems reasonable to conclude that it has wings. The AND rule lets us put these two conclusions together and conclude that, by default, birds both fly and have wings.

The OR rule corresponds to reasoning by cases. If red birds typically fly ($(red \wedge bird) \rightarrow fly$) and non-red birds typically fly ($(\neg red \wedge bird) \rightarrow fly$), then birds typically fly, no matter what color they are. Note that the OR rule actually gives us only $((red \wedge bird) \vee (\neg red \wedge bird)) \rightarrow fly$ here. To conclude $bird \rightarrow fly$, we have to apply LLE as well, using the fact that $bird \Leftrightarrow ((red \wedge bird) \vee (\neg red \wedge bird))$ is a propositional tautology.

To understand cautious monotonicity, note that one of the most important properties of the material conditional is that it is *monotonic*. Getting extra information never causes us to withdraw conclusions. From $\varphi \Rightarrow \psi$, we can conclude $\varphi \wedge \varphi' \Rightarrow \psi$, no matter what φ' is (Exercise 7.1). On the other hand, default reasoning is not always monotonic. From $bird \rightarrow fly$ we do not necessarily want to conclude $bird \wedge penguin \rightarrow fly$. Discovering that a bird is a penguin should cause us to retract the conclusion that it flies.

Cautious monotonicity captures one instance when monotonicity seems reasonable. If we are willing to conclude both ψ_1 and ψ_2 from φ by default, then discovering ψ_2 should not cause us to retract the conclusion ψ_1. For example, if we believe that birds typically fly and birds typically have wings, then we should believe that birds that have wings typically fly.

Note that all the properties of **P** hold if we interpret $\rightarrow$ as $\Rightarrow$, the material conditional. However, they also hold for other interpretations as well, that allow some degree of nonmonotonicity. It is these other interpretations that I focus on here.

If Σ is a finite set of formulas in $\mathcal{L}^{def}$, I write $\Sigma \vdash_{\mathbf{P}} \varphi \rightarrow \psi$ if $\varphi \rightarrow \psi$ can be deduced from Σ using the rules and axioms of **P**, that is, if there is a sequence of formulas in $\mathcal{L}^{def}$, each of which is either an instance of REF (the only axiom in **P**), a formula in Σ, or follows from previous formulas by an application of an inference rule in **P**. Roughly speaking, $\Sigma \vdash_{\mathbf{P}} \varphi \rightarrow \psi$ is equivalent to $\vdash_{\mathbf{P}} \wedge\Sigma \Rightarrow (\varphi \rightarrow \psi)$, where $\wedge\Sigma$ denotes the conjunction of the formulas in Σ. But we cannot write this (yet), since $\wedge\Sigma \Rightarrow (\varphi \rightarrow \psi)$ is not a formula in $\mathcal{L}^{def}$. In Section 7.4, I consider a richer language that allows such formulas.

7.2 Semantics for Defaults

There have been many attempts to give semantics to formulas in $\mathcal{L}^{def}$. The surprising thing is how many of them have ended up being characterized by the axiom system **P**. In this section, I describe a number of these attempts and relate them to each other. I conclude with a semantics based on plausibility measures, which helps explain why **P** characterizes so many different approaches.

7.2.1 Probabilistic Semantics

One compelling approach to giving semantics to defaults is based on the intuition that $\varphi \rightarrow \psi$ should mean that when φ is the case, ψ is very likely. Suppose that we capture our uncertainty in terms of a probability measure μ. Roughly speaking, it seems that this should mean that $\mu(\psi|\varphi)$ is high, or at least that $\mu(\psi|\varphi)$ is significantly higher than $\mu(\neg\psi|\varphi)$. (I am being sloppy here because, strictly speaking, φ and ψ are formulas, not events, so we cannot talk about their probability. This sloppiness will be corrected shortly.)

But how high is high enough? Suppose that we want to interpret formulas in $\mathcal{L}^{def}$ in a measurable probability structure $M = (W, \mu, \pi)$. (In this section, for simplicity, I omit the actual world from the description of a simple structure, since it plays no role in default reasoning and assume that the actual world w is in W; this means $[\![\varphi]\!]_M = [\![\varphi]\!]_M \cap W$.) Suppose that we interpret $\varphi \rightarrow \psi$ as being true in M if $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\psi]\!]_M | [\![\varphi]\!]_M) > \mu([\![\neg\psi]\!]_M | [\![\varphi]\!]_M)$. It is easy to check that, under this interpretation, $M \models \varphi \rightarrow \psi$ if and only if $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\varphi \wedge \psi]\!]_M) > \mu([\![\varphi \wedge \neg\psi]\!]_M)$, or, equivalently, if and only if $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\psi]\!]_M | [\![\varphi]\!]_M) > 1/2$. Moreover, this interpretation satisfies LLE, RW, and REF (Exercise 7.2). However, this interpretation does not necessarily satisfy AND, OR, or CM, as the following example shows.

Example 7.2.1 Consider the structure M_1 from Example 6.1.1. Then

- $\mu([\![\neg p \vee \neg q]\!]_{M_1}) = \mu(\{w_2, w_3, w_4\}) = 0.75$,

- $\mu([\![\neg p]\!]_{M_1}) = \mu(\{w_3, w_4\} = .5$,
- $\mu([\![\neg q]\!]_{M_1}) = \mu(\{w_2, w_4\}) = .45$, and
- $\mu([\![\neg p \wedge \neg q]\!]_{M_1}) = \mu(\{w_4\}) = .2$.

Thus, $\mu([\![\neg p]\!]_{M_1} | [\![\neg p \vee \neg q]\!]_{M_1}) > .5$, $\mu([\![\neg q]\!]_{M_1} | [\![\neg p \vee \neg q]\!]_{M_1}) > .5$, but $\mu([\![\neg p \wedge \neg q]\!]_{M_1} | [\![\neg p \vee \neg q]\!]_{M_1}) < .5$. Thus, we would have $M_1 \models (\neg p \vee \neg q) \rightarrow \neg p$ and $M_1 \models (\neg p \vee \neg q) \rightarrow \neg q$, but $M_1 \not\models (\neg p \vee \neg q) \rightarrow (\neg p \wedge \neg q)$, violating the AND rule. Moreover, $M_1 \not\models \neg p \rightarrow \neg q$, since $\mu([\![\neg q]\!]_{M_1} | [\![\neg p]\!]_{M_1}) < .5$, so we also get a violation of CM. It is also possible to construct a violation of OR in M_1 (Exercise 7.3). ∎

Perhaps the problem is the choice of 1/2. Another thought might be to interpret $\varphi \rightarrow \psi$ as meaning $\mu(\psi|\varphi) > 1 - \epsilon$, for some fixed, small ϵ. This interpretation fares no better than the previous one. Again, it is easy to see that it satisfies LLE, RW, and REF, but not AND, CM, or OR (Exercise 7.4).

The problem here is that no fixed ϵ will work. Once we fix ϵ, it is easy to construct counterexamples. Intuitively, we want $\varphi \rightarrow \psi$ to hold if $\mu(\psi|\varphi)$ is *arbitrarily* small. Perhaps the easiest way to capture this intuition is to take, not one probability measure, but a sequence of them, and require that the probability of ψ given φ go to 1 in this sequence.

Definition 7.2.2 A *probability sequence* on W is just a sequence $(\mu_1, \mu_2, \ldots)$ of probability measures on W. A *(simple) PS structure* is a tuple $(W, (\mu_1, \mu_2, \ldots), \pi)$, where $(\mu_1, \mu_2, \ldots)$ is a probability sequence on W. Let $\mathcal{M}^{ps}$ be the class of all simple PS structures. In a simple PS structure, the truth of a formula of the form $\varphi \rightarrow \psi$ is independent of the world. If $M = (W, (\mu_1, \mu_2, \ldots), \pi)$ is a simple PS structure, define $M \models \varphi \rightarrow \psi$ if $\lim_{k\rightarrow\infty} \mu_k([\![\psi]\!]_M | [\![\varphi]\!]_M) = 1$ (where $\mu_k([\![\psi]\!]_M | [\![\varphi]\!]_M)$ is taken to be 1 if $\mu([\![\varphi]\!]_M) = 0$. ∎

This definition satisfies all the axioms and rules of axiom system **P**. In fact, it is characterized by **P**, as the following theorem shows. Given a collection $\mathcal{M}$ of structures, write $\Sigma \models_{\mathcal{M}} \varphi$ if, whenever $M \models \Sigma$—that is, if $M \models \sigma$ for every formula $\sigma \in \Sigma$—then we also have $M \models \varphi$. Thus, $\Sigma \models_{\mathcal{M}} \varphi$ holds if every structure in $\mathcal{M}$ that satisfies the formulas in Σ also satisfies φ. As a special case, if $\Sigma = \emptyset$, we just get back the standard definition of validity in $\mathcal{M}$.

Theorem 7.2.3 *If Σ is a finite set of formulas in $\mathcal{L}^{def}$, then $\Sigma \vdash_{\mathbf{P}} \varphi \rightarrow \psi$ iff $\Sigma \models_{\mathcal{M}^{ps}} \varphi \rightarrow \psi$.*

Proof Soundness follows from two theorems proved in Section 7.2.3: Theorem 7.2.10, a more general soundess result, which applies to many classes of structures, and Theorem 7.2.11, which shows that Theorem 7.2.10 applies in particular to $\mathcal{M}^{ps}$. However, there is a straightforward direct proof as well. For example, consider the AND rule. Suppose that M is a simple PS structure such that $M \models \varphi \rightarrow \psi_1$ and $M \models \varphi \rightarrow \psi_2$. Then $\lim_{k\to\infty} \mu_k([\![\psi_1]\!]_M | [\![\varphi]\!]_M) = 1$ and $\lim_{k\to\infty} \mu_k([\![\psi_2]\!]_M | [\![\varphi]\!]_M) = 1$. By definition, for all ϵ, there must be some k such that $\mu_k([\![\psi_1]\!]_M | [\![\varphi]\!]_M) \geq 1-\epsilon$ and $\mu_k([\![\psi_2]\!]_M | [\![\varphi]\!]_M) \geq 1-\epsilon$. By the inclusion-exclusion rule (3.3),

$$\begin{array}{ll} & \mu([\![\psi_1 \wedge \psi_2]\!]_M | [\![\varphi]\!]_M) \\ = & \mu([\![\psi_1]\!]_M | [\![\varphi]\!]_M) + \mu([\![\psi_2]\!]_M | [\![\varphi]\!]_M) - \mu([\![\psi_1 \vee \psi_2]\!]_M | [\![\varphi]\!]_M) \\ \geq & (1-\epsilon) + (1-\epsilon) - 1 \\ = & 1 - 2\epsilon. \end{array}$$

Thus, $\lim_{k\to\infty} \mu_k([\![\psi_1 \wedge \psi_2]\!]_M | [\![\varphi]\!]_M) = 1$, so $M \models \varphi \rightarrow (\psi_1 \wedge \psi_2)$, as desired. The proof that OR and CM also hold in PS structures is equally straightforward (Exercise 7.5).

Completeness follows from two other theorems proved in Section 7.2.3: Theorem 7.2.13, a more general completeness result, which applies to many classes of structures, and Theorem 7.2.14, which shows that Theorem 7.2.13 applies in particular to $\mathcal{M}^{ps}$. ∎

While PS structures are a technically useful tool for capturing default reasoning, it is not so clear where the sequence of probabilities is coming from. Under what circumstances would an agent use a sequence of probability measures to describe her uncertainty? In Chapter **??**, we shall see a context in which such sequences arise naturally.

7.2.2 Using Possibility Measures, Ranking Functions, and Preference Orders

Taking $\varphi \rightarrow \psi$ to hold iff $\mu(\psi|\varphi) > \mu(\neg\psi|\varphi)$ does not work, in the sense that it does not give us some properties that seem important in the context of default reasoning. Belief functions and lower probabilities fare no better than probability measures; again, they satisfy LLE, RW, and REF, but not AND, OR, or CM. Indeed, since probability measures are a special case of belief functions and sets of probability measures, the counterexamples in the previous section apply without change. We could use sequences of belief functions or sequences of sets of probability measures, just as in PS structures. This in fact would result in the desired properties, although I do not go through the exercise of showing that here. More interestingly, possibility measures, ranking functions, and preference orders have the desired properties without the need to consider sequences.

The formal definitions are just the obvious analogue of the definitions that we considered in the case of probability:

- If $M = (W, \mathrm{Poss}, \mu)$ is a simple possibility structure, then

  $$M \models \varphi \to \psi \text{ iff } \mathrm{Poss}([\![\varphi]\!]_M) = 0 \text{ or } \mathrm{Poss}([\![\varphi \wedge \psi]\!]_M) > \mathrm{Poss}([\![\varphi \wedge \neg\psi]\!]_M).$$

- If $M = (W, \kappa, \mu)$ is a simple ranking structure, then

  $$M \models \varphi \to \psi \text{ iff } \kappa([\![\varphi]\!]_M) = \infty \text{ or } \kappa([\![\varphi \wedge \psi]\!]_M) < \kappa([\![\varphi \wedge \neg\psi]\!]_M).$$

- Finally, if $M = (W, \succeq, \pi)$ is a simple preferential structure, then

  $$M \models \varphi \to \psi \text{ iff } [\![\varphi]\!]_M = \emptyset \text{ or } [\![\varphi \wedge \psi]\!]_M \succ^s [\![\varphi \wedge \neg\psi]\!]_M.$$

Theorem 7.2.4 *Let Σ be a finite set of formulas in $\mathcal{L}^{def}$. The following are equivalent:*

(a) $\Sigma \vdash_{\mathbf{P}} \varphi \to \psi$,

(b) $\Sigma \models_{\mathcal{M}^{poss}} \varphi \to \psi$,

(c) $\Sigma \models_{\mathcal{M}^{rank}} \varphi \to \psi$,

(d) $\Sigma \models_{\mathcal{M}^{pref}} \varphi \to \psi$,

(e) $\Sigma \models_{\mathcal{M}^{tot}} \varphi \to \psi$.

Proof Soundness follows from Theorem 7.2.10. However, it is again straightforward to provide a direct proof. I show that the AND rule is sound for possibility measures here, leaving the remainder of the soundness proof as an exercise (Exercise 7.7). Suppose that $M = (W, \mathrm{Poss}, \pi)$ is a possibility structure, $M \models \varphi \to \psi_1$, and $M \models \varphi \to \psi_2$. If $\mathrm{Poss}([\![\varphi]\!]_M) = 0$, then it is immediate that $M \models \varphi \to (\psi_1 \wedge \psi_2)$. So suppose that $\mathrm{Poss}([\![\varphi]\!]_M) > 0$. Let $U_j = [\![\varphi \wedge \psi_j]\!]_M$ and $V_j = [\![\varphi \wedge \neg\psi_j]\!]_M$ for $j = 1, 2$. Note that $U_1 \cup V_1 = U_2 \cup V_2 = [\![\varphi]\!]_M$. Suppose that $\mathrm{Poss}(U_1 \cap U_2) = \alpha$, $\mathrm{Poss}(U_1 \cap V_2) = \beta$, $\mathrm{Poss}(V_1 \cap U_2) = \gamma$, and $\mathrm{Poss}(V_1 \cap V_2) = \delta$. Since $(U_1 \cap U_2) \cup (U_1 \cap V_2) = U_1$, it must be the case that $\mathrm{Poss}(U_1) = \max(\alpha, \beta)$. Similarly, $\mathrm{Poss}(V_1) = \max(\gamma, \delta)$, $\mathrm{Poss}(U_2) = \max(\alpha, \gamma)$, and $\mathrm{Poss}(V_2) = \max(\beta, \delta)$. Since $\mathrm{Poss}(U_j) > \mathrm{Poss}(V_j)$ for $j = 1, 2$, $\max(\alpha, \beta) > \max(\gamma, \delta)$ and $\max(\alpha, \gamma) > \max(\beta, \delta)$. It easily follows that $\alpha > \max(\beta, \gamma, \delta)$ (Exercise 7.7). Thus, $\mathrm{Poss}(U_1 \cap U_2) > \mathrm{Poss}(V_1 \cup V_2)$, which means that

$\mathrm{Poss}([\![\varphi \wedge \psi_1 \wedge \psi_2]\!]_M > \mathrm{Poss}([\![\varphi \wedge \neg(\psi_1 \wedge \psi_2)]\!]_M$. Thus, $M \models \varphi \rightarrow (\psi_1 \wedge \psi_2)$, as desired.

Again, completeness follows from Theorems 7.2.13 and 7.2.14. ∎

For readers familiar with nonstandard analysis, this result may clarify the relationship between default reasoning and probability. Recall that one interpretation of ranking functions is that they represent order-of-magnitude reasoning. That is, given a ranking function κ, there is a probability measure μ_κ such that if $\kappa(U) = k$ iff $\mu_\kappa(U)$ is roughly ϵ^k for some infinitesimal ϵ. With this interpretation, $\kappa([\![\varphi \wedge \psi]\!]_M) > \kappa([\![\varphi \wedge \neg\psi]\!]_M)$ if $\mu([\![\psi]\!]_M | [\![\varphi]\!]_M) > 1 - \epsilon$. Thus, although Section 7.2.1 shows that we cannot give semantics to defaults in this way using a standard ϵ (AND, OR, and CM all fail), this approach works if we use an infinitesimal ϵ.

Theorem 7.2.4 provides further evidence that $\mathcal{L}^{def}$ is a relatively weak language. For example, it cannot distinguish totally-ordered preferential structures for arbitrary preferential structures; the same axioms (in $\mathcal{L}^{def}$!) characterize both. Roughly speaking, we can think of **P** as the "footprint" of default reasoning on the language $\mathcal{L}^{def}$. Since $\mathcal{L}^{def}$ is not a very expressive language, the footprints of the various semantic approaches are indistinguishable. By way of contrast, the language $\mathcal{L}_n^{\gg}$ (which can express modularity, for example) can distinguish (some of) these approaches, as can the conditional logic $\mathcal{L}_n^{\rightarrow}$ that will be defined in Section 7.4.

The approaches for giving semantics to $\mathcal{L}^{def}$ that we have considered so far take the view that $\varphi \rightarrow \psi$ means "if φ then ψ is very likely". However, there is another semantics that focuses more on interpreting $\varphi \rightarrow \psi$ as "if φ, then normally ψ". This is perhaps best seen in the context of preferential structures. Suppose that we view the $\succeq$ order as defining normality. That is, $w \succeq w'$ means that w is more "normal" than w'. For example, a world where *bird* $\wedge$ *fly* holds might be viewed as more normal than one where *bird* $\wedge \neg$*fly* holds. Given a preferential structure $M = (W, \succeq, \pi)$ and a set $U \subseteq W$, define $\mathrm{best}_M(U)$ to be the most normal worlds in U (according to the ordering $\succeq$ in M). Since $\succeq$ in general is a partial order, the formal definition is

$$\mathrm{best}_M(U) = \{w \in U : \text{ for all } w' \in U,\ w' \not\succ w\}.$$

Define a new operator $\rightarrow'$ in simple preferential structures as follows:

$$M \models \varphi \rightarrow' \psi \text{ iff } \mathrm{best}_M([\![\varphi]\!]_M) \subseteq [\![\psi]\!]_M.$$

The intuition behind this definition should be clear: $\varphi \rightarrow' \psi$ holds in M if, in the most normal worlds where φ is true, ψ is also true. By way of contrast, notice that $M \models \varphi \Rightarrow \psi$ iff $[\![\varphi]\!]_M \subseteq [\![\psi]\!]_M$ (Exercise 7.8). Thus, for $\varphi \Rightarrow \psi$ to be valid in M, ψ must hold in *all* worlds where φ holds; for $\varphi \rightarrow \psi$ to be valid in M, ψ must just hold in the most normal worlds where φ holds.

Example 7.2.5 Normally, a bird flies and hence is not a penguin; normally, penguins do not fly. This property holds in the simple preferential structure $M_2 = (\{w_1, w_2, w_3, w_4\}, \succeq, \pi)$, where π is such that

- $(M_2, w_1) \models \mathit{bird} \wedge \mathit{fly} \wedge \neg \mathit{penguin}$,
- $(M_2, w_2) \models \mathit{bird} \wedge \neg \mathit{fly} \wedge \mathit{penguin}$,
- $(M_2, w_3) \models \mathit{bird} \wedge \mathit{fly} \wedge \mathit{penguin}$,
- $(M_2, w_4) \models \mathit{bird} \wedge \neg \mathit{fly} \wedge \neg \mathit{penguin}$,

and $\succeq$ is such that $w_1 \succ w_2 \succ w_3$, $w_1 \succ w_4$, and w_4 is incomparable to both w_2 and w_3. Since $\mathrm{best}_{M_2}([\![\mathit{bird}]\!]_{M_2}) = \{w_1\} \subseteq [\![\mathit{fly}]\!]_{M_2}$ and $\mathrm{best}_{M_2}([\![\mathit{bird} \wedge \mathit{penguin}]\!]_{M_2}) = \{w_2\} \subseteq [\![\neg \mathit{fly}]\!]_{M_2}$, it follows that

$$M_2 \models (\mathit{bird} \rightarrow' \mathit{fly}) \wedge (\mathit{bird} \wedge \mathit{penguin} \rightarrow' \neg \mathit{fly}),$$

as we would hope and expect. ∎

Although $\rightarrow$ and $\rightarrow'$ may seem on the surface to be quite different, the following theorem shows that they are in fact equivalent.

Theorem 7.2.6 *In every simple preferential structure M,*

$$M \models \varphi \rightarrow \psi \text{ iff } M \models \varphi \rightarrow' \psi.$$

Proof See Exercise 7.9. ∎

Of course, since $\rightarrow$ and $\rightarrow'$ are equivalent, it follows that this semantics for $\rightarrow'$ is also characterized by **P**.

7.2.3 Using Plausibility Measures

Why should so many approaches to giving semantics to defaults be characterized by axiom system **P**? Although in three of the cases we considered in the previous section we essentially took $\varphi \rightarrow \psi$ to mean "$\varphi \wedge \psi$ is more likely than $\varphi \wedge \neg\psi$", we used quite different notions of "more likely than". Why does this approach work for possibility measures, ranking functions, and preference orders, but not for probability measures or belief functions? Plausibility measures help to explain what is going on here. The lack of structure in the plausibility framework allows us to understand exactly what structure is needed to get the properties we want.

The definition of $\rightarrow$ in plausibility structures is just what we would expect. If $M = (W, \succeq, \mathrm{Pl})$ is a simple plausibility structure, define

$$M \models \varphi \rightarrow \psi \text{ iff } \mathrm{Pl}([\![\varphi]\!]_M) = \bot \text{ or } \mathrm{Pl}([\![\varphi \wedge \psi]\!]_M) > \mathrm{Pl}([\![\varphi \wedge \neg\psi]\!]_M).$$

Just as with all the other representations of uncertainty, LLE, RW, and REF hold for this definition of uncertainty.

Lemma 7.2.7 *All plausibility structures satisfy LLE, RW, and REF.*

Proof See Exercise 7.10. It is worth noting that REF holds in plausibility structures because of Pl2 (recall that Pl2 says that $\mathrm{Pl}(\emptyset) = \bot$ and, by assumption, $\bot$ is the minimum element with respect to $\leq$) and RW holds because of Pl3 (recall that Pl3 says that $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$ if $U \subseteq V$). ∎

AND, OR, or CM do not hold in general in plausibility structures. Indeed, since probability is a special case of plausibility, the counterexample given earlier in the case of probability applies here with no change. This leads to an obvious question: What properties of plausibility would force AND, OR, and CM to hold? The proof of Lemma 7.2.7 already shows the connections between the properties of plausibility measures and interesting properties of defaults; for example, Pl3 gives us RW. It turns out that we can actually give quite an elegant characterization of the properties of plausibility measures required to model default reasoning.

Let's start with the AND rule. "Reverse engineering" shows that the following (admittedly, somewhat ugly) property is just what we need:

Pl4$'$. For all sets U, V_1, and V_2, if $\mathrm{Pl}(U \cap V_1) > \mathrm{Pl}(U \cap \overline{V_1})$ and $\mathrm{Pl}(U \cap V_2) > \mathrm{Pl}(U \cap \overline{V_2})$, then $\mathrm{Pl}(U \cap V_1 \cap V_2) > \mathrm{Pl}(U \cap \overline{(V_1 \cap V_2)})$.

More precisely, it is easy to prove the following lemma.

Lemma 7.2.8 *If $M = (W, \mathrm{Pl}, \pi)$ is a simple plausibility structure such that* Pl *satisfies Pl4$'$, then the AND rule holds in M.*

Proof Suppose that $M \models \varphi \rightarrow \psi_1$ and $M \models \varphi \rightarrow \psi_2$. If $\mathrm{Pl}([\![\varphi]\!]_M) = \bot$ then, by definition, $M \models \varphi \rightarrow (\psi_1 \wedge \psi_2)$. On the other hand, if $\mathrm{Pl}([\![\varphi]\!]_M) \neq \bot$, then $\mathrm{Pl}([\![\varphi \wedge \psi_1]\!]_M) > \mathrm{Pl}([\![\varphi \wedge \neg\psi_1]\!]_M)$ and $\mathrm{Pl}([\![\varphi \wedge \psi_2]\!]_M) > \mathrm{Pl}([\![\varphi \wedge \neg\psi_2]\!]_M)$. From Pl4$'$, it follows that

$$\mathrm{Pl}([\![\varphi \wedge \psi_1 \wedge \psi_2]\!]_M) > \mathrm{Pl}([\![\varphi \wedge \neg(\psi_1 \wedge \psi_2)]\!]_M).$$

Thus, $M \models \varphi \rightarrow (\psi_1 \wedge \psi_2)$, as desired. ∎

In the presence of Pl3, there is a much simpler property that suffices for the AND rule. It is a variant of a property that we have seen before in the context of preference orders: the qualitative property (see Section **??**). But now it must hold only for disjoint sets.

Pl4. If U_1, U_2, and U_3 are pairwise disjoint sets, $\mathrm{Pl}(U_1 \cup U_2) > \mathrm{Pl}(U_3)$, and $\mathrm{Pl}(U_1 \cup U_3) > \mathrm{Pl}(U_2)$, then $\mathrm{Pl}(U_1) > \mathrm{Pl}(U_2 \cup U_3)$.

As the following lemma shows, in the presence of Pl3, Pl4 is equivalent to Pl4′.

Proposition 7.2.9 *A plausibility measure satisfies Pl4 if and only if it satisfies Pl4′.*

Proof See Exercise 7.11. ∎

It follows from Lemma 7.2.7 and Proposition 7.2.9 that the disjoint qualitative property—Pl4—is exactly what we need to force the AND rule. In a precise sense, it is also necessary for the AND rule (Exercise 7.12). We can use plausibility measures that satisfy Pl4 to define a notion of belief: an agent believes φ precisely if φ is more plausible than $\neg\varphi$. This notion of belief is characterized by KD45, the standard axioms of belief (Exercise 7.13). (See Section **??** for further discussion of this notion of belief.)

Somewhat surprisingly, Pl4 is also just what we need to get CM and the non-vacuous case of OR. More precisely, if $M = (W, \mathrm{Pl}, \pi)$ is a simple plausibility structure and M satisfies Pl4 (and Pl3, of course), then M satisfies AND and CM, and if $M \models \varphi_1 \rightarrow \psi$, $M \models \varphi_2 \rightarrow \psi$, and either $\mathrm{Pl}([\![\varphi_1]\!]_M) \neq \perp$ or $\mathrm{Pl}([\![\varphi_2]\!]_M) \neq \perp$, then $M \models (\varphi_1 \vee \varphi_2) \rightarrow \psi$. To deal with the vacuous case of OR (where both $\mathrm{Pl}([\![\varphi_1]\!]_M) = \perp$ and $\mathrm{Pl}([\![\varphi_2]\!]_M) = \perp$), we need one more (rather innocuous) property:

Pl5. If $\mathrm{Pl}(U) = \mathrm{Pl}(V) = \perp$, then $\mathrm{Pl}(U \cup V) = \perp$.

Note that Pl5 holds for many, but not all, the notions of uncertainty we have considered so far, when viewed as plausibility measures. For example, if Poss is a possibility measure, then certainly $\mathrm{Poss}(U) = \mathrm{Poss}(V) = 0$ implies $\mathrm{Poss}(U \cup V) = 0$. The same is true for probability. On the other, it is not true of belief functions or inner measures.

A plausibility measure is said to be *qualitative* if it satisfies Pl4 and Pl5 (as well as Pl1–3). A simple plausibility structure $M = (W, \mathrm{Pl}, \pi)$ is qualitative if Pl is. Let $\mathcal{M}^{qual}$ be the class of all simple qualitative plausibility structures.

Theorem 7.2.10 *If Σ is a finite set of formulas in $\mathcal{L}^{def}$, then*

$$\Sigma \vdash_{\mathbf{P}} \varphi \rightarrow \psi \text{ iff } \Sigma \models_{\mathcal{M}^{qual}} \varphi \rightarrow \psi.$$

Proof The soundness of LLE, RW, and CM follows from Lemma 7.2.7; the soundness of AND follows from Proposition 7.2.9. The soundness of CM and OR is left to Exercise 7.14. Again, completeness follows from Theorems 7.2.13 and 7.2.14. ∎

Lemma 2.3.3 and Exercise 3.22 show that simple possibility structures, ranking structures, and preferential structures can all be viewed as simple qualitative plausibility structures. Indeed, in the remainder of this chapter, I shall be somewhat sloppy and view $\mathcal{M}^{poss}$, $\mathcal{M}^{rank}$, $\mathcal{M}^{pref}$, and $\mathcal{M}^{tot}$ as being contained in $\mathcal{M}^{qual}$. It follows immediately from Theorem 7.2.10 that **P** is sound in $\mathcal{M}^{poss}$, $\mathcal{M}^{rank}$, $\mathcal{M}^{pref}$, and $\mathcal{M}^{tot}$, since these can all be viewed as subclasses of $\mathcal{M}^{qual}$. Properties that are valid in $\mathcal{M}^{qual}$ are bound to be valid in all of its subclasses. It is because possibility measures, ranking functions, and preference orders can all be viewed as plausibility measures that satisfy Pl4 and Pl5 that, when used to give semantics to defaults, they satisfy all the properties in **P**. By way of contrast, probability measures do not satisfy Pl4; hence the obvious way of using them to give semantics to defaults does not satisfy all the properties of **P**.

What about probability sequences? A simple PS structure can also be viewed as a plausibility structure. Given a simple PS structure $M = (W, (\mu_1, \mu_2, \ldots), \pi)$, define a simple plausibility structure $M_{PS} = (W, \mathrm{Pl}, \pi)$ such that

$$\mathrm{Pl}(U) \leq \mathrm{Pl}(V) \text{ if and only if } \lim_{i\rightarrow\infty} \mu_i(V|U \cup V) = 1. \qquad (7.1)$$

It is easy to check that there is a plausibility measure with this property (Exercise 7.15); that is, there is a set D of plausibility values and a mapping $\mathrm{Pl} : 2^W \rightarrow D$ satisfying Pl3 with property (7.1). Moreover, M_{PS} is qualitative and satisfies the same defaults as M. More precisely,

Theorem 7.2.11 *Suppose that $M \in \mathcal{M}^{ps}$. Then $M_{PS} \in \mathcal{M}^{qual}$ and $M \models \varphi \rightarrow \psi$ iff $M_{PS} \models \varphi \rightarrow \psi$.*

Proof See Exercise 7.15. ∎

Thus, Theorem 7.2.10 also explains why structures in $\mathcal{M}^{ps}$ are characterized by **P**; it is because they too satisfy Pl4 and Pl5. Indeed, there is a sense in which Pl4 and Pl5 completely characterize the plausibility structures that satisfy **P**. Roughly speaking, if **P** is sound for a collection $\mathcal{M}$

of plausibility structures, then all the structures in $\mathcal{M}$ must be qualitative. (See Exercise 7.16 for details.)

To get **P** to be sound for a class $\mathcal{M}$ of structures, we have to make sure that $\mathcal{M}$ does not have "too many" structures, in particular, no structures that are not qualitative. For completeness, we have the opposite problem. We have to ensure that $\mathcal{M}$ contains "enough" structures; if $\mathcal{M}$ has too few structures, there may be additional properties valid in $\mathcal{M}$. In particular, if $\Sigma \not\vdash_{\mathbf{P}} \varphi \rightarrow \psi$, we want to ensure that there is a plausibility structure $M \in \mathcal{M}$ such that $M \models \Sigma$ and yet $M \not\models \varphi \rightarrow \psi$. The following weak condition suffices to ensure that $\mathcal{M}$ has enough structures in this sense.

Definition 7.2.12 A class $\mathcal{M}$ of simple plausibility structures is *rich* if, for every collection $\varphi_1, \ldots, \varphi_k$, $k > 1$, of mutually exclusive and consistent propositional formulas, there is a plausibility structure $M = (W, \mathrm{Pl}, \pi) \in \mathcal{M}$ such that

$$\mathrm{Pl}([\![\varphi_1]\!]_M) > \mathrm{Pl}([\![\varphi_2]\!]_M) > \ldots > \mathrm{Pl}([\![\varphi_k]\!]_M) = \bot. \blacksquare$$

The richness requirement is quite mild. It says that $\mathcal{M}$ does not place *a priori* constraints on the relative plausibilities of a collection of disjoint sets. If we think in terms of probability measures, it just says that given any collection of disjoint sets $A_1, \ldots, A_k$, there is a probability measure μ such that $\mu(A_1) > \ldots > \mu(A_k) = 0$. Every representation method consider thus far (viewed as a collection of plausibility structures) can easily be shown to satisfy this richness condition.

Theorem 7.2.13 *Each of $\mathcal{M}^{ps}$, $\mathcal{M}^{poss}$, $\mathcal{M}^{rank}$, $\mathcal{M}^{pref}$, $\mathcal{M}^{tot}$, and $\mathcal{M}^{qual}$ is rich.*

Proof This is almost immediate from the definitions; the details are left to Exercise 7.17. Note that I am viewing $\mathcal{M}^{ps}$, $\mathcal{M}^{poss}$, $\mathcal{M}^{rank}$, $\mathcal{M}^{pref}$, and $\mathcal{M}^{tot}$ as subsets of $\mathcal{M}^{qual}$ here, so that the richness condition as stated applies to them. $\blacksquare$

More importantly, richness is a necessary and sufficient condition to ensure that the axiom system **P** is complete.

Theorem 7.2.14 *A set $\mathcal{M}$ of qualitative plausibility structures is rich if and only if $\Sigma \models_{\mathcal{M}} \varphi \rightarrow \psi$ implies $\Sigma \vdash_{\mathbf{P}} \varphi \rightarrow \psi$ for all finite sets Σ of formulas in $\mathcal{L}^{def}$ and defaults $\varphi \rightarrow \psi$.*

Proof The proof that richness is necessary for completeness is not hard; see Exercise 7.18. The proof that richness is sufficient for completeness is sketched (with numerous hints) in Exercise 7.19. $\blacksquare$

To summarize, the results of this section say that (for a method of representing uncertainty that can be associated with a subclass of plausibility structures), **P** is sound as long as the representation method satisfies Pl4 and Pl5 (the significant property here is disjoint qualitativeness, since Pl5 holds for all standard formalisms), and **P** is complete if the associated class of plausibility structures is rich, again, a rather mild restriction.

7.3 Beyond System P

As I said before, the axiom system **P** has been viewed as characterizing the "conservative core" of default reasoning. Is there a reasonable, principled way of going beyond System **P** to obtain inferences that do not follow from treating $\rightarrow$ as material implication? In particular, is it possible to ignore irrelevant information and to allow subclasses to inherit properties from superclasses? The following examples give a sense of the issues involved.

Example 7.3.1 If birds typically fly and penguins typically do not fly (although penguins are birds), it seems reasonable to conclude that red penguins do not fly. Thus, if we let

$$\Sigma_1 = \{\mathit{bird} \rightarrow \mathit{fly}, \mathit{penguin} \rightarrow \neg \mathit{fly}, \mathit{penguin} \rightarrow \mathit{bird}\},$$

we might hope that from Σ_1 we could conclude $\mathit{penguin} \land \mathit{red} \rightarrow \neg\mathit{fly}$. (Notice that Σ_1 says only that penguins are typically birds, rather than all penguins are birds. This is because we cannot express universal statements in $\mathcal{L}^{def}$. The point here could be made equally well if we replaced $\mathit{penguin} \rightarrow \mathit{bird}$ by $\mathit{penguin} \Rightarrow \mathit{bird}$ in Σ_1.) However, $\Sigma_1 \not\vdash_{\mathbf{P}} \mathit{penguin} \land \mathit{red} \rightarrow \neg fly$ (Exercise 7.20(a)). Intuitively, this is because it is conceivable that although penguins typically do not fly, red penguins might be unusual penguins, and so might in fact fly. That is, no matter how much we would like to treat redness as irrelevant, it in fact might be relevant to whether or not penguins fly. The "conservative core" does not let us conclude that red penguins do not fly because of this possibility.

Note that if Σ_1' is the result of replacing all occurrences of $\rightarrow$ in Σ_1 by $\Rightarrow$, then $M \models \Sigma_1'$ only if $[\![\mathit{penguin}]\!]_M = \emptyset$. Hence, $\Sigma_1' \models \neg\mathit{penguin}$. On the other hand, there is a structure M in, for example, $\mathcal{M}^{rank}$ such that $M \models \Sigma_1$, $M \models (\mathit{penguin} \land \mathit{red}) \rightarrow \neg\mathit{fly}$, and $[\![\mathit{penguin} \land \mathit{red}]\!]_M \neq \emptyset$ (Exercise 7.21). (Of course, there are also structures in $\mathcal{M}_n^{pref}$, $\mathcal{M}_n^{poss}$, $\mathcal{M}^{ps}$, or $\mathcal{M}_n^{qual}$ with these properties.) This shows that we do gain something by considering $\rightarrow$ rather than $\Rightarrow$, although that "something" cannot be expressed in the logic.

Now suppose that we add to Σ_1 the default "birds are typically warm-blooded". (In fact, all birds are warm-blooded, just as all penguins are birds, but the default suffices for my purposes.) Let

$$\Sigma_2 = \Sigma_1 \cup \{\mathit{bird} \rightarrow \mathit{warm\text{-}blooded}\}.$$

We might hope to get what has been called in the literature *exceptional subclass inheritance*: although penguins are an exceptional subclass of birds (in that they do not fly, although birds typically do), we would hope that they would still inherit the property of warm-bloodedness from birds. This does not follow for the material conditional, nor can we prove it for $\rightarrow$ in system **P** (since all the properties of **P** hold for the material conditional). That is, $\Sigma_2 \not\vdash_{\mathbf{P}} (\mathit{penguin} \land \mathit{bird}) \rightarrow \mathit{warm\text{-}blooded}$ (Exercise 7.20(b)). Although we might want to think of "penguin-ness" as being irrelevant to being warm-blooded, it nevertheless may be, which intuitively is why this conclusion cannot be drawn in **P**. If penguins are atypical bird in one respect, perhaps they are atypical birds in other respects.

But suppose that $\Sigma_3 = \Sigma_1 \cup \{\mathit{yellow} \rightarrow \mathit{easy\text{-}to\text{-}see}\}$: yellow things are easy to see. Now we might expect that yellow penguins are easy to see. However, $\Sigma_3 \not\vdash_{\mathbf{P}} \mathit{penguin} \land \mathit{yellow} \rightarrow \mathit{easy\text{-}to\text{-}see}$ (Exercise 7.20(c)). Note that this type of exceptional subclass inheritance is not quite that with Σ_2. Whereas penguins are atypical birds, there is no reason to expect them to be atypical yellow objects. Nevertheless, **P** will not let us conclude that they inherit the property of being easy to see.

One last example. Suppose that $\Sigma_4 = \Sigma_2 \cup \{\mathit{robin} \rightarrow \mathit{bird}\}$. Although we cannot conclude (in **P**) from Σ_2 that penguins are typically warmblooded, we might hope to conclude from Σ_4 that robins are warmblooded. After all, as far as Σ_4 is concerned, robins are completely unexceptional birds, and birds are typically warmblooded. Unfortunately, it is not hard to show that $\Sigma_4 \not\vdash_{\mathbf{P}} \mathit{robin} \rightarrow \mathit{warm\text{-}blooded}$, nor does it help to replace *robin* by $\mathit{robin} \land \mathit{bird}$ (Exercise 7.20(d)). ∎

In light of these examples, it is perhaps not surprising that there has been a great deal of effort devoted to finding principled methods of going beyond **P**. However, it has been difficult to find one that gives all and only the "reasonable" inferences. Indeed, it is difficult to characterize exactly what the reasonable inferences are. The results of the previous section point to one source of the difficulties. We might hope to find (1) an axiom system $\mathbf{P}^+$ that is stronger than **P** (in the sense that everything provable in **P** is also provable in $\mathbf{P}^+$, and $\mathbf{P}^+$ allows us to make some additional "reasonable" inferences) and (2) a class $\mathcal{M}$ of structures with respect to which $\mathbf{P}^+$ is sound and complete. If we can view the structures in $\mathcal{M}$ as plausibility structures, then they must all satisfy Pl4 and Pl5, to guarantee that **P** is sound with respect to $\mathcal{M}$. However, $\mathcal{M}$ cannot be rich, for then **P** would also be complete; we would not get any additional inferences.

Richness is not a very strong assumption. One way of avoiding it that has been taken in the literature is the following. Given a class $\mathcal{M}$ of structures, recall that $\Sigma \models_{\mathcal{M}} \varphi$ if $M \models \Sigma$ implies $M \models \varphi$ for every structure $M \in \mathcal{M}$. Rather than considering *every* structure that satisfies Σ, the idea is to consider a "preferred" structure that satisfies Σ, and check whether φ holds in that preferred structure. Essentially, this approach takes the idea used in preferential structures of considering the most preferred worlds and lifts it to the level of structures.

Here are two examples of how this general approach works. The first uses ranking structures (which are, after all, just as a special case of plausibility structures). Starting with a fixed finite set Φ of primitive propositions, let W_Φ consist of all the truth assignments to the primitive propositions in Φ. Let $\mathcal{M}_\Phi^{rank}$ consist of all simple ranking structures of the form $(W_\Phi, \kappa, \pi_\Phi)$, where $\pi_\Phi(w) = w$ (this makes sense since the worlds in W_Φ are truth assignments). Define a partial order $\succeq$ on ranking functions on W_Φ by defining $\kappa_1 \succeq \kappa_2$ if $\kappa_1(w) \leq \kappa_2(w)$ for all $w \in W_\Phi$. Thus, κ_1 is preferred to κ_2 if every world is no more surprising according to κ_1 that it is according to κ_2. We can lift $\succeq$ to a partial order on ranking structures in $\mathcal{M}_\Phi^{rank}$ by defining $(W_\varphi, \kappa_1, \pi_\Phi) \succeq (W_\varphi, \kappa_2, \pi_\Phi)$ if $\kappa_1 \succeq \kappa_2$.

Given a finite set Σ of formulas in $\mathcal{L}^{def}(\Phi)$, let $\mathcal{M}_\Sigma^{rank}$ consist of all the ranking structures in $\mathcal{M}_\Phi^{rank}$ that satisfy all the defaults in Σ. Although $\succeq$ is a partial order, it turns out that if $\mathcal{M}_\Sigma^{rank} \neq \emptyset$, then there is a unique structure $M_\Sigma \in \mathcal{M}_\Sigma^{rank}$ that is most preferred. That is, $M_\Sigma \succeq M$ for all $M \in \mathcal{M}_\Sigma^{rank}$ (Exercise 7.22). Intuitively, M_Σ makes worlds as unsurprising as possible, while still satisfying the defaults in Σ. For $\varphi \in \mathcal{L}^{def}$, we then define $\Sigma \mathrel{|\!\!\approx}^Z \varphi$ if either $\mathcal{M}_\Sigma = \emptyset$ or $M_\Sigma \models \varphi$. (The superscript Z is there because, historically, this approach has been called *System Z*.) That is, $\Sigma \mathrel{|\!\!\approx}^Z \varphi$ if φ is true in the most preferred structure of all the structures satisfying Σ.

Since **P** is sound in ranking structures, we certainly have $\Sigma \mathrel{|\!\!\approx} \varphi$ if $\Sigma \vdash_{\mathbf{P}} \varphi$. But the System Z approach has some additional desirable properties. For example, as desired, red penguins continue not to fly; that is, in the notation of Example 7.3.1, $\Sigma_1 \mathrel{|\!\!\approx}^Z$ *penguin* $\wedge$ *red* $\rightarrow \neg$*fly*. More generally, System Z can ignore "irrelevant" attributes and deals well with some of the other issues raised by Example 7.3.1, as the following lemma shows.

Lemma 7.3.2 *Let* $\Sigma_a = \{\varphi_1 \rightarrow \varphi_2, \varphi_2 \rightarrow \varphi_3\}$ *and let* $\Sigma_b = \{\varphi_1 \rightarrow \varphi_2, \varphi_2 \rightarrow \varphi_3, \varphi_1 \rightarrow \neg\varphi_3, \varphi_1 \rightarrow \varphi_4\}$.

(a) $\Sigma_a \mathrel{|\!\!\approx}^Z \varphi_1 \wedge \psi \rightarrow \varphi_3$ *if* $\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \psi$ *is satisfiable.*

(b) $\Sigma_b \approx\!\!|^{Z} \varphi_1 \wedge \psi \rightarrow \neg\varphi_3 \wedge \varphi_4$ *if if* $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4 \wedge \psi$ *is satisfiable.*

Proof For part (a), suppose that $\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \psi$ is satisfiable. Then $\mathcal{M}^{rank}_{\Sigma_a} \neq \emptyset$, since both defaults in Σ_a are satisfied in a structure where all worlds in which $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ is true have rank 0 and all others have rank 1. Suppose that $M_{\Sigma_a} = (W, \kappa_1, \pi)$. In M_{Σ_a}, it is easy to see that all worlds satisfying $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ have rank 0 and all worlds satisfying $\varphi_1 \wedge \neg\varphi_2$ or $\varphi_2 \wedge \neg\varphi_3$ have rank 1, since they violate a default in Σ_a (Exercise 7.23(a)). Since, by assumption, $\varphi_1 \wedge \neg\varphi_2 \wedge \varphi_3 \wedge \psi$ is satisfiable, there is a world of rank 0 satisfying this formula. Moreover, since any world satisfying $\varphi_1 \wedge \neg\varphi_3 \wedge \psi$ must satisfy either $\varphi_1 \wedge \neg\varphi_2$ or $\varphi_2 \wedge \neg\varphi_3$, it follows that $\kappa_1([\![\varphi_1 \wedge \neg\varphi_3 \wedge \psi]\!]_{M_{\Sigma_a}}) \leq 1$. Thus, $\kappa_1([\![\varphi_1 \wedge \varphi_3 \wedge \psi]\!]_{M_{\Sigma_a}}) < \kappa_1([\![\varphi_1 \wedge \neg\varphi_3 \wedge \psi]\!]_{M_{\Sigma_a}})$, so $M_{\Sigma_a} \models \varphi_1 \wedge \psi \rightarrow \varphi_3$.

For part (b), if $\mathcal{M}^{rank}_{\Sigma_b} = \emptyset$, then the result is trivially true. Otherwise, suppose that $M_{\Sigma_b} = (W, \kappa_2, \pi)$. It is easy to see that all worlds in M_{Σ_b} satisfying $\neg\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ have rank 0, all worlds satisfying $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4$ have rank 1, and all worlds satisfying $\varphi_1 \wedge \varphi_3$ or $\varphi_1 \wedge \neg\varphi_4$ have rank 2 (Exercise 7.23(b)). Since, by assumption, $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4 \wedge \psi$ is satisfiable, there is a world of rank 1 satisfying this formula. It follows that $\kappa_2([\![\varphi_1 \wedge \psi \wedge \neg\varphi_3 \wedge \varphi_4]\!]_{M_{\Sigma_b}}) < \kappa_2([\![\varphi_1 \wedge \psi \wedge (\varphi_3 \vee \neg\varphi_4)]\!]_{M_{\Sigma_b}})$, so $M_{\Sigma_b} \models \varphi_1 \wedge \psi \rightarrow \neg\varphi_3 \wedge \varphi_4$. ∎

Although the System Z approach has some attractive properties, it does not give us all we might want. For example, returning to Example 7.3.1, notice that we have neither $M_{\Sigma_2} \approx\!\!|^{Z} (\textit{penguin} \wedge \textit{bird}) \rightarrow \textit{warm-blooded}$ nor $M_{\Sigma_3} \approx\!\!|^{Z} (\textit{penguin} \wedge \textit{yellow}) \rightarrow \textit{easy-to-see}$ (Exercise 7.24). The next approach I consider has these properties.

This approach uses PS structures. Given a collection Σ of defaults, let Σ^k consist consist of the statements that result by replacing each default $\varphi \rightarrow \psi$ in Σ by the $\mathcal{L}^{QU}$ formula $\ell(\psi|\varphi) \geq 1 - 1/k$. Let $\mathcal{P}^k$ be the set of probability measures that satisfy the formulas in Σ^k. That is, $\mathcal{P}^k = \{\mu : (W_\Phi, \mu, \pi_\Phi) \models \Sigma^k\}$. If $\mathcal{P}^k \neq \emptyset$, let μ^{me}_k be the probability measure of maximum entropy in $\mathcal{P}^k$. (It can be shown that there is a unique probability measure of maximum entropy in this set, since it is defined by linear inequalities, but that is beyond the scope of this book.) As long as $\mathcal{P}^k \neq \emptyset$ for all $k \geq 1$, this procedure gives us a probability sequence $(\mu^{me}_1, \mu^{me}_2, \ldots)$. Let $M^{me}_\Sigma = (W_\Phi, (\mu^{me}_1, \mu^{me}_2, \ldots), \pi_\Phi)$. We say $\Sigma \approx\!\!|^{me} \varphi$ if either there is some k such that $\mathcal{P}^k = \emptyset$ (in which case $\mathcal{P}^{k'} = \emptyset$ for all $k' \geq k$) or $M^{me}_\Sigma \models \varphi$.

P is again sound for the maximum-entropy approach.

Proposition 7.3.3 *If* $\Sigma \vdash_{\mathbf{P}} \varphi$ *then* $\Sigma \approx\!\!|^{me} \varphi$.

Proof Suppose that $\Sigma \vdash \varphi$. It is easy to show that if $\mathcal{P}^k = \emptyset$ for some $k > 0$, then there is no structure $M \in \mathcal{M}^{ps}$ such that $M \models \Sigma$. On the other hand, if $\mathcal{P}^k \neq \emptyset$ for all $k \geq 1$, then $M_\Sigma^{me} \models \Sigma$ (Exercise 7.25). The result now follows immediately from Theorem 7.2.3. ∎

Standard properties of maximum entropy can be used to show that $\approx\!\!|^{me}$ has a number of additional attractive properties. In particular, it is able to ignore irrelevant attributes and sanctions inheritance across exceptional subclasses, giving the desired result in all the cases considered in Example 7.3.1.

Lemma 7.3.4 *Let* $\Sigma_a = \{\varphi_1 \to \varphi_2, \varphi_2 \to \varphi_3\}$, $\Sigma_b = \{\varphi_1 \to \varphi_2, \varphi_2 \to \varphi_3, \varphi_1 \to \neg\varphi_3, \varphi_1 \to \varphi_4\}$, $\Sigma_c = \{\varphi_1 \to \varphi_2, \varphi_2 \to \varphi_3, \varphi_1 \to \neg\varphi_3, \varphi_2 \to \varphi_4\}$, *and* $\Sigma_d = \{\varphi_1 \to \varphi_2, \varphi_2 \to \varphi_3, \varphi_1 \to \neg\varphi_3, \varphi_5 \to \varphi_4\}$.

(a) $\Sigma_a \models\!\!\approx^{me} \varphi_1 \wedge \psi \to \varphi_3$ *if* $\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \psi$ *is satisfiable.*

(b) $\Sigma_b \models\!\!\approx^{me} \varphi_1 \wedge \psi \to \neg\varphi_3 \wedge \varphi_4$ *if if* $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4 \wedge \psi$ *is satisfiable.*

(c) $\Sigma_c \models\!\!\approx^{me} \varphi_1 \to \varphi_4$ *if* $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4$ *is satisfiable.*

(d) $\Sigma_d \models\!\!\approx^{me} \varphi_1 \wedge \varphi_5 \to \varphi_4$ *if* $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4 \wedge \varphi_5$ *is satisfiable.*

Notice that parts (a) and (b) are just like Lemma 7.3.2, while part (c) is a special case of part (d). (Taking $\varphi_5 = \varphi_2$ in part (d) gives part (c).) While the proof of Lemma 7.3.4 is beyond the scope of this book, I can explain the basic intuition. It depends on the fact that maximum entropy makes things "as independent as possible". For example, given a set of constraints of the form $\ell(\psi|\varphi) = \alpha$ and a primitive proposition q that does not appear in any of these constraints, the structure that maximizes entropy subject to these constraints also satisfies $\ell(\psi|\varphi \wedge q) = \alpha$. Now consider the set Σ_2 of defaults. Interpreting these defaults as constraints, we have $\mu_n^{me}(\mathit{warm\text{-}blooded}|\mathit{bird}) \approx 1-1/n$ (most birds fly) and $\mu_n^{me}(\mathit{bird}|\mathit{penguin}) \approx 1-1/n$ (most birds are penguins). By the observation above, we also have $\mu_n^{me}(\mathit{warm\text{-}blooded}|\mathit{bird} \wedge \mathit{penguin}) \approx 1-1/n$. Thus,

$$
\begin{array}{ll}
 & \mu^{me}(\mathit{warm\text{-}blooded}|\mathit{penguin}) \\
= & \mu_n^{me}(\mathit{warm\text{-}blooded}|\mathit{bird} \wedge \mathit{penguin}) \times \mu_n^{me}(\mathit{bird}|\mathit{penguin}) \\
\approx & (1-1/n)^2 \\
\approx & 1-2/n.
\end{array}
$$

(In the last step, I am ignoring the $1/n^2$ term, since it is negligible compared to $1/2n$ for n large.) Thus, we get $\Sigma_2 \models\!\!\approx^{me} \mathit{penguin} \to \mathit{warm\text{-}blooded}$, as desired.

The maximum-entropy approach may seem somewhat *ad hoc*. While it seems to have a number of attractive properties, why is it the appropriate thing to use for nonmonotonic reasoning? One defense of it runs in the spirit of the usual defense of maximum entropy. If we view Σ_n as a set of constraints, the probability measure μ_n^{me} is the one that satisfies the constraints and gives the least "additional information" over and above this fact. But then why consider a sequence of measures like this at all? Some further motivation for the use of such a sequence will be given in Chapter **??**. But even if we accept the use of maximum entropy for now, there is still a problem of characterizing its properties in this context, something no one has yet been able to do. Besides the attractive properties described in Lemma 7.3.4, the approach may have some not-so-attractive properties as well, just as maximum entropy itself has unattractive properties in some contexts. Without a characterization, it is hard to feel completely comfortable using this approach.

7.4 Conditional Logic

$\mathcal{L}^{def}$ is a rather weak language. For example, although we can use it to say that a certain default holds, we cannot use it to say that a default does *not* hold, since $\mathcal{L}^{def}$ does not allow negated defaults. There is no great difficulty extending the language to allow negated and nested defaults, and many agents as well. Let $\mathcal{L}_n^{\rightarrow}$ be the language defined by starting with primitive propositions, and closing off under $\wedge$, $\neg$, and $\rightarrow_i$, $i = 1, \ldots, n$. Formulas in $\mathcal{L}_n^{\rightarrow}$ can describe logical combination of defaults (e.g., $(p \rightarrow_1 q) \vee (p \rightarrow_1 \neg q)$), negated defaults (e.g., $\neg(p \rightarrow_1 q)$), and nested defaults (e.g., $(p \rightarrow_1 q) \rightarrow_2 r$).

Formulas in $\mathcal{L}_n^{\rightarrow}$ in $\mathcal{M}_n^{ps}$, $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{pref}$, and $\mathcal{M}_n^{qual}$ (where $\mathcal{M}_n^{ps}$ and $\mathcal{M}_n^{qual}$ are the obvious generalizations of $\mathcal{M}^{ps}$ and $\mathcal{M}^{qual}$ to n agents) can be given semantics by extending the definition in the single-agent case in the obvious way. For example, if $M = (W, \mathcal{POSS}, \pi) \in \mathcal{M}_n^{poss}$, then

$$\begin{array}{l}(M, w) \models \varphi \rightarrow_i \psi \text{ iff } \mathrm{Poss}_{w,i}([\![\varphi]\!]_M \cap W_{w,i}) = 0 \text{ or} \\ \quad \mathrm{Poss}_{w,i}([\![\varphi \wedge \psi]\!]_M \cap W_{w,i}) > \mathrm{Poss}_{w,i}([\![\varphi \wedge \neg\psi]\!]_M \cap W_{w,i}),\end{array}$$

where $\mathcal{POSS}_i(w) = (W_{w,i}, \mathrm{Poss}_{w,i})$. Note that now we must write $(M, w) \models \varphi \rightarrow_i \psi$ rather than $M \models \varphi \rightarrow \psi$, because the possibility measure depends on the world (and the agent).

It should be clear from the definitions that formulas in $\mathcal{L}_n^{\rightarrow}$ can be expressed in $\mathcal{L}_n^{\gg}$.

Proposition 7.4.1 *For every structure M in $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{pref}$, $\mathcal{M}_n^{rank}$, and*

$\mathcal{M}_n^{qual}$,

$$(M,w) \models \varphi \rightarrow_i \psi \textit{ iff } (M,w) \models \neg(\ell_i(\varphi) > \ell_i(\mathit{false})) \vee \ell_i(\varphi \wedge \psi) > \ell_i(\varphi \wedge \neg\psi).$$

Proof Immediate from the definitions. ∎

What about the converse? Can all formulas in $\mathcal{L}_n^{\gg}$ can be expressed in $\mathcal{L}_n^{\rightarrow}$? In $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{rank}$, and $\mathcal{M}_n^{pref}$, they can.

Proposition 7.4.2 *For every structure M in $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{rank}$, and $\mathcal{M}_n^{pref}$,*

$$(M,w) \models \ell_i(\varphi) > \ell_i(\psi) \textit{ iff } (M,w) \models \neg(\varphi \rightarrow_i \mathit{false}) \wedge \neg((\varphi \vee \psi) \rightarrow_i \neg\psi)).$$

Proof See Exercise 7.26. ∎

The key step in the proof of Proposition 7.4.2 involves showing that $M \models \ell_i(\varphi) > \ell_i(\psi)$ iff $M \models \ell_i(\varphi \wedge \neg\psi) > \ell_i(\psi)$. While this property holds for structures M in $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{rank}$, and $\mathcal{M}_n^{pref}$, it does not hold in $\mathcal{M}_n^{qual}$ in general, so Proposition 7.4.2 does not extend to $\mathcal{M}_n^{qual}$. In fact, there is no formula in $\mathcal{L}_n^{\rightarrow}$ that is equivalent to $\ell_i(\varphi) > \ell_i(\psi)$ in all structures in $\mathcal{M}_n^{qual}$ (Exercise 7.27). Thus, in $\mathcal{M}_n^{qual}$, the language $\mathcal{L}_n^{\gg}$ is strictly more expressive than $\mathcal{L}_n^{\rightarrow}$.

Although we can translate $\mathcal{L}_n^{\rightarrow}$ to $\mathcal{L}_n^{\gg}$ and then use AX_n^{ord} (in the case of plausibility measures and preference orders) or AX_n^{tot} (in the case of possibility measures and ranking functions, since they lead to total orders), to reason about defaults, it is desirable to be able to characterize default reasoning directly in the language $\mathcal{L}_n^{\rightarrow}$. Of course, the characterization will depend to some extent on whether the underlying order is partial or total. As we shall see, there is one other relevant property as well.

Let **C** consist of the following axioms and inference rules.

Prop. All substitution instances of propositional tautologies.

C1. $\varphi \rightarrow_i \varphi$.

C2. $((\varphi \rightarrow_i \psi_1) \wedge (\varphi \rightarrow_i \psi_2)) \Rightarrow (\varphi \rightarrow_i (\psi_1 \wedge \psi_2))$.

C3. $((\varphi_1 \rightarrow_i \psi) \wedge (\varphi_2 \rightarrow_i \psi)) \Rightarrow ((\varphi_1 \vee \varphi_2) \rightarrow_i \psi)$.

C4. $((\varphi \rightarrow_i \psi_1) \wedge (\varphi \rightarrow_i \psi_2)) \Rightarrow ((\varphi \wedge \psi_2) \rightarrow_i \psi_1)$.

MP. From φ and $\varphi \Rightarrow \psi$ infer ψ.

RC1. From $\varphi \Leftrightarrow \varphi'$ infer $(\varphi \rightarrow_i \psi) \Rightarrow (\varphi' \rightarrow_i \psi)$.

RC2. From $\psi \Rightarrow \psi'$ infer $(\varphi \rightarrow_i \psi) \Rightarrow (\varphi \rightarrow_i \psi')$.

Axiom system **C** can be viewed as a generalization of **P**. The richer language lets us replace a rule like AND by the axiom C2. Similarly, C1, C3, C4, RC1, and RC2 are the analogues of REF, OR, CM, LLE, and RW, respectively. We need Prop and MP, as usual, to deal with propositional reasoning.

Theorem 7.4.3 **C** *is a sound and complete axiomatization of* $\mathcal{L}_n^{\rightarrow}$ *with respect to both* $\mathcal{M}_n^{pref}$ *and* $\mathcal{M}_n^{qual}$*.*

Proof As usual, soundness is straightforward (Exercise 7.28) and completeness is beyond the scope of this book. ∎

In the language $\mathcal{L}^{def}$, we could not distinguish between notions of likelihood based on partial orders and ones based on total orders. Conditional logic does allow us to make this distinction, and others as well. Consider the following two axioms.

C5. $(\varphi \rightarrow_i \psi_1) \land \neg(\varphi \rightarrow_i \neg\psi_2) \Rightarrow ((\varphi \land \psi_2) \rightarrow_i \psi_1)$.

C6. $\neg(\mathit{true} \rightarrow_i \mathit{false})$.

Note that C5 is almost the same as C4, except that the clause $\varphi \rightarrow_i \psi_2$ in C4 is replaced by $\neg(\varphi \rightarrow \neg\psi_2)$ in C5. C5 expresses a property called *rational monotonicity* in the literature; it is what distinguishes notions of uncertainty where the underlying notion of likelihood is total from ones where it is only partial. C5 does not hold in general in $\mathcal{M}_n^{qual}$ or $\mathcal{M}_n^{pref}$ (Exercise 7.29), but it does hold in $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{rank}$, and $\mathcal{M}_n^{tot}$. C6 corresponds to a property called *normality* in the literature. It holds for a plausibility measure Pl if $\mathrm{Pl}(W) > \perp$. This property holds for the plausibility measures arising from ranking functions, possibility measures, and probability sequences. As the following theorem shows, these axioms suffice to characterize reasoning about $\mathcal{L}^{\rightarrow}$ in $\mathcal{M}_n^{ps}$, $\mathcal{M}_n^{poss}$, $\mathcal{M}_n^{rank}$, and $\mathcal{M}_n^{tot}$.

Theorem 7.4.4

(a) **C**+{*C6*} *is a sound and complete axiomatization of* $\mathcal{L}_n^{\rightarrow}$ *with respect to* $\mathcal{M}_n^{ps}$*.*

(b) **C** +{*C5, C6*} *is a sound and complete axiomatization of* $\mathcal{L}_n^{\rightarrow}$ *with respect to* $\mathcal{M}_n^{tot}$*,* $\mathcal{M}_n^{rank}$*, and* $\mathcal{M}_n^{poss}$*.*

7.5 Reasoning About Counterfactuals

We can use the language $\mathcal{L}^{\rightarrow}$ to reason about counterfactuals as well as defaults. Now the interpretation of a formula such as $\varphi \rightarrow \psi$ is "if φ

were the case, then ψ would be true". In this section, $\varphi \rightarrow \psi$ gets this counterfactual reading.

Under what circumstances should such a counterfactual formula be true at a world w? Certainly if φ is already true at w (so that φ is not counter to fact) then it seems reasonable to take $\varphi \rightarrow \psi$ to be true at w if ψ is also true at w. But what if φ is not true at w? In that case, one approach is to consider the world(s) "most like w" where φ is true, and see if ψ is true there as well. But which worlds are "most like w"?

I am not going to try to characterize similarity here. Rather, I just show how we can use tools that we already have at our disposal to at least *describe* when one world is similar to another; this, in turn, gives us a way of giving semantics to counterfactuals. In fact, as I now show, all the approaches discussed in Section 7.2 can be used to give semantics to counterfactuals.

Let's start with preference orders. We can associate with each world w an ordering $\succeq_w$, where $w_1 \succeq_w w_2$ means that w_1 is at least as close to, or at least as similar to, w as w_2. We would expect w to be more like itself than any other world; that is, $w \succeq_w w'$ for all $w, w' \in W$. Note that this means we cannot use simple structures: the ordering really depends on the world.

A *counterfactual preferential structure* is a preferential structure (for one agent) $M = (W, \mathcal{O}, \pi)$ that satisfies the following condition:

Cfac$^{\succeq}$. If $\mathcal{O}(w) = (W_w, \succeq_w)$, then $w \in W_w$ and is the maximum element with respect to $\succeq_w$ (so that w is closer to itself than any other world): formally, $w \succeq_w w$ and $w \succ_w w'$ for all $w' \neq w$.

Let $\mathcal{M}_c^{pref}$ consist of all (single-agent) counterfactual preferential structures. We can generalize this to n agents in the obvious way.

We already have a definition for $\rightarrow$ in preferential structures, according to which, roughly speaking, $\varphi \rightarrow \psi$ holds if $\varphi \wedge \psi$ is more likely than $\varphi \wedge \neg\psi$. However, this does not seem to accord with the intuition that I gave earlier for counterfactuals. Fortunately, Theorem 7.2.4 show that there is another equivalent definition that could have been used, given by the operator $\rightarrow'$. Indeed, under the reinterpretation of $\succeq_w$, the operator $\rightarrow'$ gives us exactly what we want.

To make this precise, I generalize the definition of best_M so that it can depend on the world. Define

$$\text{best}_{M,w}(U) = \{w' \in U \cap W_w : \text{ for all } w'' \in U \cap W_w,\ w'' \not\succ_w w'\}.$$

Earlier, $\text{best}_M(U)$ was interpreted as "the most normal worlds in U"; now it should be interpreted as "the worlds in U closest to w". The proof of

Theorem 7.2.4 shows that in a general preferential structure M (whether or not it satisfies Cfac$^{\succeq}$)

$$(M,w) \models \varphi \rightarrow \psi \text{ iff } \mathrm{best}_{M,w}([\![\varphi]\!]_M) \subseteq [\![\psi]\!]_M.$$

That is, $\varphi \rightarrow \psi$ holds at w if all the worlds closest to or most like w that satisfy φ also satisfy ψ.

Note that in a counterfactual structure, W_w is *not* the set of worlds the agent considers possible. W_w in general includes worlds that the agent knows perfectly well to be impossible. For example, suppose that in the actual world w, the lawyer's client was drunk and it was raining. The lawyer wants to make the case that, even if his client hadn't been drunk and it had been sunny, the car would have hit the cow. (Actually, he may want to argue that there is a reasonable probability that the car would have hit the cow, but I shall ignore counterfactual probabilities here; they can be modeled using the tools of Chapter 7.) Thus, we want to consider worlds $w' \in W_w$ that are closest to w where it is sunny and the client is sober and driving his car. But these are worlds that are currently known to be impossible. This means that the interpretation of W_w in preferential structures depends on whether we use the structure for default reasoning or counterfactual reasoning.

Nevertheless, since counterfactual structures are a subclass of preferential structures, all the axioms in **C** are valid (when specialized to one agent). We get one additional property that corresponds to the condition Cfac$^{\succeq}$:

C7. $\varphi \Rightarrow (\psi \Leftrightarrow (\varphi \rightarrow \psi))$

C7 is in fact the property that I discussed earlier, which says that if φ is already true at w, then the counterfactual $\varphi \rightarrow \psi$ is true at w if and only if ψ is true at w.

Theorem 7.5.1 **C**+$\{C7\}$ *is a sound and complete axiomatization of* $\mathcal{L}^{\rightarrow}$ *with respect to* $\mathcal{M}_c^{pref}$.

Proof I leave it to the reader to check that C7 is valid in counterfactual preferential structures (Exercise 7.30). The validity of all the other axioms in **C** follows from Theorem 7.4.3. Again, completeness is beyond the scope of the book. ∎

Of course, rather than allowing arbitrary partial orders in counterfactual structures, we can restrict to total orders. In this case, C5 becomes an axiom.

Not surprisingly, all the other approaches that were used to give semantics to defaults can also be used to give semantics to counterfactuals. Indeed, the likelihood interpretation also makes sense for counterfactuals. We can still interpret "if φ were true, then ψ would be true" as meaning "the likelihood of ψ given φ is much higher than that of $\neg\psi$ given φ". While we cannot take "ψ given φ" as meaning probabilistic conditioning, since the probability of φ may well be 0 (in fact, the antecedent φ in a counterfactual may well be a formula that the agent knows to be false), it does make sense if we think in terms of possibility, ranking, or plausibility. All we need is an analogue to the condition Cfac$^{\succeq}$. The analogues are not hard to come up with. For example, for ranking structures, the analogue is

Cfac$^{\kappa}$. If $\mathcal{RANK}(w) = (W_w, \kappa_w)$, then $w \in W_w$ and $\kappa_w(w) < \kappa_w(W_w - \{w'\})$.

Similarly, for plausibility structures, the analogue is

Cfac$^{\mathrm{Pl}}$. If $\mathcal{PL}(w) = (W_w, \mathrm{Pl}_w)$, then $w \in W_w$ and $\mathrm{Pl}_w(w) > \mathrm{Pl}_w(W_w - \{w\})$.

I leave it to the reader to check that counterfactual ranking structures and counterfactual plausibility structures satisfy C7, and to come up with the appropriate analogue to Cfac$^{\succeq}$ in the case of probability sequences and possibility measures (Exercises 7.31 and 7.32).

Exercises

7.1 Show that if $\varphi \Rightarrow \psi$ is valid, then so is $\varphi \wedge \varphi' \Rightarrow \psi$, no matter what φ' is.

7.2 Suppose that $M \in \mathcal{M}^{meas}$. For this exercise, define $M \models \varphi \rightarrow \psi$ if $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\psi]\!]_M | [\![\varphi]\!]_M) > \mu([\![\neg\psi]\!]_M | [\![\varphi]\!]_M)$. Show that the following are equivalent:

(a) $M \models \varphi \rightarrow \psi$

(b) $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\varphi \wedge \psi]\!]_M) > \mu([\![\varphi \wedge \neg\psi]\!]_M)$

(c) $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\psi]\!]_M | [\![\varphi]\!]_M) > 1/2$.

Moreover, show that this interpretation satisfies LLE, RW, and REF.

7.3 Show that the OR rule is violated in the structure M_1 of Example 7.2.1.

7.4 Suppose that $M \in \mathcal{M}^{meas}$. Fix $\epsilon > 0$. For this exercise, define $M \models \varphi \rightarrow \psi$ if $\mu([\![\varphi]\!]_M) = 0$ or $\mu([\![\psi]\!]_M | [\![\varphi]\!]_M) > 1 - \epsilon$. Show that this interpretation satisfies LLE, RW, and REF, but not AND, CM, or OR.

7.5 Show directly that OR and CM hold in PS structures (that is, without using Theorems 7.2.10 and 7.2.11).

* **7.6** Show that

$$\{p \wedge q \rightarrow r, p \rightarrow \neg r\} \vdash_{\mathbf{P}} p \rightarrow \neg q.$$

7.7 Show directly that OR and CM hold in possibility structures (that is, without using Theorems 7.2.10 and 7.2.11). In addition, complete the proof of the soundness of the AND rule by showing that if $\max(\alpha, \beta) > \max(\gamma, \delta)$ and $\max(\alpha, \gamma) > \max(\beta, \delta)$, then $\alpha > \max(\beta, \gamma, \delta)$.

7.8 Show that $M \models \varphi \Rightarrow \psi$ iff $[\![\varphi]\!]_M \subseteq [\![\psi]\!]_M$ for every structure M.

7.9 Prove Theorem 7.2.6.

7.10 Prove Lemma 7.2.7.

7.11 Prove Proposition 7.2.9.

7.12 Show that Pl4 is necessary for the AND rule in the following three (related) senses.

(a) Suppose that (W, Pl) is a plausibility space that does not satisfy Pl4. Show that there exists an interpretation π such that the plausibility structure (W, Pl, π) does not satisfy the AND rule.

(b) Suppose that $M = (W, \mathrm{Pl}, \pi)$ is a plausibility structure such that $\pi(w) \neq \pi(w')$ if $w \neq w'$ and Pl does not satisfy Pl4. Again, show that M does not satisfy the AND rule.

(c) Suppose that $M = (W, \mathrm{Pl}, \pi)$ is a plausibility measure such that Pl does not satisfy Pl4 when restricted to sets definable by formulas. That is, there exist formulas φ_1, φ_2, and φ_3 such that $[\![\varphi_1]\!]_M$, $[\![\varphi_2]\!]_M$, and $[\![\varphi_3]\!]_M$ are pairwise disjoint, $\mathrm{Pl}([\![\varphi_1 \vee \varphi_2]\!]_M) > \mathrm{Pl}([\![\varphi_3]\!]_M)$, $\mathrm{Pl}([\![\varphi_1 \vee \varphi_3]\!]_M) > \mathrm{Pl}([\![\varphi_2]\!]_M)$, and $\mathrm{Pl}([\![\varphi_1]\!]_M) \not> \mathrm{Pl}([\![\varphi_1 \vee \varphi_2]\!]_M)$. Again, show that M does not satisfy the AND rule.

Show that the requirement in part (b) that $\pi(w) \neq \pi(w')$ if $w \neq w'$ is necessary by demonstrating a plausibility structure that does not satisfy Pl4 and yet satisfies the AND rule. (Of course, for this plausibility structure, it must be the case that there are two distinct worlds that satisfy the same truth assignment.)

7.13 Consider a simple plausibility structure $M = (W, \mathrm{Pl}, \pi)$, where Pl satisfies Pl4. Define a modal operator K (but now think of K as belief, not knowledge) in M by defining $(M, w) \models K\varphi$ iff $\mathrm{Pl}([\![\varphi]\!]_M) > \mathrm{Pl}([\![\neg\varphi]\!]_M)$. That is, the agent believes φ if φ is more plausible than $\neg\varphi$. Show that this definition satisfies the axioms of KD45. (It can actually be shown that KD45 is a sound and complete axiomatization with respect to this semantics, but that is beyond the scope of this book.)

7.14 Show that CM and OR are sound in $\mathcal{M}^{qual}$.

* **7.15** This exercise fills in the details of the proof of Theorem 7.2.11. Fix a PS structure $M = (W, (\mu_1, \mu_2, \ldots), \pi)$.

(a) Define an ordering $\succeq'$ on subsets of W such that $U \succeq' V$ if $\lim_{i\to\infty} \mu_i(U|U \cup V) = 1$. Show that $\succeq'$ is reflexive and transitive.

(b) Show by means of a counterexample that $\succeq'$ is not necessarily antisymmetric.

(c) Define a relation $\sim$ on subsets of W by defining $U \sim V$ if $U \succeq' V$ and $V \succeq' U$. Show that $\sim$ is reflexive, symmetric, and transitive.

(d) Define $[U] = \{V : V \sim U\}$. Since $\sim$ is an equivalence relation, show that for all $U, U' \subseteq W$, either $[U] = [U']$ or $[U] \cap [U'] = \emptyset$. Let $W/{\sim} = \{[U] : U \subseteq W\}$.

(e) Define an ordering $\succeq$ on $W/{\sim}$ by defining $[U] \succeq [V]$ iff there exist some $U \in [U]$ and $V \in [V]$ such that $U \succeq V$. Show that $\succeq$ is a partial order (that is, reflexive, antisymmetric, and transitive).

(f) Show that $[\emptyset]$ consists of all sets U such that there exists some N such that $\mu_n(U) = 0$ for all $n > N$ and that $[\emptyset]$ is the element $\perp$ in the partially ordered domain $W/{\sim}$. (A trivial argument shows that $[W]$ is the element $\top$.)

(g) Define a plausibility measure Pl on W by taking $\mathrm{Pl}(U) = [U]$, for $U \subseteq W$. Show that Pl satisfies Pl3, Pl4, and Pl5.

(h) Let $M_{PS} = (W, \mathrm{Pl}, \pi)$. By part (g), $M_{PS} \in \mathcal{M}^{qual}$. Show that $M \models \varphi \to \psi$ iff $M_{PS} \models \varphi \to \psi$.

7.16 Show that Pl4 and Pl5 completely characterize the plausibility structures that satisfy **P** in the following sense: Let $\mathcal{M}$ be a collection of simple plausibility structures such that for each structure $M = (W, \mathrm{Pl}, \pi) \in \mathcal{M}$, if $w \neq w' \in W$, then $\pi(w) \neq \pi(w')$. Show that if there is a structure in $\mathcal{M}$

that does not satisfy Pl4 or Pl5, then **P** is not sound in $\mathcal{M}$. (Note that the argument in the case of Pl4 is just part (b) of Exercise 7.12; a similar argument works for Pl5. In fact, variants of this results corresponding to parts (a) and (c) of Exercise 7.12 can also be proved.)

7.17 Prove Theorem 7.2.13.

* **7.18** This exercise proves the first half of Theorem 7.2.14, that richness is necessary for completeness.

(a) Let $\varphi_1, \ldots, \varphi_n$ be a collection of mutually exclusive and consistent propositional formulas. Let Σ consist of the default $\varphi_n \rightarrow \mathit{false}$ and the defaults $\varphi_i \vee \varphi_j \rightarrow \varphi_i$ for all $1 \leq i < j \leq n$. Show that $(W, \mathrm{Pl}, \pi) \models \Sigma$ if and only if there is some j with $1 \leq j \leq n$ such that

$$\mathrm{Pl}([\![\varphi_1]\!]_M) > \mathrm{Pl}([\![\varphi_2]\!]_M) > \cdots > \mathrm{Pl}([\![\varphi_j]\!]_M) = \cdots = \mathrm{Pl}([\![\varphi_n]\!]_M) = \perp.$$

(b) Suppose $\mathcal{M}$ is not rich. Let $\varphi_1, \ldots, \varphi_n$ be the formulas that provide a counterexample to richness and let Σ be the set of defaults defined in part (a). Show that if $(W, \mathrm{Pl}, \pi) \in \mathcal{M}$ satisfies the defaults in Σ, then $\mathrm{Pl}([\![\varphi_{n-1}]\!]_M) = \perp$.

(c) Using part (b), show that $\Sigma \models_{\mathcal{M}} \varphi_{n-1} \rightarrow \mathit{false}$.

(d) Show that $\Sigma \not\vdash_{\mathbf{P}} \varphi_{n-1} \rightarrow \mathit{false}$. (Hint: show that there exists a plausibility structure satisfying all the defaults in Σ but not $\varphi_{n-1} \rightarrow \mathit{false}$, and then use Theorem 7.2.4.)

This shows that if $\mathcal{M}$ is not rich, then **P** is not complete with respect to $\mathcal{M}$. Although $\Sigma \models_{\mathcal{M}} \varphi_{n-1} \rightarrow \mathit{false}$, we cannot prove $\varphi_{n-1} \rightarrow \mathit{false}$ from Σ in **P**.

** **7.19** This exercise provides a proof of the second half of Theorem 7.5.1, that richness is sufficient for completeness. Suppose that there is some Σ and $\varphi \rightarrow \psi$ such that $\Sigma \models_{\mathcal{M}} \varphi \rightarrow \psi$ but $\Sigma \not\vdash_{\mathbf{P}} \varphi \rightarrow \psi$. Show that $\mathcal{M}$ is not rich as follows.

(a) By Theorem 7.2.4, $\Sigma \not\models_{\mathcal{M}^{pref}} \varphi \rightarrow \psi$. Thus, there is a simple preferential structure $M = (W, \succeq, \pi)$ that satisfies the defaults in Σ but not $\varphi \rightarrow \psi$. In fact, we can do better: we can assume that $\succeq$ is a total order. Show that there is a preferential structure structure $M = (W, \succeq, \pi)$ such that $W = \{w_1, \ldots, w_n\}$, $w_i \succ w_j$ and $\pi(w_i) \neq \pi(w_j)$ if $i < j$, $M \models \Sigma$, and $M \not\models \varphi \rightarrow \psi$.

(b) Now use the preferential structure M of part (a) to construct a sequence of formulas that will be a counterexample to the richness of $\mathcal{M}$. Show that there exist mutually exclusive propositional formulas $\varphi_1, \ldots, \varphi_n$ such that $(M, w_i) \models \varphi_i$. (Hint: use the fact that $\pi(w_i) \neq \pi(w_j)$ if $i \neq j$.)

(c) Let $\varphi_{n+1} = \neg(\varphi_1 \vee \ldots \vee \varphi_n)$. Show that $\varphi_1, \ldots, \varphi_{n+1}$ are mutually exclusive.

(d) Show that if $M' = (W', \mathrm{Pl}', \pi')$ is a simple plausibility structure such that $\mathrm{Pl}'([\![\varphi_1]\!]_{M'}) > \cdots > \mathrm{Pl}'([\![\varphi_{n+1}]\!]_{M'}) = \perp$, then M' satisfies the defaults in Σ but not $\varphi \rightarrow \psi$.

(e) Show that $\mathcal{M}$ is not rich. (Hint: show $\mathcal{M}$ does not contain a structure M' such that $M' \models \Sigma$.)

7.20 This exercise refers to Example 7.3.1.

(a) Show that $\Sigma_1 \not\vdash_{\mathbf{P}}$ *penguin* $\wedge$ *red* $\rightarrow \neg fly$.

(b) Show that $\Sigma_2 \not\vdash_{\mathbf{P}}$ (*penguin* $\wedge$ *bird*) $\rightarrow$ *warm-blooded*.

(c) Show that $\Sigma_3 \not\vdash_{\mathbf{P}}$ *penguin* $\wedge$ *yellow* $\rightarrow$ *easy-to-see*.

(c) Show that $\Sigma_4 \not\vdash_{\mathbf{P}}$ *robin* $\rightarrow$ *warm-blooded*. and that $\Sigma_4 \not\vdash_{\mathbf{P}}$ *robin* $\wedge$ *bird* $\rightarrow$ *warm-blooded*.

(Hint: for part (a), by Theorem 7.2.4, it suffices to find a preferential structure—or a possibility structure or a ranking structure—satisfying all the formulas in Σ_1, but not *penguin* $\wedge$ *red* $\rightarrow \neg$*fly*. A similar approach works for parts (b) and (c).)

7.21 This exercise compares $\rightarrow$ and $|rimp$. Referring again to Example 7.3.1, let Σ_1' be the result of replacing all occurrences of $\rightarrow$ in Σ_1 by $\Rightarrow$. Show that $M \models \Sigma_1'$ if and only if $[\![\textit{penguin}]\!]_M = \emptyset$. Note that it follows that $\Sigma_1' \models \neg$*penguin*. On the other hand, show that there is a structure $M \in \mathcal{M}^{rank}$ such that $M \models \Sigma_1$, $M \models$ (*penguin* $\wedge$ *red*) $\rightarrow \neg$*fly*, and $[\![\textit{penguin} \wedge \textit{red}]\!]_M \neq \emptyset$.

* **7.22** Show that if $\mathcal{M}_\Sigma^{rank} \neq \emptyset$, then there is a unique structure $M_\Sigma \in \mathcal{M}_\Sigma^{rank}$ that is most preferred, in that $M_\Sigma \succeq M$ for all $M \in \mathcal{M}_\Sigma^{rank}$.

7.23 Complete the proof of Lemma 7.3.2, by showing that

(a) in M_{Σ_a}, all worlds satisfying $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ have rank 0 and all worlds satisfying $\varphi_1 \wedge \neg\varphi_2$ or $\varphi_2 \wedge \neg\varphi_3$ have rank 1; and

(b) in M_{Σ_b}, all worlds satisfying $\neg\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3$ have rank 0, all worlds satisfying $\varphi_1 \wedge \varphi_2 \wedge \neg\varphi_3 \wedge \varphi_4$ have rank 1, and all worlds satisfying $\varphi_1 \wedge \varphi_3$ or $\varphi_1 \wedge \neg\varphi_4$ have rank 2.

7.24 Show that $M_{\Sigma_2} \not\approx^Z (\mathit{penguin} \wedge \mathit{bird}) \rightarrow \mathit{warm\text{-}blooded}$ and that $M_{\Sigma_3} \not\approx^Z (\mathit{penguin} \wedge \mathit{yellow}) \rightarrow \mathit{easy\text{-}to\text{-}see}$.

7.25 Complete the proof of Proposition 7.3.3 by showing that

(a) if $\mathcal{P}^k = \emptyset$ for some $k > 0$, then there is no structure $M \in \mathcal{M}^{ps}$ such that $M \models \Sigma$,

(b) if $\mathcal{P}^k \neq \emptyset$ for all $k \geq 1$, then $M_\Sigma^{me} \models \Sigma$.

7.26 Prove Proposition 7.4.2.

7.27 Let $W = \{a, b, c\}$. Define two plausibility measures, Pl_1 and Pl_2, on W. Each of these plausibility measures assigns to each subset of W a triple of integers. Define a straightforward ordering on triples: $(i, j, k) \leq (i', j', k')$ if $i \leq i'$, $j \leq j'$, and $k \leq k'$; $(i, j, k) < (i', j', k')$ if $(i, j, k) \leq (i', j', k')$ and $(i', j', k') \not\leq (i, j, k)$. Pl_1 is defined so that $\mathrm{Pl}_1(\emptyset) = (0, 0, 0)$, $\mathrm{Pl}_1(a) = (1, 0, 0)$, $\mathrm{Pl}_1(b) = (0, 1, 0)$, $\mathrm{Pl}_1(c) = (0, 0, 1)$, $\mathrm{Pl}_1(\{a, b\}) = (1, 1, 1)$, $\mathrm{Pl}_1(\{a, c\}) = (2, 0, 1)$, $\mathrm{Pl}_1(\{b, c, \}= (0, 2, 1)$, and $\mathrm{Pl}_1(\{a, b, c\} = (2, 2, 2)$. Pl_2 is identical to Pl_1 except that $\mathrm{Pl}_2(\{a, b\}) = (2, 2, 1)$. Let $\Phi = \{p_a, p_b, p_c\}$ and define π so that $\pi(d)(p_e) = \mathbf{true}$ iff $d = e$ (so that p_a is true only at world a, p_b is true only at world b, and p_c is true only at world c). Let $M_j = (W, \mathcal{PL}_1^j, \pi)$, where $\mathcal{PL}_1^j(w) = (W, \mathrm{Pl}_j)$, for $j = 1, 2$.

(a) Show that both M_1 and M_2 are in $\mathcal{M}^{qual}$; that is, show that Pl_1 and Pl_2 satisfy Pl4 and Pl5.

(b) Show that if U and V are disjoint subsets of W, then $\mathrm{Pl}_1(U) > \mathrm{Pl}_1(V)$ iff $\mathrm{Pl}_2(U) > \mathrm{Pl}_2(V)$.

(c) Show as a consequence that $(M_1, w) \models \varphi$ iff $(M_2, w) \models \varphi$ for all formulas $\varphi \in \mathcal{L}_1^{\rightarrow}$ and all $w \in W$.

(d) Note, however, that $(M_1, w) \models \neg(\ell_1(p_a \vee p_b) > \ell_1(p_b \vee p_c))$ while $(M_2, w) \models \ell_1(p_a \vee p_b) > \ell_1(p_b \vee p_c)$.

This exercise shows that $\ell_1(p_a \vee p_b) > \ell_1(p_b \vee p_c)$ is not equivalent to any formula in $\mathcal{L}_1^{\rightarrow}$. For if it were equivalent to some formula φ, then by part (d), we would have $(M_1, a) \models \neg\varphi$ and $(M_2, a) \models \varphi$. However, part (c) shows that this cannot happen.

7.28 Prove that system **C** is a sound axiomatization of $\mathcal{L}_n^{\rightarrow}$ with respect to both $\mathcal{M}_n^{pref}$ and $\mathcal{M}_n^{qual}$.

7.29 Show that C5 does not hold in general in $\mathcal{M}_n^{qual}$ or $\mathcal{M}_n^{pref}$, by providing a counterexample.

7.30 Show that C7 is valid in counterfactual preferential structures.

7.31 Show that counterfactual ranking structures (that is, ranking structures satisfying Cfac^{κ}) and counterfactual plausibility structures (that is, plausibility structures satisfying $\mathrm{Cfac}^{\mathrm{Pl}}$) satisfy C7.

7.32 Construct conditions analogous to $\mathrm{Cfac}^{\succeq}$ appropriate for possibility structures and PS structures, and show that the resulting classes of structures satisfy C7.

7.33 In counterfactual preferential structures, there may in general be more than one world closest to w satisfying φ. In this exercise I consider counterfactual preferential structures where, for each formula φ and world w, there is always a unique closest world to w satisfying φ.

(a) $M = (W, \mathcal{O}, \pi)$, where $\mathcal{O}(w) = (W_w, \succeq_w)$, is a *totally ordered (counterfactual) structure* if, for all $w \in W$, $\succeq_w$ is a *total order*—that is, for all $w' \neq w''$, either $w' \succ_w w''$ or $w'' \succ_w w'$. Show that in totally ordered structures, there is always a unique closest world to w satisfying φ for each world w and formula φ.

(b) Show that in totally-ordered counterfactual structures, the following axiom is valid:

C8. $(\varphi \rightarrow \psi) \vee \varphi(\rightarrow \neg\psi)$.

In fact, it can be shown that C8 characterizes totally-ordered counterfactual structures (although doing so is beyond the scope of this book).

(c) Show that C5 follows from C8 and all the other axioms and inference rules in **C**.

Notes

There has been a great deal of discussion in the philosophical literature about *conditional statements*. These are statements of the form "if φ then ψ", and include counterfactuals as a special case. Stalnaker [1992] gives a short and readable survey of the philosophical issues involved.

There have been many approaches to giving semantics to defaults. Some of the early and most influential approaches include Reiter's *default logic* [1980], McCarthy's *circumscription* [1980], McDermott and Doyle's *nonmonotonic logic* [1980], and Moore's *autoepistemic logic* [1985]. See Reiter's overview paper [1984] and the book by Marek and Trusczynski [1993] for a discussion of the issues.

The approach discussed in this chapter, characterized by axiom system **P**, was introduced by Kraus, Lehmann, and Magidor [1990] (indeed, the axioms and rules of **P** are often called the *KLM properties* in the literature) and Makinson [1989], based on ideas that go back to Gabbay [1985]. Kraus, Lehmann, and Magidor and Makinson gave semantics to default formulas using preferential structures. Pearl [1989] gave probabilistic semantics to default formulas using what he called *epsilon semantics*, an approach that actually was used independently and earlier by Adams [1975] to give semantics to conditionals. The formulation given here using PS structures was introduced by Goldszmidt, Morris, and Pearl [1993], and was shown by them to be equivalent to Pearl's original notion of epsilon semantics. Geffner [1992b] showed that this approach is also characterized by **P**.

Dubois and Prade [1991] were the first to used possibility measures for giving semantics to defaults; they showed that **P** characterized reasoning about defaults using this semantics. Goldszmidt and Pearl [1992] did the same for ranking functions. Friedman and I used plausibility measures to explain why all these different approaches are characterized by **P**. In particular, Theorems 7.2.10, 7.2.11, 7.2.13, and 7.2.14 are from [Friedman and Halpern 1998].

There has been a great deal of effort applied to going beyond axiom system **P**. The basic observation that many of the approaches to nonmonotonic reasoning (in particular, ones that go beyond **P**) can be understood in terms of choosing a preferred structure that satisfies some defaults is due to Shoham [1987]. The system Z approach and maximum entropy approach discussed in Section 7.3 were introduced by Pearl [1990] and Goldszmidt, Morris, and Pearl [1993], respectively; see these papers for further details. Two other approaches that have many of the properties of the maximum-entropy approach are due to Geffner [1992a] and Bacchus, Grove, Halpern, and Koller [1996]; the latter approach will be discussed further in Chapter **??**.

The language $\mathcal{L}^{\rightarrow}$ was introduced by Lewis [1973]. Lewis first proved the connection between $\rightarrow$ and $\rightarrow'$ given in Theorem 7.2.6; he also showed that $>$ could be captured by $\rightarrow$ in preferential orders as described in Proposition 7.4.2. (Lewis assumed that the preference order was total; the fact that the same connection holds even if the order is partial was observed in [Halpern 1997a].) The soundness and completeness of **C** for preferential structures (Theorem 7.4.3) was proved by Burgess [1981]; the result for plausibility structures is proved in [Friedman and Halpern 1998].

Stalnaker [1968] first gave semantics to counterfactuals using what he called *selection functions*. A selection function f takes as arguments a world w and a formula φ; $f(w, \varphi)$ is taken to be the world closest to w satisfying φ. (Notice that this means that there is a unique closest world, as in Exercise 7.33.) Stalnaker and Thomason [1970] provided a complete axiomatization for counterfactuals using this semantics. The semantics for counterfactuals using preferential orders presented here is due to Lewis [1973].