1

The Logic of Defeasible Inference

1.1 FIRST-ORDER LOGIC

It was mentioned that first-order logic (henceforth FOL) was originally developed for the representation of mathematical reasoning. Such a representation required the establishment of a high standard of rigor, meant to guarantee that the conclusion follows from the premises with absolute deductive cogency. In this respect, FOL turned out to be nothing but a stunning success. The account of deductive reasoning provided by FOL enjoys a number of important mathematical properties, which can also be used as a crucial benchmark for the assessment of alternative logical frameworks. (The reader interested in an introduction to the nuts and bolts of FOL can consult any of the many excellent introductory texts that are available, such as Enderton, 1972.)

From the point of view of abstract consequence relations, FOL provides an implementation of the so-called *no-counterexample* account: A sentence ϕ is a consequence of a set Γ of sentences if and only if one cannot reinterpret the (nonlogical part of the) language in which Γ and ϕ are formulated in such a way as to make all sentences in Γ true and ϕ false. An inference from premises $\psi_1, \ldots, \psi_k$ to a conclusion ϕ is *valid* if ϕ is a consequence of $\{\psi_1, \ldots, \psi_k\}$, i.e., if the inference has no counterexample.

For this to be a rigorous account of logical consequence, the underlying notion of interpretation needs to be made precise, along with a (noncircular, possibly stipulative) demarcation of the logical and nonlogical vocabulary. This was accomplished by Alfred Tarski in 1935, who precisely defined the notion of truth of a sentence on an interpretation (see Tarski, 1956, for a collection of Tarski's technical papers). In so doing,

Tarski overcame both a technical and a philosophical problem. The technical problem had to do with the fact that in FOL quantified sentences are obtained from components that are not, in turn, sentences, so that a direct recursive definition of truth for sentences breaks down at the quantifier case. To overcome this problem Tarski introduced the auxiliary notion of *satisfaction*. The philosophical obstacle had to do with the fact that the notion of *truth* was at the time considered suspiciously metaphysical among logicians trained within the environment of the Vienna Circle. This was a factor, for instance, in Gödel's reluctance to formulate his famous undecidability results in terms of truth (see, for instance, Feferman, 1998).

Tarski's analysis yielded a mathematically precise definition for the no-counterexample consequence relation of FOL, which is usually denoted by the symbol "$\models$": We say that ϕ is a consequence of a set Γ of sentences, written $\Gamma \models \phi$, if and only if ϕ is true on every interpretation on which every sentence in Γ is true. At first glance, there would appear to be something intrinsically infinitary about $\models$. Regardless of whether Γ is finite or infinite, to check whether $\Gamma \models \phi$ one has to "survey" infinitely many possible interpretations and check whether any of them is a counterexample to the entailment claim, i.e., whether any of them is such that all sentences in Γ are true on it while ϕ is false.

However, surprisingly, in FOL the infinitary nature of $\models$ is only apparent. As Gödel (1930) showed, the relation $\models$, although defined by universally quantifying over all possible interpretations, can be analyzed in terms of the existence of finite objects of a certain kind, viz., formal proofs. A *formal proof* is a finite sequence of sentences, each of which is an *axiom*, an *assumption*, or is obtained from previous ones by means of one of a finite number of inference rules, such as *modus ponens*. If a sentence ϕ occurs as the last line of a proof, then we say that the proof is a *proof of* ϕ, and we say that ϕ is *provable from* Γ, written $\Gamma \vdash \phi$, if and only if there is a proof of ϕ all of whose assumptions are drawn from Γ. In practice, in FOL, one provides a small and clearly defined number of primitive inferential principles (such as axioms and rules) and then posits that a conclusion ϕ follows from a set Γ of premises if ϕ can be obtained from some of the premises by repeated application of the inferential principles. Many different axiomatizations of FOL exist, and a particularly simple and elegant one can be found in Enderton (1972).

Gödel's famous completeness theorem of 1930 states that the two relations, $\models$ and $\vdash$, are extensionally equivalent: For any ϕ and Γ, $\Gamma \models \phi$ if and only if $\Gamma \vdash \phi$. This is a remarkable feature of FOL, which has a number of consequences. One of the deepest consequences follows from

the fact that proofs are finite objects, and hence that $\Gamma \vdash \phi$ if and only if there is a *finite* subset Γ_0 of Γ such that $\Gamma_0 \vdash \phi$. This, together with the completeness theorem, gives us the *compactness theorem*: $\Gamma \models \phi$ if and only if there is a finite subset Γ_0 of Γ such that $\Gamma_0 \models \phi$. There are many interesting equivalent formulations of the theorem, but the following one is perhaps the most often cited. Say that a set of sentences is *consistent* if they can all be made simultaneously true on some interpretation; then the compactness theorem says that a set Γ is consistent if and only if each of its finite subsets is by itself consistent.

Another important consequence of Gödel's completeness theorem is the following form of the Löwenheim–Skolem theorem: If all the sentences in Γ can be made simultaneously true on some interpretation, then they can also be made simultaneously true on some (other) interpretation whose universe is no larger than the set $\mathbb{N}$ of the natural numbers.

Together, the compactness and the Löwenheim–Skolem theorems are the beginning of one of the most successful branches of modern symbolic logic: model theory. The compactness and the Löwenheim–Skolem theorems characterize FOL; as shown by Per Lindström in 1969, any logical system (meeting certain "regularity" conditions) for which both compactness and Löwenheim–Skolem hold is no more expressive than FOL (see Ebbinghaus, Flum, and Thomas, 1994, Chap. XIII, for an accessible treatment).

Gödel's completeness theorem also reflects on the question of whether and to what extent one can devise an effective procedure to determine whether a sentence ϕ is valid or, more generally, if $\Gamma \models \phi$ for given Γ and ϕ. First, some terminology. We say that a set Γ of sentences is *decidable* if there is an effective procedure, i.e., a mechanically executable set of instructions that determines, for each sentence ϕ, whether ϕ belongs to Γ or not. Notice that such a procedure gives both a positive and a negative test for membership of a sentence ϕ in Γ. A set of sentences is *semidecidable* if there is an effective procedure that determines if a sentence ϕ is a member of Γ, but might not provide an answer in some cases in which ϕ is not a member of Γ. In other words, Γ is semidecidable if there is a positive, but not necessarily a negative, test for membership in Γ. Equivalently, Γ is semidecidable if it can be given an effective listing, i.e., if it can be mechanically generated. These notions can be generalized to relations among sentences of any number of arguments. For instance, it is an important feature of the axiomatizations of FOL, such as that of Enderton (1972), that both the set of axioms and the relation that holds among $\phi_1, \ldots, \phi_k$ and ψ when ψ can be inferred from $\phi_1, \ldots, \phi_k$ by one application of the

rules, are decidable. As a result, the relation that holds among $\phi_1, \ldots, \phi_k$ and ϕ whenever $\phi_1, \ldots, \phi_k$ is a proof of ϕ is also decidable.

The import of Gödel's completeness theorem is that if the set Γ is decidable (or even only semidecidable), then the set of all sentences ϕ such that $\Gamma \models \phi$ is semidecidable. Indeed, one can obtain an effective listing for such a set by systematically generating all proofs from Γ. The question arises of whether, in addition to this positive test, there might not be a negative test for a sentence ϕ being a consequence of Γ. This *decision problem* [*Entscheidungsproblem*] was originally proposed by David Hilbert in 1900, and it was solved in 1936 independently by Alonzo Church and Alan Turing. The Church–Turing theorem states that, in general, it is not decidable whether $\Gamma \models \phi$, or even whether ϕ is valid. (It's important to know that for many, even quite expressive, fragments of FOL the decision problem is solvable; see Börger, Grädel, and Gurevich, 1997, for details.) We should also notice the following fact that will be relevant in Section 1.3; say that a sentence ϕ is *consistent* if $\{\phi\}$ is consistent, i.e., if its negation $\neg\phi$ is not valid. Then the set of all sentences ϕ such that ϕ is consistent is not even semidecidable, for a positive test for such a set would yield a negative test for the set of all valid sentences, which would so be decidable, against the Church–Turing theorem.

1.2 CONSEQUENCE RELATIONS

In the previous section, we considered the no-counterexample consequence relation $\models$ by saying that $\Gamma \models \phi$ if and only if ϕ is true on every interpretation on which every sentence in Γ is true. In general, it is possible to consider the abstract properties of a relation of consequence between sets of sentences and single sentences. Let $\mathrel{|\!\sim}$ be any such relation. We identify the following properties, all of which are satisfied by the consequence relation $\models$ of FOL:

Supraclassicality: If $\Gamma \models \phi$ then $\Gamma \mathrel{|\!\sim} \phi$;
Reflexivity: If $\phi \in \Gamma$ then $\Gamma \mathrel{|\!\sim} \phi$;
Cut: If $\Gamma \mathrel{|\!\sim} \phi$ and $\Gamma, \phi \mathrel{|\!\sim} \psi$ then $\Gamma \mathrel{|\!\sim} \psi$;
Monotony: If $\Gamma \mathrel{|\!\sim} \phi$ and $\Gamma \subseteq \Delta$ then $\Delta \mathrel{|\!\sim} \phi$.

Supraclassicality states that if ϕ follows from Γ in FOL, then it also follows according to $\mathrel{|\!\sim}$; i.e., $\mathrel{|\!\sim}$ extends $\models$ (the relation $\models$ is trivially supraclassical). Of the remaining conditions, the most straightforward is Reflexivity: It says that if ϕ belongs to the set Γ, then ϕ is a consequence of Γ. This is a very minimal requirement on a relation of logical consequence. We certainly would like all sentences in Γ to be inferable from Γ. It's not

clear in what sense a relation that fails to satisfy this requirement can be called a *consequence* relation.

Cut, a form of transitivity, is another crucial feature of consequence relations. Cut is as a conservativity principle: If ϕ is a consequence of Γ, then ψ is a consequence of Γ together with ϕ only if it is already a consequence of Γ alone. In other words, adjoining to Γ something that is already a consequence of Γ does not lead to any *increase* in inferential power. Cut can be regarded as the statement that the "length" of a proof does not affect the degree to which the assumptions support the conclusion. Where ϕ is already a consequence of Γ, if ψ can be inferred from Γ together with ϕ, then ψ can also be obtained by means of a longer "proof" that proceeds indirectly by first inferring ϕ. It is immediate to check that FOL satisfies Cut.

It is worth noting that many forms of probabilistic reasoning fail to satisfy Cut, precisely because the degree to which the premises support the conclusion is inversely correlated to the length of the proof. To see this, we adapt a well-known example. Let Ax stand for "x was born in Pennsylvania Dutch country," Bx stand for "x is a native speaker of German," and Cx stand for "x was born in Germany." Further, let Γ comprise the statements "most As are Bs," "most Bs are Cs," and Ax. Statements of the form "most As are Bs" are interpreted probabilistically as saying that the conditional probability of B given A is, say, greater than 50%; likewise, we say that Γ supports a statement ϕ if Γ assigns ϕ a probability $p > 50\%$.

Then Γ supports Bx, and Γ together with Bx supports Cx, but Γ by itself does not support Cx. Because Γ contains "most As are Bs" and Ax, it supports Bx (in the sense that the probability of Bx is greater than 50%); similarly, Γ together with Bx supports Cx; but Γ by itself cannot support Cx. Indeed, the probability of someone who was born in Pennsylvania Dutch country being born in Germany is arbitrarily close to zero. Examples of inductive reasoning such as the one just given cast some doubt on the possibility of coming up with a logically well-behaved relation of probabilistic consequence.

Special considerations apply to Monotony. Monotony states that if ϕ is a consequence of Γ then it is also a consequence of any set containing Γ (as a subset). The import of Monotony is that one cannot preempt conclusions by adding new premises to the inference. It is clear why FOL satisfies Monotony: Semantically, if ϕ is true on every interpretation on which all sentences of Γ are true, then ϕ is also true on every interpretation on which all sentences in a larger set Δ are true (similarly, proof theoretically, if there is a proof of ϕ, all of whose assumptions are drawn from Γ,

then there is also a proof of ϕ – indeed, the same proof – all of whose assumptions are drawn from Δ).

Many people consider this feature of FOL as inadequate to capture a whole class of inferences typical of everyday (as opposed to mathematical or formal) reasoning and therefore question the descriptive adequacy of FOL when it comes to representing commonsense inferences. In everyday life, we quite often reach conclusions tentatively, only to retract them in the light of further information. Here are some typical examples of essentially nonmonotonic reasoning patterns.

TAXONOMIES. Taxonomic knowledge is essentially hierarchical, with superclasses subsuming smaller ones: Poodles are dogs, and dogs are mammals. In general, subclasses inherit features from superclasses: All mammals have lungs, and because dogs are mammals, dogs have lungs as well. However, taxonomic knowledge is seldom strict, in that feature inheritance is prone to exceptions: Birds fly, but penguins (a special kind of bird) are an exception. Similarly, mammals don't fly, but bats (a special kind of mammal) are an exception.

It would be unwieldy (to say the least) to provide an exhaustive listing of all the exceptions for each subclass–superclass pair. It is therefore natural to interpret inheritance *defeasibly*, on the assumption that subclasses inherit features from their superclasses, unless this is explicitly blocked. For instance, when told that Stellaluna is a mammal, we infer that she does not fly, because mammals, by and large, don't fly. But the conclusion that Stellaluna doesn't fly can be retracted when we learn that Stellaluna is a bat, because bats are a specific kind of mammal, and they do fly. So we infer that Stellaluna does fly after all. This process can be further iterated. We can learn, for instance, that Stellaluna is a baby bat and that therefore she does not know how to fly yet. Such complex patterns of defeasible reasoning are beyond the reach of FOL, which is, by its very nature, monotonic.

CLOSED WORLD. Some of the earliest examples motivating defeasible inference come from database theory. Suppose you want to travel from Oshkosh to Minsk and therefore talk with your travel agent who, after querying the airline database, informs you that there are no direct flights. The travel agent doesn't actually *know* this, as the airline database contains explicit information only about existing flights. However, the database incorporates a *closed-world assumption* to the effect that the database is complete. But the conclusion that there are no direct connections between Oshkosh and Minsk is defeasible, as it could be retracted on expansion of the database.

DIAGNOSTICS. When complex devices fail, it is reasonable to assume that the failure of a smallest set of components is responsible for the observed behavior. If the failure of any two out of three components A, B, and C, can explain the device's failure, it is assumed that not all three components simultaneously fail, an assumption that can be retracted in the light of further information (e.g., if replacement of A and B fails to restore the expected performance).

For these and similar reasons, people have striven, over the past 25 years or so, to devise nonmonotonic formalisms capable of representing defeasible inference. We will take a closer look at these formalisms in Section 1.3, but for now we want to consider the issue from a more abstract point of view.

When one gives up Monotony in favor of descriptive adequacy, the question arises of what formal properties of the consequence relation are to take the place of Monotony. Two such properties have been considered in the literature for an arbitrary consequence relation $\mid\!\sim$:

Cautious Monotony: If $\Gamma \mid\!\sim \phi$ and $\Gamma \mid\!\sim \psi$, then $\Gamma, \phi \mid\!\sim \psi$;
Rational Monotony: If $\Gamma \not\mid\!\sim \neg\phi$ and $\Gamma \mid\!\sim \psi$, then $\Gamma, \phi \mid\!\sim \psi$.

Both Cautious Monotony and the stronger principle of Rational Monotony are special cases of Monotony and are therefore not in the foreground as long as we restrict ourselves to the classical consequence relation $\models$ of FOL.

Although superficially similar, these principles are quite different. Cautious Monotony is the converse of Cut: It states that adding a consequence ϕ back into the premise set Γ does not lead to any *decrease* in inferential power. Cautious Monotony tells us that inference is a cumulative enterprise: We can keep drawing consequences that can in turn be used as additional premises, without affecting the set of conclusions. Together with Cut, Cautious Monotony says that if ϕ is a consequence of Γ then for any proposition ψ, ψ is a consequence of Γ if and only if it is a consequence of Γ together with ϕ. In other words, as pointed out by Kraus, Lehman, and Magidor (1990, p. 178), if the new facts turned out already to be expected to be true, nothing should change in our belief system. It also turns out that Cautious Monotony has a nice semantic characterization: The just-cited article by Kraus et al. (1990) provides a system **C** (with Cautious Monotony among its axioms), which is proved sound and complete with respect to entailment over suitably defined *preferential models*, having a preferential ordering $\prec$ between states. In fact, it has been often pointed out that Reflexivity, Cut, and Cautious Monotony are critical properties

for any well-behaved nonmonotonic consequence relation (see Gabbay, Hogger, and Robinson, 1994; Stalnaker, 1994).

The status of Rational Monotony is much more problematic. As we observed, Rational Monotony can be regarded as a strengthening of Cautious Monotony, and, like the latter, it is a special case of Monotony. A case for Rational Monotony is forcefully made in Lehman and Magidor (1992, p. 20), as follows. Let p, q, and r be distinct propositional variables, and suppose that $p \mathrel{|\!\sim} q$ (for instance, because it is explicitly contained in our knowledge base); then we would intuitively expect also $p, r \mathrel{|\!\sim} q$, as r cannot possibly provide any information about whether p is satisfied or not (and in particular $p \not\mathrel{|\!\sim} \neg r$). Observe that there are relevance considerations at work here. The reason that $p, r \mathrel{|\!\sim} q$ appears plausible to us is that the sentences involved are atomic and therefore none of them is relevant for the truth of any of the others.

We will come back to this issue of relevance in Section 1.6, but for now we observe that there are reasons to think that Rational Monotony might not be a correct feature of a nonmonotonic consequence relation after all. Stalnaker (1994, p. 19) adapts a counterexample drawn from the literature on conditionals. Consider three composers: Verdi, Bizet, and Satie. Suppose that we initially accept (correctly but defeasibly) that Verdi is Italian, whereas Bizet and Satie are French. Suppose now that we are told by a reliable source of information that Verdi and Bizet are compatriots. This leads us no longer to endorse the propositions that Verdi is Italian (because he could be French), and that Bizet is French (because he could be Italian); but we would still draw the defeasible consequence that Satie is French, because nothing that we have learned conflicts with it. By letting $I(v)$, $F(b)$, and $F(s)$ represent our initial beliefs about the nationality of the three composers, and $C(v, b)$ represent that Verdi and Bizet are compatriots, the situation could be represented as follows:

$$C(v, b) \mathrel{|\!\sim} F(s).$$

Now consider the proposition $C(v, s)$ that Verdi and Satie are compatriots. Before learning that $C(v, b)$ we would be inclined to reject the proposition $C(v, s)$ because we endorse $I(v)$ and $F(s)$, but after learning that Verdi and Bizet are compatriots, we can no longer endorse $I(v)$, and therefore we no longer reject $C(v, s)$. The situation then is as follows:

$$C(v, b) \not\mathrel{|\!\sim} \neg C(v, s).$$

However, if we added $C(v, s)$ to our stock of beliefs, we would lose the inference to $F(s)$: In the context of $C(v, b)$, the proposition $C(v, s)$ is

equivalent to the statement that all three composers have the same nationality, and this leads us to suspend our assent to the proposition $F(s)$. In other words, and contrary to Rational Monotony,

$$C(v, b), C(v, s) \not\vdash F(s).$$

Thus we have a counterexample to Rational Monotony. On the other hand, there appear to be no reasons to reject Cautious Monotony, which is in fact a characteristic feature of our reasoning process. In this way we come to identify four crucial properties of a nonmonotonic consequence relation: Supraclassicality, Reflexivity, Cut, and Cautious Monotony.

1.3 NONMONOTONIC LOGICS

As was mentioned, over the past 25 years or so, a number of socalled *nonmonotonic* logical frameworks have emerged, expressly devised for the purpose of representing defeasible reasoning. The development of such frameworks represents one of the most significant developments both in logic and artificial intelligence and has wide-ranging consequences for our philosophical understanding of argumentation and inference.

Pioneering work in the field of nonmonotonic logics was carried out beginning in the late 1970s by (among others) J. McCarthy, D. McDermott, J. Doyle, and R. Reiter (see Ginsberg, 1987, for a collection of early papers in the field). With these efforts, the realization (which was hardly new) that ordinary FOL was inadequate to represent defeasible reasoning was for the first time accompanied by several proposals of formal frameworks within which one could at least begin to talk about defeasible inferences in a precise way, with the long-term goal of providing for defeasible reasoning an account that could at least approximate the degree of success achieved by FOL in the formalization of mathematical reasoning. The publication of a monographic issue of the *Artificial Intelligence Journal* in 1980 can be regarded as the "coming of age" of defeasible formalisms.

The development of nonmonotonic logics has been guided all along by a rich supply of examples. Many of these examples share the feature of an attempted *minimization* of the extension of a particular predicate (a minimization that is not, in general, representable in FOL, or at least not in a natural way). For instance, recall the travel agent example that was used in the preceding section in discussing the closed-world assumption: What we have in this example is an attempt to *minimize* the extension of the predicate "flight between." And, of course, such a minimization needs

to take place not with respect to what the database explicitly contains but with respect to what it implies.

The idea of minimization is at the basis of one of the earliest nonmonotonic formalisms, McCarthy's *circumscription*. Circumscription makes explicit the intuition that, all other things being equal, extensions of predicates should be *minimal*. Again, consider principles such as "all normal birds fly." Here we are trying to minimize the extension of the abnormality predicate and assume that a given bird is normal unless we have positive information to the contrary. Formally, this can be represented using second-order logic. In second-order logic, in contrast to FOL, one is allowed to explicitly quantify over predicates, forming sentences such as $\exists P \forall x\, Px$ ("there is a universal predicate") or $\forall P(Pa \leftrightarrow Pb)$ ("a and b are indiscernible"). In circumscription, given predicates P and Q, we abbreviate $\forall x(Px \rightarrow Qx)$ ("all Ps are Qs") as $P \leq Q$, and likewise we abbreviate $P \leq Q \wedge Q \not\leq P$ as $P < Q$. If $A(P)$ is a formula containing occurrences of a predicate P, then the circumscription of P in A is the following second-order sentence $A^*(P)$:

$$A(P) \wedge \neg \exists Q[A(Q) \wedge Q < P].$$

$A^*(P)$ says that P satisfies A and that no smaller predicate does. Let Px be the predicate "x is abnormal," and let $A(P)$ be the sentence "all normal birds fly." Then the sentence "Tweety is a bird," together with $A^*(P)$ implies the sentence "Tweety flies," for the circumscription axiom forces the extension of P to be empty, so that "Tweety is normal" is automatically true. In terms of consequence relations, circumscription allows us to define, for each predicate P, a nonmonotonic relation $A(P) \mathrel{|\!\sim} \phi$ that holds precisely when $A^*(P) \models \phi$. (This basic form of circumscription has been generalized, for in practice, one needs to minimize the extension of a predicate while allowing the extension of certain other predicates to vary.) From the point of view of applications, however, circumscription has a major shortcoming because of the second-order nature of $A^*(P)$. In general, second-order logic does not have a complete inference procedure: The price one pays for the greater expressive power of second-order logic is that there are no complete axiomatizations, as we have for FOL. It follows that it is impossible to determine whether $A(P) \mathrel{|\!\sim} \phi$ [except in special cases in which $A^*(P)$ happens to be in fact equivalent to a first-order sentence (see Lifschitz, 1987)].

There is another family of approaches to defeasible reasoning that makes use of a *modal* apparatus, most notably *autoepistemic logics*. Modal