Chapter 9

Inference Rules

In this chapter, several representative topics are discussed to show the difference between the inference rules of NARS and those in other theories.

9.1 Deduction

Deduction is the type of inference that has been studied most thoroughly. However, there are still problems when the knowledge and resources of the system are insufficient. Here the *reference class* problem is discussed as an example.

9.1.1 Deduction with reference classes

How do we predict whether an individual has a certain property, if direct observation is impossible? A useful method is to look for a "reference class." The class should include the individual as an instance, and we should know something about how often the instances of the class have the desired property, or whether its typical instances have it. Then, the prediction can be done by letting the instance "inherit" the information from the class.

In reasoning under uncertainty, there are (at least) two groups of approaches that use this type of inference: non-monotonic logics [Touretzky, 1986], and probabilistic reasoning systems [Pearl, 1988].

In non-monotonic logics, if the only relevant knowledge is "S is an instance of R" and "Normally, R's instances have the property Q," a defeasible conclusion is "S has the property Q."

In probabilistic reasoning systems, under the subjective interpretation of probability, if the only relevant knowledge is "S is an instance of R" and "The probability for R's instances to have the property Q is p," a plausible conclusion would be "The probability for S to have the property Q is p."

Now a problem appears: if S belongs to two classes R_1 and R_2 at the same time, and the two classes lead to different predictions about whether (or how probable) S has the property Q, what conclusion can we reach? In different contexts, the problem is referred to as "multiple inheritance problem," "multiple extension problem," or "reference class problem." [Grosof, 1990, Kyburg, 1983, Neufeld, 1989, Pearl, 1988, Poole, 1985, Reichenbach, 1949, Touretzky, 1986].

Though the above theories treat the problem differently, they have something in common: None of them suggest a general solution to the problem, though they agree on a special case: if R_2 is a (proper) subset of R_1, R_2 is the correct reference class to be used.

Let us see two examples.

1. "Since Clyde is a royal elephant, and royal elephants are not gray, Clyde is not gray. On the other hand, we could argue that Clyde is a royal elephant, royal elephants are elephants, and elephants are gray, so Clyde is gray. Apparently there is a contradiction here. But intuitively we feel that Clyde is not gray, even though he is an elephant, because he is a special type of elephant: a royal elephant." [Touretzky, 1986].

2. "If you know the survival rate for 40-year old American male to be 0.990, and also that the survival rate for 40-year old American male white-collar workers to be 0.995, then, other things being equal, it is the latter that should constrain your beliefs and enter your utility calculations concerning the particular 40 year old male white-collar worker John Smith." [Kyburg, 1983].

Let us call this principle "specificity priority principle." It looks quite reasonable, and it is not hard to find many examples to show that

we do apply such a principle in common sense reasoning. However, the following questions are still open:

1. Why is the principle correct? Can it be justified by more basic postulates?

2. Beside specificity, what are the "other things" that influence the priority of a reference class?

3. When neither reference class is more specific than the other, what should be done?

For the first question, Reichenbach made it a matter of definition by "regarding the individual case as the limit of classes becoming gradually narrower and narrower" [Reichenbach, 1949]; Pearl said it is because "the influence of the remote ancestors is summarized by the direct parents." [Pearl, 1988].

For the second question, Reichenbach said we need to have complete statistical knowledge on the reference class, that is, the probability for R to be Q should be supported by good statistical data [Reichenbach, 1949]. In non-monotonic logics, this corresponds to sufficient evidence which can determine what properties a *normal* instance of the class has.

For the third question, few words are said, except Reichenbach's suggestion to "look for a larger number of cases in the narrowest common class at your disposal." [Reichenbach, 1949]

9.1.2 A thought experiment

Let us reconstruct Kyburg's example in the following way: Imaging that you are working for a life insurance company, and you need to predict whether John Smith can live to 40. You have John's personal information, and for some special reasons (such as you just woke up from a 200-year-long sleep or you are actually an extraterrestrial spy), you have no background knowledge about the survival rates at 40 for various groups of people. Fortunately, you have access to personal files of some Americans, who are alive or died in recent years, and you decide to make the prediction by the "reference class method" defined above.

At first, knowing that John is a male, you begin to build the first reference class R_1 by picking up some files randomly. R_1 consists of two subsets: P_1 includes the positive evidence for John's survival, that is, American males who are more than 40 years old (including those who are already deceased), and N_1 includes the negative evidence, that is, those who died before 40. You should keep in mind that American males who are alive and younger than 40 (including John himself) are neither positive evidence nor negative evidence for the prediction, so they do not belong to R_1.

If you weight everyone equally (and why wouldn't you?), your prediction should be determined by the relative size of P_1 and N_1. Let us say $|P_1| > |N_1|$. Therefore you predict that John Smith can live to 40.

After returning the files, you have a new idea: why not consider the fact that John is, among other things, a white-collar worker? So you build another reference class R_2 similarly. Let us assume, unfortunately, this time you find that $|P_2| < |N_2|$. Here you meet the reference class problem: to see John as a "male" and a "male white-collar worker" will lead to different predictions.

If we apply the *specificity priority principle* here, the result should be dominated by R_2, since "male white-collar worker" is a proper subset of "male." However, it is easy to find a situation to show that sometimes the result is counter-intuitive. If you have looked through 1000 files, and all of them are males and live to 40, and after that you find 1 male white-collar worker who died at 35, will you predict that John will die before 40? It seems very unlikely.

Does this mean that the specificity priority principle is wrong? Of course not. Sample size is obviously one of the "other things" that influence the priority of a reference class. One sample is far from enough to tell us about how a "typical" or "normal" instance looks like, or to support a statistical assertion on the instances. In such a case, the principle is inapplicable, since there is another relevant difference between the two reference classes, beside their specificities.

If you have to make predictions in such an environment, what will you do? Let us consider a simple psychological experiment. Assuming R_1 includes positive evidence only (that is, $R_1 = P_1$; no male is found to have died before 40), but R_2 includes negative evidence only (that is, $R_2 = N_2$, no male white-collar worker is found to be alive at 40).

Even before really carrying out such an experiment on human subjects, I am confident to make the following prediction: If $|P_1|$ is fixed at a big number (say 1000), and $|N_2|$ is increased one by one, starting from 1, the predictions made by subjects will be positive before $|N_2|$ reaching a certain point, and negative after reaching that point. That critical point may vary from person to person, but is always smaller than $|P_1|$.

The "sample size effect" can also be used to answer the following question: If a more specific reference class is always better, why do not we simply use the *most specific reference class*, defined by all available properties of John Smith? The reason is simple: in most situations such a class is *empty* — nobody is similar to John to such an extent. With more and more properties used to define a reference class, the extension of the class becomes narrower and narrower. As a result, fewer and fewer samples can be found to support or discredit our prediction. From this point of view, specificity is not preferred.

Previously, I talked about the reference classes R_1 and R_2, as if they are accurately defined. Obviously this is a simplification. Though we can ignore the boundary cases for "male," the fuzziness in "white-collar worker" cannot be neglected so easily. As argued by fuzzy set theory [Zadeh, 1965] and prototype theory [Rosch, 1973], whether an instance belongs to a concept is usually a matter of degree. This membership function is also related to the current issue: if John can be referred to as a "white-collar worker," but not a typical one, the influence of R_2 will be reduced.

From the above analysis, we can see that the previous solutions from non-monotonic logic and probability theory ignored several important factors when handling deduction with reference classes.

9.1.3 The NARS solution

Now Let us see how NARS treats the reference class problem.

Putting the previous example into Narsese, the premises are:

$$\begin{array}{ll} J_1: & \{S\} \rightarrow R_1 \ <f_1, c_1> \\ J_2: & \{S\} \rightarrow R_2 \ <f_2, c_2> \\ J_3: & (\#x \rightarrow R_1) \Rightarrow (\#x \rightarrow Q) \ <f_3, c_3> \\ J_4: & (\#x \rightarrow R_2) \Rightarrow (\neg(\#x \rightarrow Q)) \ <f_4, c_4> \end{array}$$

Since John shares one property with R_1 ("male") and two properties with R_2 ("male" and "white-collar worker"), we have $w_1 = w_1^+ = 1$ and $w_2 = w_2^+ = 2$. It follows that (assuming $k = 1$) $f_1 = f_2 = 1$, $c_1 = 1/2$, and $c_2 = 2/3$. Under the assumption that R_1 consists of 1000 positive samples, we have $f_3 = 1$ and $c_3 = 1000/1001$. Let us say that R_2 includes negative samples only, but leaves the number of samples, n, as a variable, to see how it affects the final evaluation of $\{S\} \rightarrow Q$. Therefore, we have $f_4 = 1$ and $c_4 = n/(n + 1)$.

Applying the deduction rule, from J_1, J_3 and J_2, J_4, respectively, we get

$$J_5: \ \{S\} \rightarrow Q \ <1, c_1c_3>$$
$$J_6: \ \{S\} \rightarrow Q \ <0, c_2c_4>$$

Since the knowledge that "John is male" is used to evaluate both J_1 and J_2, and they are used in the derivation of J_5 and J_6, respectively, the evidence for J_5 and J_6 is correlated. Therefore, they cannot be merged by the revision rule. Instead, the choice rule is applied to pick the judgment that has a higher confidence as the conclusion. Which reference class will win the competition?

By solving the inequality $c_1c_3 > c_2c_4$ (that is, $(1/2) \times (1000/1001) > (2/3) \times (n/(n + 1))$), we can see that

1. When $0 < n < 3$, R_1 is selected. The specificity priority of R_2 is undermined by the fact that the sample size of R_2 is too small.

2. When $n \geq 3$, R_2 is selected. The specificity priority can be established even by a pretty small sample size: with $|R_1| = 1000$ and $|R_2| = 3$, the prediction is still determined by R_2 due to its specificity.

If John is not a typical white-collar worker (i.e., $f_2 < 1$), R_2's confidence is smaller than c_2c_4, so it may need a bigger n for R_2 to be dominant.

Therefore, when NARS is selecting a reference class, several factors are balanced against one another, including specificity, typicality, sample size, and so on. It provides a generalization of the specificity priority principle, by taking more relevant factors into consideration.

NARS' approach is more general than the specificity priority principle in another way. The including of reference classes is only a special

case for two judgments to be based on correlated evidence. It follows that the specific priority principle is a special case of NARS' choice rule.

How about competing reference classes that do not involve correlated evidence? Let us say in the previous examples, R_1 is still for "male," but R_2 is changed for "smoker and white-collar worker." If the deduced judgments J_5 and J_6 are not based on correlated evidence in some other ways, the two judgments will be combined by the revision rule of NARS. Other things being equal, R_2 has a higher priority, since it matches better with John's properties. However, in this case a higher priority only means a higher *weight* in determining the frequency of the conclusion. The judgment from the other reference class is not ignored. In this situation, the reference class competing is solved not by *choosing one of them*, but by *combining the two*.

Let us see how NARS treats the "Nixon Diamond" discussed in the study of non-monotonic logics [Touretzky, 1984]. This example assumes we know that Nixon is a Quaker, and Quakers are pacifists. We also know that Nixon is a Republican, and Republicans are not pacifists. From the above knowledge alone, should we predict Nixon to be a pacifist or not? Putting into the previous framework, in this problem we have "Nixon" as S, "Quaker" as R_1, "Republican" as R_2, and "Pacifist" as Q. By deduction, two conflicting judgments J_5 ("Nixon is a pacifist") and J_6 ("Nixon is not a pacifist") can be derived as in the previous example.

Since we can assume the un-correlation of evidence of the judgments (R_1 and R_2 have no known relation), J_5 and J_6 will be combined by the revision rule, and the result depends on the truth value of the premises.

1. If $f_1 = f_2$, $c_1 = c_2$, $f_3 = 1 - f_4$, and $c_3 = c_4$, for the conclusion we will get $f = 0.5$. That is, when the positive evidence and the negative evidence exactly balance with each other, the system is indifferent between a positive prediction and a negative prediction.

2. If $c_1 > c_2$, and the other conditions as in (1), we will get $f > 0.5$. That is, when Nixon shares more property with Quaker, the system will put more weight on the conclusions suggested by the evidence about Quaker.

3. If $f_3 > 1 - f_4$ or $c_3 > c_4$, and the other conditions as in (1), we will get $f > 0.5$. That is, when we have stronger statistical data about Quaker, the system will put more weight on the conclusions suggested by the evidence about Quaker, too.

In any situation, what NARS does is to combine the evidence from both sources. Even if "Quaker" is given a higher priority, the evidence provided by "Republican" still has its effect on the result. On the other hand, this kind of conflict does not always (though sometimes it does) cause complete indifference or ambiguity, as it does in non-monotonic logics [Touretzky, 1986].

In summary, compared with non-monotonic logics and probability theory, the processing of the reference class problem in NARS has the following characteristics:

1. While still following the specificity priority principle, several factors, such as sample size and degree of membership, are taken into account to quantitatively determine the priority of a reference class, and all the factors are projected into a common dimension, that is, the amount of evidence.

2. The specificity priority principle has been generalized into a "confidence priority principle," which will pick a judgment with the highest confidence among the competing ones, supported by correlated evidence. As discussed above, specificity is one way to get a high confidence, while the inclusion relation between reference classes causes evidence correlation.

3. When conflicting judgments come from different sources, the revision rule is applied to combine them by summarizing the evidence. This operation is not directly available in non-monotonic logics and probability theory.

Why cannot similar things be done in non-monotonic logics and probability theory? One of the major reasons is that the *confidence* (or equivalently, *amount of evidence*) measurement cannot be easily introduced there. From the view point of NARS, the confidence of all the default rules (in non-monotonic logics) and probability assignments (in

probability theory) is 1, that is, they cannot be revised by accommodating its current evaluation to new evidence.

Therefore, the reference class problem provides another piece of evidence for the previous criticism on non-monotonic logic (Section 8.1) and probabilistic logic (Section 8.3), as general solutions to the reasoning under uncertainty problem.

9.2 Induction

Induction is a major topic in this book, because it has been called "the glory of science and the scandal of philosophy," as well as because the basic ideas of NARS to a large extent were formed during my study on the problem of induction.

9.2.1 The problem of induction

The term "induction" is usually used to denote the inference that derives *general* knowledge from *specific* knowledge. There are some people who call all non-deductive inferences "induction," but in this way the category includes too many heterogeneous instances to be studied fruitfully.

There are three major academic traditions in the study of induction. The *philosophical/logical* study concentrates on the formalization and justification of induction; the *psychological* study concentrates on the description and explanation of induction in the human mind; and the *computational* study concentrates on the implementation of induction in computer systems.

Though Aristotle mentioned induction as the method by which general primary premises can be obtained, he did not develop a theory for this type of inference, as he did for deduction. It was Bacon who for the first time proposed a systematical inductive method, with the hope that it could provide a general methodology for empirical science [Cohen, 1989]. However, such an approach was seriously challenged by Hume, who argued that the inferences that extend past experience to future situations cannot have a logical justification [Hume, 1748]. After Hume, most philosophical and logical work on induction are about the

justification of the process. The mainstream approach is to use probability theory, with the hope that though inductive conclusions cannot be absolutely true, they can have certain probabilities [Carnap, 1950].

In recent years, the study of induction has been enriched by AI researchers. With computer systems as tools and platforms, different formalizations and algorithms are proposed and tested. In terms of the formal language used, we can further divide the existing approaches in this domain into three "families."

The first family uses propositional logic and probability theory. Let us say that S is a proposition space and P is a probability distribution function on it. Induction is defined in this situation as the operation of determining $P(H|E)$, where H is a hypothesis and E is available evidence, and both belong to S. The inference — or more precisely, calculation — is carried out according to probability theory in general, and Bayes' theorem in particular. This family is the mainstream of the philosophical and logical tradition of induction study [Keynes, 1921, Carnap, 1950, Good, 1983], and it has been inherited by the Bayesian school in AI [Korb, 1995, Pearl, 1988].

The second family uses first-order predicate logic. Let us say that K is the background knowledge of the system, and E is available evidence (both K and E are sets of proposition). Induction is defined in this situation as the operation of finding a proposition H that implies E and is also consistent with K. Because the inference from H and K to E is deduction, induction thus defined, as the inference from E and K to H, is often referred to as "reverse deduction." This family is very influential in machine learning [Michalski, 1993].

The third family uses term logic. Though Aristotle discussed induction briefly in his work [Aristotle, 1989], it was Peirce who first defined different types of inference in term logic, roughly in the following manner [Peirce, 1931]:

deduction	**induction**	**abduction**
$M \rightarrow P$ $S \rightarrow M$ ——— $S \rightarrow P$	$M \rightarrow P$ $M \rightarrow S$ ——— $S \rightarrow P$	$P \rightarrow M$ $S \rightarrow M$ ——— $S \rightarrow P$

One interesting fact is that though Peirce's distinction of deduction, induction, and abduction is widely accepted, his formalization in term logic is seldom followed. Instead, the above definition is rephrased within the frame of predicate logic [Michalski, 1993]. We will see the subtle difference between these two formalizations later.

Obviously, NARS belongs to the term-logic family. Now let us see how NARS answers the questions about the aspects of induction.

9.2.2 To represent inductive conclusions

As mentioned previously, NARS represents all knowledge, including inductive conclusions, in Narsese. The simplest sentence has the form of "$S \rightarrow P$," with a truth value attached, which is determined according to the past experience of the system.

To decide truth according to the available evidence or according to a set of axioms is fundamentally different. In the former situation, no decision is final in the sense that it cannot be revised by future evidence. Each piece of evidence contributes, to a certain extent, to the evaluation of truth value. Therefore, truth value is always a matter of degree in a system like NARS.

This opinion is against a well-known conclusion proposed by Popper. He claimed that there is an asymmetry between verifiability and falsifiability — "a positive decision can only temporarily support the theory, for subsequent negative decisions may always overthrow it" [Popper, 1959].

The crucial point here is: what is the content of a *general statement*, or, in Popper's words, a *theory*?

According to my opinion, "Ravens are black" is a general statement, for which a black raven is a piece of positive (affirmative) evidence, and a non-black (e.g., white) raven is a piece of negative (rejective) evidence — the former verify an inheritance relation "$raven \rightarrow [black]$" to a certain extent, while the latter falsify it, also to a certain extent. When we say that "All ravens are black," it means that according to our experience, the inheritance relation between the two terms only has positive evidence, but no negative evidence. In this case, the truth value of the statement is still a matter of degree, though the frequency value happens to be at its maximum, 1.

What Popper referred to as theory are *universal statements*. Accordingly, when we say "All ravens are black," we mean that all ravens in the whole universe, known or unknown, are black. Such a statement can only be true or false, and there is no middle ground (if we ignore the fuzziness of the terms). We know the statement is false as soon as we find a non-black raven, but we need to exhaust all ravens in the universe to know it is true.

Such a formalization of inductive conclusions is shared by the Baconian tradition of induction [Cohen, 1989]. According to an approach proposed by Cohen, induction is a sequence of tests with increasing complexity, and the (Baconian) probability of a hypothesis indicates how many tests the hypothesis passed in the process.

If we accept the above definition of scientific theory, all conclusions of Popper and Cohen follow logically. However, why should we accept the definition? As a matter of fact, many empirical scientific theories have counterexamples, and we do not throw them away [Kuhn, 1970]. It is even more obvious when we consider our common-sense knowledge. A general statement like "Ravens are black" works well as our guide of life, even when we know that it has counterexamples. Such a statement can be applied to predict new situations, though its truth value is determined by past experience. We do hope to establish theories that have no known counterexamples, but it does not mean that theories with known counterexamples cannot be used for various practical purposes. Only in mathematics, where truth values are determined according to fixed axioms, do universal statements become available.

The above argument also serves as a criticism to the AI induction projects within the framework of binary logic [Korb, 1995]. To define induction as "finding a pattern to fit *all* data" makes it a luxury that can only be enjoyed in a laboratory. Though such an approach can produce research results, these results are hardly applicable to practical situations. Also, this over-idealization makes the process fundamentally different from the generalizations happening in the human mind. It is not even appropriate to justify this approach as "a preliminary step toward more complex studies," because when giving up the idea that "an inductive conclusion can be falsified once for all," the situation will become so different that the previous results are hardly useful at all.

Because in NARS truth values are determined by available evidence, we need to first precisely define what is counted as evidence and how evidence is quantitatively measured.

Though it is natural to say that a black raven is a piece of positive evidence for "Ravens are black," and a white raven is its negative evidence, Hempel points out that such a treatment leads to counter-intuitive results [Hempel, 1943]. If "Ravens are black" is formulated as $(\forall x)(Raven(x) \rightarrow Black(x))$, a green shirt will also be counted as a piece of positive evidence for the sentence, because it confirms the "logically equivalent" sentence $(\forall x)(\neg Black(x) \rightarrow \neg Raven(x))$ ("Non-black things are non-ravens"). Such a result is highly counterintuitive, and may cause many problems (for example, a green shirt is also a piece of positive evidence for "Ravens are white," for exactly the same reason).

Here I will not discuss the various solutions proposed for this paradox. It is enough to say that almost all of those attempts are still within the framework of predicate logic, whereas in the following we can see that the problem does not appear in term logics like NARS.

As we already know, in Narsese "Ravens are black" can be represented as "$(\#x \rightarrow raven) \Rightarrow (\#x \rightarrow [black])$." For this statement, "black ravens" are positive evidence, "non-black ravens" are negative evidence, and "non-ravens" are not directly relevant (according to the definition of evidence for implication statements in Chapter 5). On the other hand, "Non-black things are non-ravens" can be represented in Narsese as "$(\neg(\#x \rightarrow [black])) \Rightarrow (\neg(\#x \rightarrow raven))$." For it, "non-black non-raven" are positive evidence, "non-black ravens" are negative evidence, and "black things" are not directly relevant.

Comparing the two statements, we see that, in the terminology of NARS, they have the same negative evidence, but different positive evidence (as discussed in Section 5.1.4). In a binary logic, the truth value of a statement only indicates whether there is any negative evidence, so these two statements have the same truth value, or are "equivalent." In a logic where truth value is determined by both positive and negative evidence, they may have different truth values, and are no longer equivalent.

Therefore, what Hempel's paradox reveals is that "equivalent statements" in a binary logic do not necessarily have the same truth value

when the system is extended into a multi-valued logic. This problem does not appear in NARS, because here the evidence for "Ravens are black" and "Non-black things are non-ravens" are different (therefore they usually have different truth values). In NARS, the existence of a green shirt is not directly relevant to whether ravens are black, just as our intuition tells us.

9.2.3 To generate inductive conclusions

The induction rule defined in Section 3.3.4 has the following form:

$$\{M \rightarrow P <f_1,\, c_1>,\; M \rightarrow S <f_2,\, c_2>\} \vdash S \rightarrow P <f_1,\, \frac{f_2 c_1 c_2}{f_2 c_1 c_2 + k}>$$

This section explains why the rule is defined in this way.

To show how the rule works on a concrete example, let us to back to the example used in Section 3.3.6, and let P be "*swimmer*," and S be "*bird*." To see if the rule makes intuitive sense, let us at first consider the following special situations.

1. When $f_1 = c_1 = f_2 = c_2 = 1$, M is a piece of (idealized) positive evidence for the conclusion. According to the previous definitions, in this case we have $w^+ = w = 1$ for the conclusion — that is, $f = 1,\; c = 1/(1 + k)$. For the "Birds are swimmers" example, here M is a swimmer bird, such as a swan.

2. When $f_1 = 0,\; c_1 = f_2 = c_2 = 1$, M is a piece of (idealized) negative evidence for the conclusion. According to the previous definitions, in this case we have $w^- = w = 1$ for the conclusion — that is, $f = 0,\; c = 1/(1 + k)$. For the "Birds are swimmers" example, here M is a non-swimmer bird, such as a robin.

3. When $f_2 = 0$, M is not an instance of S. In this case, no matter it is an instance of P or not, it provides no evidence for the conclusion, therefore $w = 0,\; c = 0$, and f is undefined. For the "Birds are swimmers" example, here M is not a bird (but a dolphin, for example).

4. When c_1 or c_2 is 0, one of the premises gets no evidential support, so the conclusion gets no evidential support either. That means

$w = 0,\ c = 0$. For the "Birds are swimmers" example, here either whether M is a bird or whether M is swimmer is completely unknown.

From these boundary conditions of the truth value function for induction, if all the variables take boolean values (either 0 or 1), we get $f = f_1$ and $w = and(f_2, c_2, c_1)$, here *and* is the Boolean conjunction of the arguments.

To generalize the Boolean function into real numbers, *and* is replaced by multiplication, and that gives us $w = f_2 c_2 c_1$. Using the equation $c = w/(w + k)$, finally we get the truth-value function used in the induction rule.

Because in NARS the truth value indicates the relation between a statement and available evidence, induction is "ampliative" in the sense that its conclusions are more general than its premises, but it is also "summative" in the sense that the conclusions claim no more support than they actually get from the premises. Therefore the traditional distinction between these two types of induction does not apply here [Cohen, 1989, Popper, 1959] in its original form.

On the other hand, the distinction between "truth-preserving" and "ampliative" inferences appears in a different form. In NARS, the confidence of deductive conclusions have a upper bound of 1, and we already know that the upper bound for induction is $1/(1 + k)$, which is smaller than 1. If all premises are absolutely certain, so are their deductive conclusions, but this does not hold for their inductive conclusions.

It needs to be stressed again that the truth value of the conclusion indicates the support provided by the evidence, rather than whether the statement corresponds to a fact in the outside world. An adaptive system behaves according to its beliefs, not because they guarantee success (which is impossible, as Hume argued), but because it has to rely on its experience to survive, even though the experience may be biased or outdated — this is what "adaptation" means.

In summary, my solution to Hume's problem is to justify induction (and all other inference rules) according to an experience-grounded semantics and the notion of adaptation.

9.2.4 To conduct inductive inference

Another feature that distinguishes the induction rule of NARS from other induction systems is that the rule is able to generate and evaluate an inductive conclusion at the same time.

Traditionally, the generating and evaluating of inductive conclusions (or hypotheses) are treated as two separated processes. The most well-known arguments on this issue were provided by Carnap and Popper, though they hold opposing opinions on induction in general [Carnap, 1950, Popper, 1959]. The consensus is that from given evidence, there is no effective procedure to generate all the hypotheses supported by the evidence, therefore the discovery of a hypothesis is a *psychological* process, which contains an "irrational element" or "creative intuition." On the contrary, the evaluation of a given hypothesis, according to given evidence, is a *logical* process, following a well-defined algorithm.

The above opinion is in fact implicitly based on the specific language in which the inductive process is formalized. In probability theory, there is no way to get a unique hypothesis H from given evidence E for the purpose of induction, because for every proposition X in the proposition space, $P(X|E)$ can be calculated, at least in principle. In first-order predicate logic, there are usually many hypotheses H that imply the given evidence E, and also are consistent with background knowledge K. In both cases, some heuristics can be used to pick up an inductive conclusion that has some desired properties (simplicity, for instance), but these heuristics are not derived from the definition of the induction rule [Mitchell, 1980, Haussler, 1988].

In term logic, the situation is different. Here premises of an inductive inference must be a pair of judgments that share a common subject, and the premises uniquely determine an inductive conclusion. (Of course, there is also a symmetric inductive conclusion if we exchange the order of the premises.) Therefore, in NARS we do not need an "irrational element" or domain-dependent heuristics, and the discovery of a hypothesis, in the current sense, also follows logic.

In NARS, induction is unified with other types of inferences, in the sense that the premises used by the induction rule may be generated by the deduction (or abduction, and so on) rule, and that the conclusions

of the induction rule may be used as premises by the other rules. In particular, the revision rule may merge an inductive conclusion with a deductive (or abductive, and so on) conclusion.

Therefore, though NARS has an induction rule, it is not an "inductive logic," in the sense that it solves problems by induction only. An answer reported by NARS to the user is usually the cooperative result of several rules in a multi-step inference process. Though there are other "multi-strategy" inference models (which combine different types of inference), using first-order predicate logic [Michalski, 1993], attribute-value language [Giraud-Carrier and Martinez, 1995], or hybrid (symbolic-connectionist) representation [Sun, 1995], the term logic model, proposed by Peirce and extended in NARS, puts different types of inference in the same framework in a more natural, elegant, and consistent manner.

From the above discussion, we see that conclusions in NARS are based on different amounts of evidence, and, generally speaking, conclusions based on more evidence are preferred, because of their relative stability. However, since NARS is designed to be an open system, future evidence is always possible, therefore there is no way for the system to get "complete evidence" for an inductive conclusion.

A reasonable retreat is to use all evidence known to the system — the so-called "total evidence" [Carnap, 1950]. Unfortunately, this is also impossible, because NARS has insufficient resource. The system has to answer questions under a time pressure, which makes exhaustive search in knowledge space not affordable.

Moreover, in NARS the time pressure is variable, depending to the request of the user and the existence of other information-processing tasks. In this situation, even a predetermined "satisfying threshold" becomes inapplicable — such a threshold is sometimes too low and sometimes too high.

As described in Chapter 6, the control mechanism used in NARS is similar to an "anytime algorithm" [Dean and Boddy, 1988]. If the system is asked to evaluate the truth value of a statement, it reports the best conclusion (i.e., with the highest confidence) as soon as such a conclusion is found, then continues to look for a better one, until no resources are available for this task. In this way, from the user's point of view, the system may change its mind from time to time,

when new evidence is taken into consideration. The system will never say that "This is the final conclusion and I will stop working on the problem."

The above discussion is directly related to the "acceptance" problem in inductive logic [Kyburg, 1994]. As put by Cohen, "what level of support for a proposition, in the light of available evidence, justifies belief in its truth or acceptance of it as being true?" [Cohen, 1989]. In NARS, there is no such a thing as "accepted as being true." Judgments are true to different extents, and the system always follows the best-supported conclusion (compared with its rivals), no matter what its truth value is — the standard is relative and dynamic, not absolute and static. In this way, an inductive conclusion also benefits from the refutation of competing conclusions, which is stressed by the Baconian tradition of induction [Cohen, 1989] — though its truth value may not change in this process, its relative ranking becomes higher.

According to the definition given be Peirce, the difference among deduction, abduction, and induction is the position of the shared term in the two premises. This property of term logic makes it possible for NARS to combine different types of inference in a "knowledge-driven" manner. In each inference step, the system does not decide what rule to use, then look for corresponding knowledge. Instead, it picks up a task and a belief which share a term, and decides what rule to apply according to the position of the shared term (as described in Chapter 6). In general, an inference process in NARS consists of many steps. Each step carries out a certain type of inference, such as deduction, abduction, induction, and so on. These steps are linked together in run-time in a context-dependent manner, so the process does not follow a predetermined algorithm.

Therefore, NARS is not an "inductive machine" which uses an effective algorithm to generate inductive conclusions from given evidence. Carnap's argument against the possibility of this kind of machine [Carnap, 1950] is still valid. However, this argument does not prevent us from building a computer system that can do induction. The system does not have a general purpose induction algorithm, but can solve problems under its knowledge and resource constraints, and in the problem-solving activities there are inductive steps.

9.3 Abduction

Since in NARS abduction and induction are duals of each other (because extension and intension are duals of each other), most of the previous discussions on induction also apply to abduction. In the following, I will not repeat them, but focus on the special issues that distinguish my approach toward abduction from the other approaches.

9.3.1 Two definitions of abduction

Approaches of defining abduction can be classified into two types: syllogistic and inferential. An inferential definition identifies abduction as a type of inference *process* that carries out a certain cognitive function, such as explanation or hypothesis generation, while a syllogistic definition specifies it as a type of inference *step* with a specific pattern [Flach and Kakas, 2000].

As defined in NAL-1 and NAL-5, in NAL the distinction among deduction, abduction, and induction is formally specified at the inference-step level, according to the position of the shared term in the premises. Such a formal definition makes discussions about them clear and concrete.

To use a formal definition to distinguish various inference types does not prevent us from attributing them with different cognitive functions. Given the definition used in NAL, it is valid to say that among the three, only deduction can produce conclusive results, while the other two only produce tentative results. Both abduction and induction can be seen as "reversed deduction," and the former usually corresponds to explanation, and the latter to generalization. These descriptions are similar to the ones proposed as inferential definitions of the three types. However, in NAL these descriptions are *secondary*, derived from the syllogistic definition. This approach has the advantage of avoiding ambiguity and oversimplification in the definition, and at the same time preserve the intuitive meaning of the terms (i.e., deduction, abduction, and induction) associated with different types of inference.

Though abduction defined in NAL usually can be interpreted as "explanation," to define "abduction" as "explanation" at the inference-process level is a quite different decision. This is the case because what

we called "explanation" in everyday thinking may include complex cognitive processes where multiple types of inference are involved. Therefore, to abstract such a process into a consistent and non-trivial pattern is not an easy thing to do, if it is possible at all.

For the same reason, to define abduction as "inference toward the best explanation" makes things even harder, because besides the derivation of explanations, this definition further requires evaluation of explanations and comparison of competing candidates. In this process, many other factors should be taken into account, such as simplicity, surprisingness to the system, and relevance to the given context. If we cover all of these issues under "abduction," it becomes such a complex process that few concrete conclusion can be made. Such a definition is not wrong, but not very useful.

9.3.2 Multi-valued vs. binary

In the framework of binary logic, abduction is usually defined formally as "reverse deduction" which starts from a given conclusion and background knowledge to find a premise that is consistent with the background knowledge, and derives the conclusion deductively.

Such a definition is logically sound, and can lead to fruitful results. However, it ignores certain factors that are crucial for a system working with insufficient knowledge and resources.

In empirical science and everyday life, we usually do not throw away theories that have known counterexamples and inexplicable phenomena. If we do that, there is hardly anything left. Since we usually have insufficient knowledge in these domains, we have to live with imperfect knowledge, because they are still far better than random guesses.

When selecting among competing explanations and hypothesis, measurement of (positive and negative) evidence becomes necessary — if no explanation is perfect, then the one with more positive evidence and less negative evidence is preferred, which is what is measured by the *frequency* defined in NAL. Since evidence may come from time to time, incremental revision becomes inevitable, which requires the amount of evidence to be represented in some way, and this is how the *confidence* measurement becomes necessary.

These measurements enrich our understanding of the inference rules. In the truth-value functions, we can see that the fundamental difference between deductive inference and non-deductive (such as abductive or inductive) inference is in the confidence (not the frequency) of the conclusion. In deduction, if both premises are completely true, so is the conclusion. However, in abduction and induction, the confidence of the conclusion is much lower in this situation, meaning that the conclusion is tentative even when the premises are certain, and can be revised by new evidence.

To ignore quantity of evidence means it will be hard for the system to distinguish hypotheses that have a little of negative evidence from those that have a lot. Even for a hypothesis for which only positive evidence has been found, the amount of evidence still matters — a hypothesis confirmed only once is quite different from a hypothesis conformed a million times. For these reasons, to study abduction in binary logic is not wrong, but not very useful. Unfortunately, it is still the most common approach to the topic of "abduction" in the current AI research.

9.4 Implication

This section addresses two topics in higher-order inference, both about the implication relation.

9.4.1 Implication and relevance

A look at the grammar of Narsese reveals the origin of the intuition behind the design: first-order NAL is closely related to set theory,[1] and higher-order NAL is closely related to propositional logic — both contain logical constants for *negation* ("¬"), *conjunction* ("∧"), *disjunction* ("∨"), *implication* ("⇒"), and *equivalence* ("⇔").

Though the intuitive meanings of the above constants are similar in these two logic systems, there is a fundamental difference. In propositional logic, all of the five logic constants are truth-functional operators that form compound propositions, whose truth values are

[1] Their relation will be discussed in the next chapter.

fully determined by those of their components. In NAL, on the contrary, the five constants belong to two different categories. The first three are *term operators* that form compound (higher-order) terms; the last two are (higher-order) *relations* (i.e., *copulas*), which are not purely truth-functional. Consequently, when P and Q are both Narsese statements, so do $(P \Rightarrow Q)$ and $((\neg P) \vee Q)$, but the latter two are no longer equivalent to each other in NAL.

In propositional logic, $P \Rightarrow Q$ is equivalent to $(\neg P) \vee Q$. Though this equivalence is useful for various purposes, it suffers from the well-known "implication paradox," which says that $P \Rightarrow Q$ is true when P is false (Q can be anything) or Q is true (P can be anything). Though logically consistent with propositional logic, this result is highly counter-intuitive, and it gives people a feeling that some important thing is missing in the definition of implication in propositional logic — P and Q should be somehow *relevant* to each other, which is assumed by the "if ... then ..." structure in natural languages [Copi, 1982].

A whole branch of logic, *relevant logic* [Read, 1989], has been developed specially for this issue. I will not review that type of logic here, but to mention a key property of it, that is, as far as I know, all the works in that branch of logic are still within the framework of predicate logic and propositional logic. On the other hand, in NAL, this problem does not appear in the first place.

In NAL, *implication*, in its idealized form, is defined to be a reflexive and transitive binary relation from one statement to another. In its realistic form, it is multi-valued, with its truth value defined according to available evidence. Here evidence is measured by comparing the sufficient and necessary conditions of the two statements.

As a term logic using syllogistic rules, in NAL the two premises of an inference step must share a common component, otherwise no conclusion can be derived. Also, the conclusion shares terms with the premises, respectively. As a result, the three must be related to one another in their meanings, according to EGS — two terms are related in their meanings as soon as they appear in the same belief of the system.

Specially, in the induction rule introduced in NAL-5, $(P \Rightarrow Q)$ can be derived from P and Q only if the two premises are based on the same (implicitly represented) evidence. From an arbitrary pair of statements, nothing will be derived — to know their truth values is not enough.

Even when $(P \Rightarrow Q)$ is derived from P and Q, its confidence is low, because the rule is induction. Only when P and Q have been repeatedly supported by the same evidence for many times (and the evidence is different at each time), can $(P \Rightarrow Q)$ then become more confident (by merging the individual conclusions with the revision rule).

In NAL, $(P \Rightarrow Q)$ and $((\neg P) \vee Q)$ are no longer equivalent, but still related to each other. Especially, they have the same negative evidence (that is, when P is true and Q is false). Here the situation is exactly the same as the one revealed by the previous analysis on the "confirmation paradox," that is, since in binary logic a truth value only indicates the existence of negative evidence, statements with exactly the same scope of negative evidence are treated as equivalent. In multi-valued logic with EGS, however, these statements may have different truth values if they have different positive evidence. In NAL, two statements are equivalent if they have the same scope and amount of both positive evidence and negative evidence.

Now we can see that, in this sense, both "confirmation paradox" and "implication paradox" are problems of *binary predicate logics with MTS*, but not problems of *logic* in general. To properly capture the intuitive meaning of concepts like "confirmation," "implication," and so on, we need a multi-valued term logic with an experience-grounded semantics.

9.4.2 Implication and causation

Causal inference is a very important cognitive function, and it has attracted researchers from AI [Pearl, 2000], psychology [Cheng, 1997], and philosophy [Sosa and Tooley, 1993].

What differs NAL from the other approaches in causal inference is: though NAL also attempts to capture all aspects of causal inference, it does not treat "causal inference," as well as the related concepts like "causal relation" and "causation," as logical constants, in the sense that there are inference rules especially responsible for causal inference.

Unlike the Bayesian interpretation of causal inference [Pearl, 2000] (in which causal relation is formalized by conditional probability), in NAL causal inference is carried out by formal inference rules on a formal language. Even in this framework, how to define "causal relation"

is still a controversy. Logically speaking, it has been defined as sufficient condition, necessary condition, sufficient-and-necessary condition, or even something more complicated, by different people [Copi, 1982, Sosa and Tooley, 1993]. In practical applications, there are debates on the "cause" of all kinds of events in every newspapers everyday.

In my opinion, this situation indicates that "causation" is a concept we use to organize our experience, and especially for the prediction of the future [Anderson, 1990]. Since in different domains predictions are made in different ways, the meaning of this concept is context dependent. For this reason, we should not expect a physicist, a biologist, an economist, and a historian to agree on the accurate definition of "cause."

Even when this is the situation, it is still possible to provide a common logical foundation for all the different usage of the concept "causation" (and the related concepts). In NAL, causal inference, or prediction in general, is seen as consisting of two basic aspects, a logical one and a temporal one. The logical factor is represented by the implication relation (and its variant, the equivalence relation), which indicates when a statement can be derived from another one. The temporal factor is represented by the temporal orders introduced in NAL-7. According to the common usage of the term, a "cause" of an event E should be a *pre*condition of it, though it depends on the context whether it is a sufficient one, a necessary one, an equivalent one, or even something more complicated.

In this way, the NAL logical constants provide the "greatest common factor" of all kinds of causal relations, which are treated in NARS as "ordinary relations" (as defined in Section 4.4), with meanings learned from the experience of the system. They can include additional considerations on causation, such as the distinctions between "causation" and "covariation," between "causes" and "enabling conditions," and so on [Cheng, 1997, Sosa and Tooley, 1993], though none of these considerations are included in the logical constants of NAL.

In general, a question with the form of "? $\Rightarrow Q$" is a task looking for an *explanation* of statement Q, and a question with the form of "$P \Rightarrow$?" is a task looking for a *consequence* of statement P. Causal explanation and causal consequence are just special cases of the above general higher-order inference tasks. As described before, both tasks,

as well as yes/no question "$(P \Rightarrow Q)$?," can be answered by a judgment "$P \Rightarrow Q <f, c>$." If there are multiple candidate answers, the choice rule is used to pick the best one. If the question is not about the implication relation in general, but about a special causal relation "*cause*," then the questions will be like "? $\circ\!\!\rightarrow (\perp \; cause \; \diamond \; Q)$," "? $\circ\!\!\rightarrow (\perp \; cause \; P \; \diamond)$," and "$((P \times Q) \circ\!\!\rightarrow cause)$?," respectively, and the answer will depend on the current meaning of "*cause*."

Like the other statements, an implication statement can be derived in more than one way. For example, when a causal judgment "$P \Rightarrow Q <f, c>$" is derived by induction, it may correspond to a causal hypothesis obtained from observed regularity; when it is derived by deduction, it may correspond to a causal hypothesis obtained according to an underlying mechanism; when it is derived by abduction, it may correspond to a causal hypothesis obtained through an explanation. If there is more than one way to support a hypothesis, it will get a higher confidence value after the revision rule merges evidences from different sources, though none of the sources are absolutely necessary for the conclusion to be confident. The same is true for temporal implication/equivalence relations, and the various causal relations obtained in experience. Therefore, in NARS there is no separate rule set for "causal inference" — all inference rules may contribute to the selection and evaluation of various "causes" and "effects."

As an answer to Hume's question on the validity of causal induction [Hume, 1748], in NARS causal inference is a way for the system to organize its experience, rather than a way to find "natural causal laws." The truth value of such a conclusion measures its available evidential support, not its distance to the "objective truth" — even if the system gets such a truth, it cannot confirm the case, given its insufficient knowledge and resources. Under AIKR, all causal beliefs in the system may be revised by future evidence, and none of them fully specifies the causes or consequences of any event. Nevertheless, the causal beliefs still serve a crucial role in the adaptation process of the system.