Journal of Artificial General Intelligence 1 (2009) 25-53 Submitted 2009-2-18; Revised 2009-7-29
DOI: 10.2478/v10229-011-0003-7

Formalization of Evidence: A Comparative Study

Pei Wang Pei.Wang@Temple.Edu
Temple University
Philadelphia, USA

Editor: Ute Schmid

Abstract

This article analyzes and compares several approaches of formalizing the notion of evidence in the context of general-purpose reasoning system. In each of these approaches, the notion of evidence is defined, and the evidence-based degree of belief is represented by a binary value, a number (such as a probability), or two numbers (such as an interval). The binary approaches provide simple ways to represent conclusive evidence, but cannot properly handle inconclusive evidence. The one-number approaches naturally represent inconclusive evidence as a degree of belief, but lack the information needed to revise this degree. It is argued that for systems opening to new evidence, each belief should at least have two numbers attached to indicate its evidential support. A few such approaches are discussed, including the approach used in NARS, which is designed according to the considerations of general-purpose intelligent systems, and provides novel solutions to several traditional problems on evidence.

Keywords: evidence, degree of belief, logic, probability, weight of evidence, revision, ignorance, evidential reasoning, general-purpose system

1. Introduction

> *It is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence.* (Clifford, 1877)

Though the notion of "evidence" is widely used in AI publications, exactly what counts as evidence is an issue that has not been sufficiently discussed, and there are still many open problems (McDermott, 1987). Like many other basic notions, evidence has been formalized in different ways. Most of the formalizations came from the study of mathematics, logic, or philosophy (Achinstein, 1983), each with its assumptions and implications. When they are introduced into AI research, very often people only focus on the technical details, but do not pay enough attention to the theoretical issues involved. Also, there are not many discussions that compare the alternative formalizations of evidence, to show their comparative strength and weakness in AI systems.

This article aims at a systematic analysis and comparison of several representative formalizations of the notion of "evidence" in AI research. Especially, we are going to focus on domain-independent usages of the notion, and ignore the domain-specific usages.[1]

Informally speaking, when *evidence* is mentioned, it is always with respect to some *belief* of a system, for which it provides justification or reason. Since evidence is defined with

[1] For example, in legal discussions, the notion of "evidence" is used with some special conventions.

respect to belief, it is not a special type of knowledge or fact. Whether certain knowledge is "evidence" or not depends on the belief under consideration, rather than only on the property of the knowledge itself.

When designing an AI system, we normally hope the system to establish its beliefs according to the evidence provided by the knowledge or experience of the system. Therefore, it becomes desired to accurately specify the relationship between each belief of the system and the evidence supporting it. The following questions can be asked about this relationship:

- For a given belief, what counts as evidence?
- For a given belief, is there conclusive evidence?
- For a given belief, is there qualitative difference among evidence?
- For a given belief, is there quantitative difference among evidence?
- How much evidence is sufficient for a system to accept or to reject a belief?
- When new evidence comes, how to revise the related beliefs?
- For derived beliefs, how to evaluate their evidential support?

A formalization of evidence will allow the above questions to be answered accurately.

To start, let us set up a general framework in which different formalizations of evidence can be compared. First, we assume the system has a collection of "beliefs" (for the current discussion, they can also be called "hypotheses") that determines the system's responses and behaviors. We further assume there is a *belief language* L_B whose sentences are the beliefs to be evaluated, and an *evidence language* L_E whose sentences or words represent candidate evidence.[2] In some approaches the two languages are the same.

To represent evidential support for beliefs, we assume there is a "degree" associated to each belief, indicating whether, or to what extent, the system accepts the belief. This *degree of belief* should depend on relevant evidence, which is the available information that contributes to the status of belief of the system. Therefore, the degree of belief is the value of a function d that takes a belief B (a sentence in L_B) and its evidence E (sentences or words in L_E) as arguments. Depending on the value range of $d(B, E)$, most of the existing approaches explored in AI can be divided into three groups:

- **Binary-value** — the degree of belief is a binary value, that is, the system either accepts a belief, or rejects it,
- **One-number** — the degree of belief is a number, indicating the extent to which the belief is accepted or held by the system,
- **Two-number** — the degree of belief is a pair of numbers, which can be interpreted as an interval, or two independent measurements.

[2] Generated from a language, either the set of possible beliefs or the set of possible evidence can be infinite, so cannot be exhaustively listed in advance. This framework is needed for general-purpose systems that are always open to new knowledge and problems. Halpern and Pucella (2006) assume the hypothesis set and the evidence set are both finite, and the former is mutually exclusive and exhaustive. Since it assums a predetermined set of problem, this kind of model is usually improper for general-purpose systems.

In the following, we are going to analyze each of them, as well as to compare them, in the context of general-purpose AI (that is, Artificial General Intelligence, or AGI) systems with evidential reasoning capability. The major conclusion to be argued is that for the purpose of AGI, it takes two numbers to properly represent a degree of belief. Therefore, we will not discuss approaches that use more than two numbers for the representation, since they would be consistent with the above conclusion anyway.

Some people may think it is too rigid to use the same uncertainty measurement for all beliefs, and wonder why not "to use as many numbers as needed" to represent the system's state of belief — after all, in fields like statistics, people usually analyze the problem first, then decide how many numbers will be used to represent the uncertainty in it. This methodology is indeed preferred when individual practical problems are analyzed, or problem-specific solutions are designed, *by a human being*. However, for it to work in an AGI system that has to deal with novel problems, it requires, at least, a decision-making procedure to decide how many numbers are needed to represent uncertainty for each problem the system meets, as well as a translation procedure to map one representation into another for cross-problem inference (such as analogy). Since there is no well-established way to formalize the above procedures in a domain-independent manner[3], "to use as many numbers as needed" does not qualify as an alternative to the ones covered in this paper.

For the same reason, the approaches of uncertainty reasoning favored in traditional AI research may not work well in AGI systems, because most traditional AI systems are designed for special problems, and the successes there do not guarantee similar successes in AGI systems, where the problems to be solved are often beyond the restrictions made by the traditional approaches. Consequently, the "AGI context" must be kept in mind to understand the following analysis.

Since this paper focuses on the formal definition and representation of evidence in AGI systems, it will not address the other aspects of evidential reasoning, such as the collection and organization of evidence, or the details of inference rules and inference control mechanism, though they are important aspects of evidential reasoning. To address all these issues is simply impossible for a single journal article.

2. Binary-value Approaches

The most typical case of binary degree of belief can be found in logic-based systems.

In a system based on traditional binary logic, such as Aristotle's Syllogistic or First-Order Predicate Logic (FOPL), the truth-value of a proposition is either *true* or *false*. Since it is rational for the system to only believe true propositions, in this context "evidence" basically means "proof". For given belief B and evidence E, B is acceptable if and only if it can be derived from E, so

$$d(B, E) \equiv (E \vdash B)$$

In this case, the belief language and the evidence language are the same, with their sentences being binary propositions.

[3] Such a formalization, if possible, may require a fixed uncertainty representation itself, since the above procedures are also evidential reasoning processes. Consequently, we will go back to the same problem at the meta-level, where a uniform representation is required.

This kind of evidence is *conclusive*, in the sense that it determines the truth-value of a proposition (and therefore, the system's degree of belief on it) once for all. Consequently, the beliefs of the system increase monotonically with the coming of new evidence, and there is no need to re-evaluate the accepted beliefs, as far as its supporting evidence remains. For the system to be practically useful, its evidence should be *consistent*, that is, it cannot contain, or derive, a proposition together with its negation, otherwise the evidence will support any arbitrary proposition.

Though this approach is simple and elegant, it is not enough for most AI systems, where evidence is usually *inconclusive*, meaning that though the evidence contributes to a degree of belief, it cannot decide the truth-value of the proposition, and therefore the degree of belief may change when new evidence comes. This is usually the case when the type of inference from the evidence to the belief is not *deduction*, but *induction.* (Kyburg, 1983a)

For example, should we believe a general statement "Ravens are black", after a finite number of black ravens are observed? After all, as Hume (1748) pointed out, in this kind of induction, since the statement says more than the past observations, it cannot be proved to be true from the observations alone.

As far as the current discussion is concerned, there are two major approaches attempting to solve Hume's problem:

Incremental-confirmation: Though induction cannot provide conclusive evidence for a general belief, it can incrementally confirm it with inconclusive evidence.

Hypothetico-deduction: Beliefs on general statements are not justified by the existence of *confirming* evidence, but by the lack of *falsifying* evidence.

Though each of the two has its applicable situations, it also has well-known problems.

The incremental-confirmation approach will eventually move beyond binary logic (since it suggests a numerical measurement as degree of belief), though some of its key issues have been discussed qualitatively, within the framework of binary logic. We will address these issues here, and leave the quantitative issues to the next section.

For incremental-confirmation to work, every statement should have explicitly defined *positive* and *negative* evidence (though sometimes they are called by other names).

A well-known definition is "Nicod's Criterion", proposed by French mathematician Jean Nicod. According to it, for "Ravens are black", black ravens are positive evidence, non-black ravens are negative evidence, and non-ravens are irrelevant (Hempel, 1965). Let us be more accurate about this definition. First, it treats a general statement "Ravens are black" as a universally quantified proposition in FOPL, S_1:

$$(\forall x)(Raven(x) \rightarrow Black(x))$$

then every constant in the domain falls into exactly one of three sets with respect to S_1:

$$\begin{array}{rrcl} \text{positive-evidence:} & P_{S_1} & = & \{x \mid Raven(x) \wedge Black(x)\} \\ \text{negative-evidence:} & N_{S_1} & = & \{x \mid Raven(x) \wedge \neg Black(x)\} \\ \text{irrelevant-objects:} & I_{S_1} & = & \{x \mid \neg Raven(x)\} \end{array}$$

Please note that while our belief language is still the one used in FOPL with proposition as beliefs, each piece of evidence is not a proposition anymore, but an object in the domain, which is specified by a proposition.

Though this definition of evidence seems clear and natural, Hempel revealed a paradox by considering a logically equivalent proposition S_2:

$$(\forall x)(\neg Black(x) \rightarrow \neg Raven(x))$$

which can be read as "Whatever is not black is not a raven", and according to Nicod's criterion, for S_2:

$$\begin{array}{rrcl} \text{positive-evidence:} & P_{S_2} & = & \{x \mid \neg Black(x) \wedge \neg Raven(x)\} \\ \text{negative-evidence:} & N_{S_2} & = & \{x \mid \neg Black(x) \wedge Raven(x)\} \\ \text{irrelevant-objects:} & I_{S_2} & = & \{x \mid Black(x)\} \end{array}$$

Compare the two cases, we see that Nicod's criterion gives the two propositions different positive evidence, though the same negative evidence. Since S_1 and S_2 are equivalent propositions (i.e., having the same truth-value), they should have the same evidence. Therefore, Nicod's criterion of evidence fails to specify the same evidence for logically equivalent statements.

If we modify Nicod's criterion by also letting P_{S_2} be positive evidence for S_1, too (and do the symmetric treatment for S_2), then the two equivalent propositions S_1 and S_2 will have the same evidence, that is, positive evidence $P_{S_1} \cup P_{S_2}$ and negative evidence N_{S_1} (which is the same as N_{S_2}).

However, now any red pencil (which is neither black nor a raven, so is in P_{S_2}) becomes confirming evidence for "Ravens are black". This counterintuitive consequence is what Hempel called "Confirmation Paradox" (which is also known as "Hempel's Paradox" and "Raven Paradox").

Since the notion of logical equivalence is central to classical logic, Hempel (1965) felt that "the equivalence condition has to be regarded as a necessary condition for the adequacy of any definition of confirmation". That means to revise Nicod's criterion as above, and to accept any non-black non-raven as confirming evidence for "Ravens are black". After analyzing several alternatives which lead to even worse situations, Hempel concluded that we should rather accept the seemingly counterintuitive result.

There has been a large literature on this paradox, which this article will not attempt to survey. Instead, let us just consider what Hempel's solution means to AI systems. If an AI system were built according to this definition of evidence, then each time it saw a red pencil, a green leaf, or a yellow flower, it would consider "Ravens are black" as having been confirmed one more time. If that still does not sound ridiculous enough, then consider this: for the same reason, the above items are also confirming evidence for "Ravens are white" and even "Ravens are colorless". Furthermore, statement "Dragons are unicorns" would have all existing objects as positive evidence, since they are neither unicorns nor dragons. It is hard to imagine an AI system built according to this solution.

The confirmation paradox places the believers of incremental-confirmation between a rock and a hard place, since they have to either give up equivalence condition and violate propositional logic, or accept a highly counterintuitive and practically inapplicable definition of evidence.

The confirmation paradox does not exist if we take general beliefs as accepted by hypothetico-deduction, as suggested by Popper (1959). Using the previous terminology, this method says that we accept “Ravens are black” as far as no *negative* evidence has been encountered, and whether there is *positive* evidence does not matter — this method does not even define “positive evidence”, and in it “evidence” means “negative evidence”.

According to Popper, there is an asymmetry between verifiability (by positive evidence) and falsifiability (by negative evidence), which results from the logical form of universal statements, or “theories” in his words, that is, “a positive decision can only temporarily support the theory, for subsequent negative decisions may always overthrow it”. He further said “I never assume that by force of ‘verified’ conclusions, theories can be established as ‘true’, or even as merely ‘probable’ ” (Popper, 1959).

Compared to Nicod’s criterion and Hempel’s suggestion, Popper’s solution to the problem of evidence is more compatible with FOPL. If “Ravens are black” is represented as universal proposition

$$(\forall x)(Raven(x) \rightarrow Black(x))$$

then an observation corresponds to a constant c that instantiates the variable, and produces a particular proposition

$$Raven(c) \rightarrow Black(c)$$

If the particular proposition is false, c also makes the universal proposition false; but if c makes the particular proposition true, it tells us little about the truth-value of the universal proposition, which can still be either true or false. This is the case because a universal proposition is defined as the *conjunction* of corresponding particular proposition on every constant in the domain.

Confirmation paradox does not exist in this situation, because the S_1 and S_2 defined previously do have the same negative evidence, and positive evidence does not count, so equivalent propositions still have the same evidence, as desired. Therefore, observing a red pencil has nothing to do with our belief on “Ravens are black”, which feels right.

However, this approach also claims that observing a black raven has nothing to do with our belief on “Ravens are black”, which is counterintuitive. Assume that “Ravens are black” and “Dragons are red” both have no observed counterexamples, and we have observed many black ravens but no dragon of any color, then should the two statements be believed to the same extent? Furthermore, almost all conclusions in empirical science and everyday life have known exceptions, but the conclusions are rarely falsified, as long as they still cover much more situations, that is, have sufficient positive evidence.

Now we can see that in the hypothetico-deduction approach, the notion of *evidence* actually means *conclusive evidence.* When a general statement is represented as a universally quantified proposition, a constant can only prove it false, but can never prove it true. In this way, this approach is consistent with FOPL, but it still cannot capture inconclusive evidence, either positive or negative.

One well-known result showing people’s affinity for confirming evidence is Wason’s selection task (Wason and Johnson-Laird, 1972), a psychological experiment that has been repeated many times by different researchers. Its result shows that when people are asked to check the truthfulness of a general statement, they more often seek positive evidence than negative evidence, though according to logic only the latter is relevant. For example, when subjects are given four cards showing symbols E, K, 4, and 7, respectively, and are asked to determine whether “If a card has a vowel on one side, then it has an even number on the other side", most subjects turn the E card alone, or E and 4, while the "logical" answer is E and 7. This result is usually interpreted as a human fallacy, but it can also be argued that the human behavior can be justified, and the problem is actually in the "logic" that fails to include the natural concept of (inconclusive) positive evidence (Wang, 2001c).

To summarize the above discussion, we have seen that in classical logic, the concept of *conclusive evidence* is well-defined by deduction, but the concept of *inconclusive evidence* is hard to introduce. It should not be a surprise if we consider where the logic come from. Logic study has been dominated by deductive logic for two millennia, and by mathematical logic for a century. In those logics, inconclusive evidence plays little role — no matter how many times the Goldbach Conjecture has be verified on various numbers, it remains a "conjecture", not a "theorem", even though these verifications make people's belief on it to become stronger and stronger.

Therefore, to build AI systems in which inconclusive evidence plays an important role, it is necessary to look beyond classical logic.

One attempt to extend classical logic, within the framework of binary logic, is *nonmonotonic logic* (Reiter, 1987). In this kind of logic, a "default rule", such as "A bird normally flies", can be used to produce tentative conclusions, like "Tweety flies", from the default rule and available facts, like "Tweety is a bird". Later, when new information disqualifies the applicability of the default rule (for example, by revealing that Tweety is not a normal bird), the status of the previous conclusion is changed. In this way, default rules, which represent *normal* or *general* situations, can coexist with known counterexamples, as their exceptions. This is clearly closer to the reality of human reasoning.

However, in nonmonotonic logics the default rules are given to the system by its designer (or user), not induced from observations by the system itself, and nor are they verified by evidence. Consequently, here the induction problem and the confirmation problem are *avoided*, rather than *solved*. In these systems, new evidence only revises the degree of belief of the tentative conclusions, not that of the default rules. For instance, no matter how many birds observed cannot fly, the belief "Birds normally fly" remains valid. If such a system attempts to generate its own default rules, or to attach numerical degree of belief to them, the same problems will appear, as in classical logic. This type of logic is not powerful enough for an AGI system where *all* beliefs should be based on evidence.

Another related non-classical logic is *conditional logic* (Dubois and Prade, 1994; Milne, 1997). Though conditional statement "If P, then Q" (where P and Q are propositions) is traditionally formalized as material implication "$P \rightarrow Q$", some scholars, such as de Finetti, find reasons to represent the statement as a three-valued *conditional object*, "$Q|P$", which has the same truth-value as Q when P is true, while has a truth-value *void* (i.e., *undefined*) when P is false.

This idea is relevant to the current discussion, because the *evidence* for a belief can be conceptually considered as the *condition* of the belief, and some problems can be solved in this way. For example, if "Ravens are black" is rephrased as "If something is a raven, then it is black", and formalized as "$Black(x)|Raven(x)$", then the above three-valued truth directly corresponds to Nicod's criterion for evidence, that is, black ravens are positive evidence, non-black ravens are negative evidence, and non-ravens are irrelevant. Hempel's paradox does not appear here, because "$Black(x)|Raven(x)$" and "$\neg Raven(x)|\neg Black(x)$" are not equivalent in this three-valued logic.

Even so, it does not mean that conditional logic will satisfy the need for evidential reasoning in AGI, for two major reasons:

- In its three-valued form, conditional logic still only represents conclusive evidence, though it corresponds to an intuitively reasonable definition for inconclusive evidence (Nicod's Criterion). For an AGI system, three truth-values are not enough to distinguish the status of belief caused by various inconclusive evidence.
- Though "condition" is related to "evidence", they are not the same. The content of a condition is always *explicitly expressed* in a conditional belief, while the evidence of a belief is usually *implicitly summarized* in the degree of belief. Consequently, the two are processed in different ways in a reasoning system.

Both topics will be discussed in the next section with more details.

In summary, though nonmonotonic logics and conditional logics have their applicable situations, they are not suitable for the evidential reasoning problem discussed in this paper, because in this context inconclusive evidence usually need to be *quantitatively* represented, which is not what a binary logic can naturally do.

3. One-number Bayesian Approach

From the previous discussion, we see that for many practical problems, both positive and negative evidence should be taken into consideration when a degree of belief is determined, and it is often necessary to quantitatively compare them. This observation makes many people to believe that the proper framework to be used here is not a binary logic, but probability theory.[4]

According to certain interpretation, "probability" measures the logical relation between a hypothesis and the available evidence, and "conditional probability" exactly measures the evidential support a hypothesis gets, with the available evidence as the condition (Carnap, 1950; Rescher, 1958; Kyburg, 1994). Therefore, it seems enough to use a single value, (conditional) probability, to indicate the status of a belief. In AI, the most influential example of this opinion is the Bayesian approach proposed by Pearl (1988), which is characterized by the following commitments: (Pearl, 1990)

- willingness to accept subjective belief as an expedient substitute for raw data,
- reliance on complete (i.e., coherent) probabilistic models of beliefs,
- adherence to Bayes' conditionalization as the primary mechanism for updating belief in light of new information.

Like many other phrases, "Bayesian approach" may mean different things to different people. To avoid confusion, in the following we will use "one-number Bayesian" to indicate the above treatment of evidential reasoning in an AGI system. According to it, the system's

[4] Though fuzzy logic also uses a numerical truth-value, it is usually not based on evidence. See Zadeh (1975); Wang (1996b).

degree of belief, with given evidence, is fully represented by a conditional probability with the evidence as condition, that is,

$$d(B, E) \equiv P(B|E)$$

Consequently, the processing of evidence follows probability theory, especially Bayes' theorem.

Such a definition naturally covers both positive and negative evidence, and their difference is whether the evidence increases the probability or decreases it. That is, for belief B,

$$\begin{array}{rrcl} \text{positive-evidence:} & P_B & = & \{x \mid P(B|x) > P(B)\} \\ \text{negative-evidence:} & N_B & = & \{x \mid P(B|x) < P(B)\} \\ \text{irrelevant-information:} & I_B & = & \{x \mid P(B|x) = P(B)\} \end{array}$$

Defined in this way, both the belief language and the evidence language are events or propositions on which the probability distribution function P is defined.

Some problems in the binary approaches can be solved using probability theory. For example, Oaksford and Chater (1994) re-interpret the result of Wason's selection task according to probability theory, and consequently, "we can view behavior in the selection task as optimizing the expected amount of information gained by turning each card".

The situation is similar for the confirmation paradox. First, there are different ways to formalize the statement "Ravens are black" in the one-number Bayesian framework. One way is to simply attach a probability value to a universally quantified proposition, so the degree of belief is

$$P((\forall x)(Raven(x) \rightarrow Black(x)))$$

and another way is to interpret the statement as the previously mentioned "conditional object" (Dubois and Prade, 1994; Milne, 1997), so the degree of belief is

$$P(Black(x)|Raven(x))$$

Under the former interpretation, Fitelson and Hawthorne (2009) shows that a non-black non-raven is positive evidence for "Ravens are black", though it is "weak evidence", that is, its degree of confirmation is much lower than that of a black raven. For instance, under certain assumptions, "100 instances of black ravens would yield a likelihood ratio 169 times higher than would 100 instances of non-black non-ravens."

Under the latter interpretation, $P(Black(x)|Raven(x))$ and $P(\neg Raven(x)|\neg Black(x))$ are usually different, since the first probability is for ravens to be black, and the second for non-black things not to be raven. Consequently, a red pencil is positive evidence for the second, but has nothing to do with the first. As in conditional logic, this treatment of evidence returns to Nicod's criterion.

Either way, the Bayesian solution to the confirmation paradox seems less counterintuitive than Hempel's. It should be noticed that this result is achieved by dropping or weakening the equivalence condition, that is, equivalent propositions in predicate logic may have different (though related) probabilities when they are taken to be probabilistic. The Bayesian solution usually avoids the Raven's paradox by building problem-specific models, where whether, or how much, a red pencil contributes to the system's belief on "Ravens are black" depends on the assumptions made in the model, and the results can be justified in that way, too.

Even so, in the AGI context there are still problems left. First, as mentioned before, an AGI system cannot assume that there is already a built-in probabilistic model for every problem, and to automatically build models for various kinds of problems is not yet a feasible procedure. Furthermore, in AGI systems it is practically impossible to treat a red pencil as confirming (though weak) evidence for "Ravens are black", since the system simply cannot afford the resources to do so (there are too many non-black non-ravens to be considered for the system to scale up). On the other hand, it leads to theoretical inconsistency if this type of update is acknowledge as necessary, but not implemented.

There are other criticisms to the one-number Bayesian approach. Some people do not think it is necessary for AI systems to follow probability theory. After all, psychological study shows that the everyday human reasoning systematically violates probability theory (Tversky and Kahneman, 1974). For example, people tend to use *representativeness* as probability. One consequence is the "conjunction fallacy" — after learning certain properties of a certain person, people often judge her more likely to be (a) "a bank teller and active in the feminist movement" than (b) "a bank teller". However, since (a) is a subset of (b), according to probability theory the person should be more likely to be in (b) than in (a) (Tversky and Kahneman, 1983).

As the works of Hempel and Wason show that human reasoning does not follow FOPL, the works of Tversky and Kahneman show that it does not follow classical probability theory, neither. These results are usually interpreted as "fallacies and biases" caused by the non-optimality nature of the human mind. According to this opinion, probability theory, like FOPL, is still a proper *normative* theory of reasoning (which specifies the rules that *should be followed*), though not a proper *descriptive* theory of the process in the human mind (which specifies the rules that *are followed*).

Even when probability theory is evaluated as a normative theory, there is still no lack of controversy. First, the availability of a prior probability distribution may be problematic (Kyburg, 1983a). A traditional reason for AI researchers to refuse numeric approaches of reasoning in general, and probabilistic approaches in specific, is that we do not have the numbers to start with (McCarthy and Hayes, 1969). Even if for each individual belief we can evaluate its degree of belief in isolation, there is no guaranty that when these values are putting together they form a *coherent* probability distribution (Walley, 1996b). Actually the situation is often the opposite, and that is where the *reference class* problem comes from: according to different considerations, we often get different probability evaluations for the same belief, and probability theory, except in some special cases, does not tell us what to do in this situation (Kyburg, 1983b; Wang, 1995b).

How about to use Solomonoff's universal priori distribution (Solomonoff, 1964; Hutter, 2005)? To handle beliefs in this way raises several complicated issues beyond the scope of this paper. For the current discussion, it is enough to say that for an AI system working in practical situation, this approach has not provided a computable procedure to assign a prior probability value to a belief like "Tweety can fly".

Some people do not take the lack of prior knowledge as a big problem, because they believe we can start with a "non-informative prior", then use Bayesian conditionalization to learn from new evidence whenever it becomes available. Though putting the stress on learning is justifiable, depending on Bayesian conditionalization to do so has serious limitations. As analyzed in detail in Wang (1993, 2004), when the degree of belief is represented by a single probability value, Bayes theorem and Jeffrey's rule cannot be used to learn all kinds of knowledge that can be put into a prior probability. Since this topic is central to the current discussion and the previous analysis has not got enough attention, it is necessary for the argument to be rephrased in the following.

Assume $P_K(x)$ is a probability distribution function established according to background knowledge K on proposition space S, that is, $P_K(x)$ is defined if and only if $x \in S$, and its value is determined by the knowledge in K, which can be intuitively considered as a set of evidence.

In this context, Bayes theorem is often used for *conditionalization*, that is, to accept new event E into the background knowledge K when the event happens, so as to turn the prior (conditional) distribution based on K and conditioned on E into a posterior (unconditional) distribution based on K plus E, as

$$P_{K \cup \{E\}}(x) = P_K(x|E) = P_K(E|x)P_K(x)/P_K(E)$$

However, this usage requires $E \in S$ and $P_K(E) > 0$. If the new evidence E needs to be handled by revising K, then it cannot be treated as conditionalization, because $P_K(x|E)$ is still based on K. Since the background knowledge K is not necessarily included in the domain of the probability distribution S, it cannot be written as a condition. Therefore, $P_K(x)$ should not be written as $P(x|K)$.

In the relevant discussions, the background knowledge K is often mistaken as the condition E. Here are some typical examples:

- Pearl (1988, page 29): "$P(A|K)$ stands for a person's subjective belief in A given a body of knowledge K, ... In defining belief expressions, we often simply write $P(A)$ or $P(\neg A)$, leaving out the symbol K. This abbreviation is justified when K remains constant, since the main purpose of the quantifier P is to *summarize* K without explicating it."

- Cheeseman (1988, page 60): Bayes Theorem is expressed as

  $$P(H|E,c) = \frac{P(E,H|c)}{P(E|c)} = \frac{P(H|c)P(E|H,c)}{P(E|c)}$$

  where H is a hypothesis, E is the evidence, and c is the context. This target paper was published together with 23 commentaries by well-known researchers in the field, but none of the commenters challenged the above expression by pointing out that the context c cannot be represented as a condition, since $P(c)$ may be undefined.

Here the issue is not what terminology or notation to use, and it is usually fine to write $P_K(H|E)$ as $P(H|E;K)$, as some authors do. What matters is to understand that in either form, this conditional probability evaluation of hypothesis H depends on *two* types of "condition", E and K, and the system's degree of belief on H is actually $d(H,(K \cup \{E\}))$, rather than merely $d(H,E)$. These two types of condition have different status in the one-number Bayesian approach:

1. E is often referred to as "condition" or "observation", and is called "explicit condition" in Wang (1993, 2004). It is a proposition in the space S on which the probability distribution function P_K is defined, and usually cannot be omitted in the expression, since $P_K(H|E)$ and $P_K(H)$ are normally different. The system has no problem to do inference with different Es for the same H.

2. K is often referred to as "background knowledge" or "context", and is called "implicit condition" in Wang (1993, 2004). Its elements are not necessarily in S, so their probabilities may be undefined. It is often omitted in the expression, so $P(H|E)$ normally means $P_K(H|E)$ for certain constant K. However, even when K is omitted in the expression, it does not mean the probability function is "context-free", unless P is "objective probability" (which is rarely available to AI systems). Except in special cases, inference across probability distributions based on different Ks is not allowed in probability theory, because they correspond to different probability distribution functions, which are not necessarily consistent with each other. Inconsistent probability functions cannot be used together, since doing so violates the axioms of probability theory, which require each event (or proposition) to have a unique (prior) probability value.

This distinction between "explicit condition" and "implicit condition" is basically the same as the distinction between "condition" and "evidence" mentioned in the previous discussion on conditional logic. In $P_K(H|E)$, roughly speaking both E and K provide *evidence* for the system's belief on H. However, E is explicitly expressed, while K, after making its contribution to the degree of belief, is summarized by the probability value, and often implicitly represented. Even when we explicitly mark it as K, we can only process it in very limited ways. For example, we can say that P_{K_1} and P_{K_2} are usually different when K_1 and K_2 are different. However, in neither probability theory nor conditional logic can K be further analyzed or processed, because its content is no longer fully specified.

Let's see a concrete example. Assume we are interested in predicting the result of a one-time experiment, which has three possibilities R_1, R_2, and R_3. A probability distribution function P is used to represent our degree of belief on each possibility, so the space of events S includes R_1, R_2, and R_3 as propositions, as well as the propositions formed by them using Boolean operators $\wedge$ (*and*), $\vee$ (*or*), and $\neg$ (*not*). Initially, the background knowledge, referred to as K_1, supports each possibility to the same extent, so our belief status about the experiment result is

$$P_{K_1}(R_1) \;=\; P_{K_1}(R_2) \;=\; P_{K_1}(R_3) \;=\; 1/3$$

This belief status can be revised in two different ways when new information comes:

1. The new information shows that R_1 will not happen. In this case, Bayesian conditionalization gives us a new belief status

$$\begin{array}{lll} P_{K_2}(R_1) & = P_{K_1}(R_1|\neg R_1) & = 0 \\ P_{K_2}(R_2) & = P_{K_1}(R_2|\neg R_1) & = 1/2 \\ P_{K_2}(R_3) & = P_{K_1}(R_3|\neg R_1) & = 1/2 \end{array}$$

   here the new background knowledge K_2 include K_1 and the new information.

2. The new information doubles the evidential support for R_1, while says nothing directly about the other two. According to the current background knowledge K_3, the new belief status is
$$P_{K_3}(R_1) = 1/2 \; ; \; P_{K_3}(R_2) = P_{K_3}(R_3) = 1/4$$
The revision is not carried out by Bayesian conditionalization, since the new information cannot be represented as a "condition".

In this example, the probability of K_1, K_2, or K_3 is undefined. Some people think since they are accepted assumptions, they have probability value 1, which is a misconception, since they are not even in S.[5]

When putting in this way, we assume all researchers will agree that the two types of evidence are very different, and the one-number Bayesian approach can only handle the "learning" or "pooling" of E (such as to get $P_K(H|E_1 \wedge E_2)$ from the information provided by P_K), but not that of K (such as to get $P_{K_1 \cup K_2}(H|E)$ from the information provided by P_{K_1} and P_{K_2}), except in certain special situations. However, the previous examples show that inaccurate use of notions does happen in many places, and this conceptual confusion between different types of evidence is often unnoticed. Many people seem to think that "as far as the background knowledge is not actually *processed* as condition, there is no harm for it to be *represented* as condition". Though such a treatment does not cause calculation mistakes, this practice is responsible for the misconception that there is no fundamental difference between background knowledge and condition, so that the missing information in background knowledge can be learned later by conditionalization. Consequently, people get the impression that all types of evidence can be handled by the one-number Bayesian approach, though nobody has explicitly argued that "background knowledge" can indeed be treated as "condition" in conditional probability.

In summary, when used properly, the one-number Bayesian approach does provide a framework for the representation and processing of inconclusive evidence. Using a numerical measurement of evidential support, it works better than a binary logic in many problems. However, within this framework certain properties of evidence cannot be captured. It is not enough to use a single probability distribution for representation (since it cannot provide the information on how much the distribution should be modified by a piece of new evidence), and Bayesian conditionalization for learning (since it cannot revise background knowledge). As soon as there is a need to generally revise the background knowledge (i.e., the evidence on which the prior probability distribution is based), the one-number Bayesian approach will not work. Since an AGI system needs to handle evidence coming from different sources, it cannot assume that all evidence can be treated with respect to a chunk of background knowledge that remains constant all the time.

4. Dempster-Shafer Approach

The conclusion of the previous section can be put in a different way, that is, even if a probability value can be assigned to a belief according to the available evidence, the value

[5] S cannot be extended to include possible background knowledge, since K_1, K_2, and K_3 are not in the same conceptual space as R_1, R_2, and R_3, so the two groups should not be covered by the same probability distribution function.

does not show the system's ignorance or uncertainty about the probability evaluation itself, which is needed for its revision. This opinion is hardly new, and it has been a major motivation for new uncertain reasoning theories. One example is the Dempster-Shafer theory of evidence, or D-S theory (Dempster, 1967; Shafer, 1976).

D-S theory differs from other uncertainty management approaches in two major points:

1. In the representation of uncertainty, the theory starts at a *basic probability assignment*, $m(x)$, defined on the space of the *subsets* of competing hypotheses. Then $m(x)$ defines the *degree of belief*, $Bel(A)$, and the *degree of plausibility*, $Pl(A)$, of a set of hypotheses A. Intuitively, the $[Bel(A),\ Pl(A)]$ interval is a generalization of probability function $Pr(A)$, and the width of the interval indicates the ignorance of the system on A.

2. In the processing of uncertainty, *Dempster's rule of combination* is applied to calculate $m_1 \oplus m_2(x)$ from $m_1(x)$ and $m_2(x)$, where $m_1(x)$ and $m_2(x)$ are based on evidence from distinct sources, and $m_1 \oplus m_2(x)$ is based on the pooled evidence. The resulting interval $[Bel(x),\ Pl(x)]$ gets narrower, and when the rule is repeatedly applied, the interval eventually converges to a point, which is a probability value.

Formally, a *frame of discernment* Θ is an exhaustive and exclusive set of hypothesis. On it, the basic probability assignment $m : 2^{\Theta} \rightarrow [0, 1]$ is constrained by

$$m(\emptyset) = 0, \ \sum \{m(A) \,|\, A \subseteq \Theta\} = 1.$$

When A is a subset of Θ, its *degree of belief* $Bel(A)$ and *degree of plausibility* $Pl(A)$ are defined as the following

$$Bel(A) = \sum \{m(B) \,|\, B \subseteq A\} \ , \ Pl(A) = \sum \{m(B) \,|\, B \cap A \neq \emptyset\}$$

For a hypothesis $H \in \Theta$, $Bel(\{H\}) \leq Pl(\{H\})$, and $Pl(\{H\}) = 1 - Bel(\{\neg H\})$. When $Bel(\{H\}) = Pl(\{H\})$, both of them should be the same as $Pr(H)$, the probability, or chance, of H. Therefore, D-S theory attaches two numbers to each belief to measure its uncertainty, and the $[Bel(\{H\}),\ Pl(\{H\})]$ interval provides a more general measurement than $Pr(H)$, by allowing some ignorance, as measured by the width of the interval.

Dempster's rule specifies how to combine two basic probability assignments:

$$\begin{array}{rcl} m_1 \oplus m_2(\emptyset) & = & 0 \\ m_1 \oplus m_2(A) & = & \lambda \sum \{m_1(B) m_2(C) \,|\, B \cap C = A \neq \emptyset\} \\ \lambda & = & [1 - \sum \{m_1(B) m_2(C) \,|\, B \cap C = \emptyset\}]^{-1} \end{array}$$

This rule is used to combine the evidential support from distinct sources.

Shafer (1976) also introduces a *weight of evidence* with the following properties (Shafer, 1976, pages 7, 88):

1. Weight of evidence w is a measurement defined on bodies of evidence, and it takes values in $[0,\ \infty]$.

2. When two entirely distinct bodies of evidence are combined, the weight of the pooled evidence is the *sum* of the original ones.

Now we can summarize the relevant statements in Shafer (1976) into the following four postulates:

Postulate 1: Chance, or probability, is the limit of the proportion of positive outcomes among all outcomes (pages 9, 202).

Postulate 2: Chances, if known, should be used as degrees of belief (pages 16, 201).

Postulate 3: Evidence combination corresponds to the addition of weights of evidence (pages 8, 77).

Postulate 4: Dempster's rule should be used for evidence combination (pages 6, 57).

Though each of the above postulates sounds reasonable individually, Wang (1994a) has proved that they are inconsistent. In the following the proof is summarized.

For our current purpose, it is enough to study the simplest non-trivial frame of discernment, where $|\Theta| = 2$. Let $\Theta = \{H, H'\}$. Since Θ is exhaustive and exclusive, $H' = \neg H$. Also, we consider a simple type of evidence combination: *enumerative induction*. In this situation there are only two types of evidence: positive evidence (that supporting H) and negative evidence (that supporting $\neg H$).

For each piece of evidence, the basic probability assignment is:

$$\begin{array}{rl} \text{positive evidence:} & m(\{H\}) = s\,, \quad m(\{\neg H\}) = 0\,, \quad m(\Theta) = 1 - s \\ \text{negative evidence:} & m(\{H\}) = 0\,, \quad m(\{\neg H\}) = s\,, \quad m(\Theta) = 1 - s \end{array}$$

where s is a constant in (0, 1), indicating the extent of the support by a single piece of evidence. Later we will see that the value of s does not matter for this discussion.

The relationship between s and the weight of a single piece of evidence, w_0, can be derived from Dempster's rule (Postulate 4) and the additivity of the weight in evidence combination (Postulate 3). Shafer (1976, page 78) gives the following results:

$$s = 1 - e^{-w_0}\ ,\ w_0 = -log(1 - s)$$

Let us use t^+ and t^- for the number of pieces of positive and negative evidence, respectively, and assume the pieces are all distinct, so no evidence is repeatedly counted. Applying the above relation to this situation, Postulate 3 gives the accumulated weight of positive and negative evidence:

$$w^+ = w_0 t^+\ ,\ w^- = w_0 t^-$$

The corresponding basic probability assignments can be decided by the reverse relation, and then using Dempster's rule the belief function can be derived (Shafer, 1976, page 84):

$$Bel(\{H\}) = \frac{e^{w^+} - 1}{e^{w^+} + e^{w^-} - 1}$$

Now when the number of pieces of evidence, $t = t^+ + t^-$, goes to infinite, so does the total weight of evidence $w = w^+ + w^-$. According to Postulate 1, the probability of H is

$$q = \lim_{t\to\infty} \frac{t^+}{t}$$

At the same time, the belief function also converges (Shafer, 1976, page 198):

$$Bel_{\infty}(\{H\}) = \lim_{w \rightarrow \infty} Bel(\{H\})$$

But for the current example we find (Wang, 1994a):

$$Bel_{\infty}(\{H\}) \quad = \quad \begin{cases} 0 & \text{if } q < 0.5 \\ 0.5 & \text{if } q = 0.5 \\ 1 & \text{if } q > 0.5 \end{cases}$$

This result contradicts Postulate 2, which requires $Bel_{\infty}(\{H\}) = q$.

Therefore if q (the chance of H) exists, then by repeatedly applying Dempster's rule to combine the coming evidence, $Bel(\{H\})$ (and $Pl(\{H\})$) will converge to a point. However, that point is not q in most cases, but 0, 0.5, or 1, indicating qualitatively whether there is more positive evidence than negative evidence.

Possible solutions of this inconsistency are discussed in detail in Wang (1994a). One of them is to give belief function and Dempster's rule a new interpretation, and do not link them to probability or chance (Smets, 1991; Smets and Kennes, 1994; Baroni and Vicig, 2001). Such a semantic change resolves the inconsistency (though it was not proposed initially for this purpose), but it achieves that at the price of giving up a major objective of the theory, that is, to extend probability theory by representing ignorance as part of the uncertainty to be processed.

As far as the current discussion goes, the important issue is not how to save D-S theory from the inconsistency, but how to represent and process evidence in AGI systems. On one hand, we see that the one-number Bayesian approach is not enough here, because the amount of evidence cannot be decided from a probability distribution function. On the other hand, we still want to take a probability distribution as a special case, when the ignorance about it can be ignored.

From the above discussion, we show that if Dempster's rule is used to combine evidence, the belief function does not converge to the chance of the belief (if it exists), and in general, the belief function is not directly related to the most common measurement of uncertainty, the proposition of positive evidence among all evidence. If these properties are desired, then Dempster's rule has to be given up.

5. Two-number Bayesian Approaches

D-S theory is not the first attempt to use *two numbers* to represent a probability evaluation and the amount of its supporting evidence. Similar opinions were proposed much earlier, by researchers following different paths of thought:

> *In short, to express the proper state of our belief, not* one *number but* two *are requisite, the first depending on the inferred probability, the second on the amount of knowledge on which that probability is based.* (Peirce, 1878, page 160)

> *As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, according as the new*

> *knowledge strengthens the unfavorable or the favorable evidence; but* something *seems to have increased in either case — we have a more substantial basis upon which to rest our conclusion.* (Keynes, 1921, page 71)

According to these opinions, this *amount* of evidence (or knowledge) provides information that is not in the probability values. Intuitively, this measurement should correspond to the weight of evidence in D-S theory, or the sample size in statistics, and it should be additive when different pieces of evidence are pooled together. A larger amount of evidence should correspond to a more stable belief during revision.

There have been attempts to derive such a measurement from a probability distribution function. For example, Good (1950, 1985) defined a "weight of evidence" as the logarithm of a "Bayes factor", a function of probability. More recently, Halpern and Pucella (2006) introduced a "weight of evidence", which "is essentially a normalized likelihood". This kind of measurement, though useful for other purposes, cannot solve our current issue, because according to the previous discussion, the probability distribution function $P_K(x)$ does not contain all the information about the background knowledge K. This K is able to *derives* the probability distribution, rather than be *derived from* it, since the same probability distribution (such as "The probability of getting a head from tossing this coin is 0.5.") may come from very different background knowledge ("This coin is known to be fair" vs. "The fairness of this coin is unknown").

Another approach in the Bayesian tradition is Walley's theory of "imprecise probabilities" (Walley, 1991, 1996b). The intuition behind Walley's lower and upper probabilities of an event is similar to Dempster's original idea, as well as Shafer's belief function and plausibility function, but Walley defines them as the minimum and maximum betting rate, respectively, that a rational person is willing to pay for a gamble on the event.

To compare this approach with D-S theory, we can also apply it to enumerative induction. To relate probability to evidence, Walley assumes a situation where the chance for an event to happen has a *near-ignorance* prior distribution, and the observations of the event are independent of one another. If among t observations the event happens t^+ times, then according to Bayes' Rule, the *lower* and *upper* probabilities of the event are

$$l = t^+/(t + s_0)\ ,\ u = (t^+ + s_0)/(t + s_0)$$

respectively, where s_0 is a parameter of the prior distribution, indicating the convergence speed of the lower and upper probabilities (Walley, 1991, page 218).

An *evidence combination rule* can be derived from the additivity of evidence and the above relation between evidence and lower/upper probability. If the support of two distinct pieces of evidence to the same belief is measured by two pairs of lower/upper probabilities, $[l_1, u_1]$ and $[l_2, u_2]$, respectively, then the equivalent amounts of evidence are:

$$t_1^+ = s_0 \frac{l_1}{u_1 - l_1},\ t_1 = s_0 \frac{1 - (u_1 - l_1)}{u_1 - l_1}$$

$$t_2^+ = s_0 \frac{l_2}{u_2 - l_2},\ t_2 = s_0 \frac{1 - (u_2 - l_2)}{u_2 - l_2}$$

The *ignorance* (or *imprecision*) of a belief is defined as the difference between its lower and upper probabilities, that is, $i = u - l = s_0/(t + s_0)$, which decreases as t increases. Using it, the above relations are simplified into:

$$t_1^+ = s_0\frac{l_1}{i_1},\ t_1 = s_0\frac{1-i_1}{i_1}$$

$$t_2^+ = s_0\frac{l_2}{i_2},\ t_2 = s_0\frac{1-i_2}{i_2}$$

When the two pieces of evidence are combined, for the result we have

$$t^+ = t_1^+ + t_2^+,\ t = t_1 + t_2$$

Now the probability interval of the conclusion, $[l,\ u]$, can be calculated from the probability intervals of the premises, $[l_1,\ u_1]$ and $[l_2,\ u_2]$, according to the relation between $[l,\ u]$ and $\{t^+,\ t\}$, plus the above six equations. Assuming all ignorance values are non-zero, we get the following functions

$$l = \frac{l_1 i_2 + l_2 i_1}{i_1 + i_2 - i_1 i_2},\ u = \frac{l_1 i_2 + l_2 i_1 + i_1 i_2}{i_1 + i_2 - i_1 i_2},\ i = \frac{i_1 i_2}{i_1 + i_2 - i_1 i_2}$$

which are independent of the choice of s_0. This combination rule is not in any of Walley's writings (as far as we know), though can be directly derived from the relationship between belief and evidence in his theory.

In this simple situation, the above rule does what Dempster's rule is supposed to do, that is, to combine evidence from different sources. Furthermore, when the chance of the event does exist, the two probabilities converge to it, that is,

$$\lim_{t\to\infty}\frac{t^+}{t+s_0} = \lim_{t\to\infty}\frac{t^+}{t} = \lim_{t\to\infty}\frac{t^+ + s_0}{t+s_0}$$

In summary, as far as the current discussion is concerned, both D-S theory and Walley's theory can be seen as attempts to extend the one-number Bayesian approach, by using an interval to represent the degree of belief:

$$d(B,E) \equiv [l,u]$$

Intuitively speaking, the interval as a whole serves the role of a probability value in the one-number Bayesian approach, while the width of the interval measures another type of uncertainty that cannot be properly represented in the one-number approach. When the system gets more and more evidence, the interval gets narrower and narrower, and it eventually converges to a point. Their difference is that in Walley's theory, the point is the probability of the belief, while in D-S theory it is not (despite of the claim that it is).

In Walley's theory, the $[l,u]$ interval and the amounts of (confirming and total) observations $\{t^+,t\}$ mutually determine each other, under the assumption on the prior distribution and its parameters. Therefore if the evidence of the system directly comes from countable observations (as in enumerative induction), the degree of belief can also be represented by the this pair of number:

$$d(B,E) \equiv \{t^+,t\}$$

Similar situations exist in some other approaches of the Bayesian tradition, where the system's belief on a statement is represented not by a probability *value*, but by a probability *distribution* of a certain type (such as beta distribution) and with certain parameters, then with the coming of new observations, Bayes Theorem is used to revise the parameters, so as to change the system's belief (DeGroot, 1970). These approaches are different from the one-number Bayesian approach discussed previously, because when implemented in a computer system, they will need to attach two numbers to each statement to represent how much the system believes it, either in the form of the parameters of the probability distribution (e.g., α and β for a beta distribution), or the number of observations (i.e., $\{t^+, t\}$) from which the parameters can be determined.

This result is in agreement with the Peirce-Keynes thesis that two numbers are needed to represent the relation between evidence and belief, though we have seen that there may be different ways to define such numbers. What matters here is that the "degree of belief" should actually have *two* degrees of freedom in it, because using a single number, it is impossible to distinguish the two aspects in evidential support: its *direction* (i.e., positive vs. negative) and its *stability* (i.e., strong vs. weak). More will be said about these two factors in the following.

6. Evidence in NARS

In this section, one more two-number approach is introduced. It is part of an AGI project, NARS (Non-Axiomatic Reasoning System). The project has been described in many other places, including Wang (2006) and the publications at the project website.[6] This article makes no attempt to introduce NARS as a whole, but uses it as another example to support the conclusion that two numbers are needed for degree of belief used in AGI systems. Therefore, here we only describe the definitions of evidence and degree of belief in NARS, as well as how these definitions are related to the approaches and issues discussed earlier in the article.

NARS is an *adaptive* system that can work with *insufficient knowledge and resources*. The system solves problems in real time according to its beliefs (i.e., knowledge), while new knowledge and problems show up from time to time, with unpredictable content. As a reasoning system, its beliefs and problems are all represented in a formal language, and processed according to a set of formal inference rules. The system implements a *logic*, in the sense that the language and the rules are not *ad hoc*, but based on a clearly specified semantics, which is established to capture the rationality and validity of reasoning in intelligent systems. NARS is very different from classical or the other non-classical logics, mainly due to the assumption of insufficient knowledge and resources.[7]

A major syntactical feature that distinguishes the logic of NARS from FOPL and other conventional logics is that it is a *term logic*, in which a typical statement is in the "subject-copula-predicate" format, as in Aristotle's logic. Concretely, in NARS the basic form of

[6] The NARS website is at `http://nars.wang.googlepages.com/`. The syntax of NARS' formal language and inference rules is described in Wang (1994b, 2001a), and its semantics in Wang (2005). NARS is compared with probabilistic logics in Wang (2001b), and with fuzzy logics in Wang (1996b). The website also links to an online demonstration of NARS, with working examples.

[7] A detailed discussion on this topic can be found in Wang (2006).

knowledge is an *inheritance statement*, "$S \rightarrow P$", where S and P are the *subject term* and *predicate term* of the statement, respectively, and "$\rightarrow$" is a copula representing *inheritance*, a reflexive and transitive relation from one term to another. Intuitively, the statement says that S is a specialization of P, and P is a generalization of S. Therefore "Ravens are black" can be represented as "*raven* $\rightarrow$ *black-thing*" in NARS.[8]

In NARS, an *experience-grounded semantics* is used, which defines truth-value and meaning according to the system's *experience* (i.e., input information stream). In the idealized situation, the system's experience is a set of inheritance statement defined above. Given experience K and different terms S and P, "$S \rightarrow P$" is *true* if and only if it is in K or can be derived from it (via the transitivity of the inheritance relation). The *meaning* of a term T is defined as consisting of its *extension* $T^E = \{x \,|\, x \rightarrow T\}$ and *intension* $T^I = \{x \,|\, T \rightarrow x\}$, that is, its known specializations and generalizations.

From the above definitions, it can be proved that

$$S \rightarrow P \iff S^E \subseteq P^E \iff P^I \subseteq S^I$$

that is, a *perfect* inheritance relation means the extension of the subject is *completely* included in that of the predicate, and the intension of the predicate is *completely* included in that of the subject. Furthermore, *all* evidence is already available, so there is no future evidence to be considered.

The above result shows that an inheritance statement can also be seen as a summary of many other inheritance statements. This feature is used to naturally introduce the notion of *evidence*, so as to extend the perfect inheritance relation into an *imperfect* inheritance relation, where *conflicting evidence* and *future evidence* must be considered.

From given experience K, the meaning of the terms in it, including S and P, are determined. For statement "$S \rightarrow P$", its positive evidence are terms in $S^E \cap P^E$ and $P^I \cap S^I$ (because the statement is true as far as these terms are considered), and its negative evidence are terms in $S^E - P^E$ and $P^I - S^I$ (because the statement is false as far as these terms are considered). As a result, the amounts of positive evidence, negative evidence, and total evidence of the statement "$S \rightarrow P$" are defined as the following, respectively:

$$\begin{array}{rcl} w^+ & = & |S^E \cap P^E| + |P^I \cap S^I| \\ w^- & = & |S^E - P^E| + |P^I - S^I| \\ w & = & w^+ + w^- \;=\; |S^E| + |P^I| \end{array}$$

The truth-value of a statement consists of a pair of real numbers in [0, 1], defined by the amounts of evidence:[9]

$$\begin{array}{rcl} \textit{frequency} & = & w^+/w \\ \textit{confidence} & = & w/(w+k) \end{array}$$

where k is a positive parameter, with 1 as default in the current implementation.

[8] NARS can directly use compound terms called "intensional sets" to represent adjectives, without turning them into nouns. Therefore, "Ravens are black" can also be represented in NARS as "*raven* $\rightarrow$ [*black*]". See Wang (2006) for details, though this topic has little impact on the current discussion.

[9] According to model-theoretic semantics, truth and belief are different. On the contrary, according to experience-grounded semantics, truth-value and degree of belief are the same, both determined by the extent of evidential support. This topic is discussed in detail in Wang (2005, 2006).

Comparing the above definition of evidence to the previously discussed ones, we see that it is basically Nicod's criterion, except that in NARS both the extensional aspect and the intensional aspect of the relation are taken into account. Therefore the *properties* (generalizations, intension) of the predicate are counted as evidence of the inheritance statement, just like the *instances* (specializations, extension) of the subject. Consequently, the truth-value of NARS includes a factor that is similar to the "representativeness" discussed in Tversky and Kahneman (1974, 1983), and the "conjunction fallacy" is not necessarily a fallacy anymore. A detailed discussion of this topic is in Wang (1996a).

NARS uses a term logic partly because in it the basic statements are in the "subject-copula-predicate" format, so the above definition of evidence can be easily introduced. In predicate logics, such a definition cannot be directly applied.

NARS does not suffer from the confirmation paradox, because a red pencil is not evidence for "*raven* → *black-thing*". NARS uses *compound terms* for complex statements, and they include the *extensional difference* of terms T_1 and T_2, $(T_1 - T_2)$, defined by

$$(T_1 - T_2)^E = T_1^E - T_2^E,\ (T_1 - T_2)^I = T_1^I$$

Therefore "Whatever is not black is not a raven" can be written as

$$(\mathit{thing} - \mathit{black\text{-}thing}) \rightarrow (\mathit{thing} - \mathit{raven})$$

which has the same negative evidence (i.e., non-black ravens) as

$$\mathit{raven} \rightarrow \mathit{black\text{-}thing}$$

but they have different positive evidence. Consequently, in NARS "Ravens are black" and "Whatever is not black is not a raven" have different evidence, and therefore different truth-value and meaning, though they are still related to each other semantically.

Now it is time to summarize our analysis of the confirmation paradox. Since "confirmation" is about inconclusive positive evidence, it cannot be properly introduced into a *binary* logic, where a single constant can only provide negative evidence. For the same reason, it cannot be introduced into a new logic together with the traditional equivalence condition, which only considers negative evidence. The proper solution to this paradox is not to accept the counterintuitive conclusion (e.g., A red pencil is confirming evidence for "Ravens are black"), but to drop the old equivalence condition, because it is incompatible with the notion of confirming evidence.

Some people may think that this does not count as a solution to Hempel's paradox, but a different problem. This is true in a sense, but does not disqualify the conclusion. For many problems in the history of science, their solutions turn out to be reformulations of the problems. Hempel's initial goal was to formalize the confirmation process, and when he tried to do so in the framework of binary logic, a paradox was found. The above analysis shows that the problem exists in the fundamental assumptions of the framework, in which the concept of confirming evidence cannot be properly introduced. This is a valid solution of Hempel's problem, though not in a form expected by him and many others.

Our treatment of Wason's selection task is similar, which has been explained in Wang (2001c). To evaluate the truth-value of a statement, both positive evidence and negative evidence should be collected. Since the former is often easier to be recognized and processed, what the subjects do in this experiment can be explained and even justified. The problem in the traditional interpretation of the experiment result is that many people take FOPL as the *only* normative theory of reasoning, and therefore treat any deviation from it as a mistake. According to the previous discussion, binary logic should be applied to the selection task only when violation of a statement is explicitly sought and confirming evidence are deliberately ignored. There are such situations, such as the often mentioned "underage drinking" scenario (Griggs and Cox, 1982), but they are exceptions, not normal cases, for evidential reasoning in general.

The truth-value in NARS is intuitively related to probability. The frequency value is the success rate of the inheritance statement in the past, which is often taken as an estimation of the statement's probability when the sample size is large enough. In NARS, frequency indicates the *direction* of a belief, that is, a value near 1 means the belief is affirmative (or positive), while a value near 0 means the belief is dissenting (or negative). The confidence value is the ratio of the amount of current evidence to the amount of future evidence after the coming of evidence of a constant amount, and is therefore an increasing function of the amount of total evidence. In NARS, confidence indicates the *stability* of a belief, that is, a value near 1 means the belief is already based on a large amount of evidence (so is insensitive to new evidence), while a value near 0 means the belief is based on little evidence (so is sensitive to new evidence). Frequency and confidence are independent of each other, in the sense that given the value of one, the value of the other cannot be determined or even bounded. Used together, these two values are roughly what Peirce and Keynes suggested.[10]

As argued previously, as well as in Wang (1993, 2001b, 2004), the information in the confidence measurement of NARS is not generally available in the one-number Bayesian approach, which uses a probability value to represent a degree of belief.

Furthermore, though each truth-value in NARS can be seen as corresponding to an estimated probability value plus a function of sample size, the truth-values of various beliefs that co-exist at the same time do not correspond to a consistent probability distribution on the statement space. In NARS, each inheritance statement has its own evidence scope (defined by the extension of its subject and the intension of its predicate, as described before), and due to the assumption of insufficient resources, when the truth-value of a given statement is evaluated, the system does not attempt to consider all relevant evidence. Instead, in each inference step the truth-value of the conclusion is evaluated only according to the evidence provided by the premises. Consequently, following different inference paths, the same statement can be given different truth-values, so there is no guarantee of coherence, in the sense that the truth-value of a statement is unique, independent of how it is decided.

Though coherence among beliefs is highly desired, it is not always possible to be fully achieved. For a system working with insufficient knowledge and resources, it cannot always recognize potential conflicts among its beliefs, nor can it afford the resources to fully and immediately resolve all the conflicts it finds. Instead, such a system can only try its best to base its beliefs on available evidence that can be considered with available resources.

In NARS, when the same statement gets two different truth-values from distinct bodies of evidence,[11] the *revision rule* is used to combine the evidence. The truth-value function of

[10] NARS does not directly use amounts of evidence as truth-value, because in the design of inference rules, values in [0, 1] are easier to handle than values in $[0, \infty)$.

[11] See Wang (1995a, 2006) for how the system decides whether two bodies of evidence are distinct.

this rule is directly derived from the additivity of amount of evidence in this operation. If the bodies of evidence have overlap, the *choice rule* is used to select one of the truth-values (usually the one with a higher confidence factor). These rules contribute to the solution of the reference class problem, as described in Wang (1995b, 2006).

The truth-value of NARS can be equivalently represented as an interval, too. As defined previously, the current frequency value of a statement is the proposition of positive evidence among all available evidence, w^+/w. In the near future, with the coming of evidence of amount k, the frequency value will be in the interval

$$[w^+/(w+k), (w^+ + k)/(w+k)]$$

Since enumerative induction is a special case where $w^+ = w_0 t^+$ and $w = w_0 t$, this interval is basically the same as Walley's

$$[t^+/(t+s_0), (t^+ + s_0)/(t+s_0)]$$

though interpreted differently — in NARS, the assumption on prior distribution is not made, and all measurements are defined on available evidence.

In NARS, the system's *ignorance* about the statement is also represented by the width of the above "frequency interval", like the other interval approaches. So $i = k/(w+k) = 1-c$, where c is the confidence factor. Therefore, "confidence" and "ignorance" are opposite to each other, which is consistent with the usual usage of these two words. With the coming of new evidence, the interval becomes narrower and narrower. An interval-based revision rule that combines evidence from different sources can be derived for frequency interval from the additivity of w^+ and w during revision, and it has the same form as the proposed "combination rule" for Walley's theory in the previous section.

The representation and processing of uncertainty in NARS is more similar to Walley's theory of imprecise probabilities than to the other approaches mentioned before. These two approaches not only share many intuitions, but also have identical results on certain cases. One major difference between the two is semantic interpretation. The truth-value of NARS is defined in terms of evidence, while Walley's theory starts at people's preference among options as revealed by their betting decisions. Though the probability interval in Walley's theory can be related to additive evidence, it is not the focus of the theory, so this relation is often omitted completely in descriptions of the theory, such as Walley (1996a). Walley's theory is proposed as an extension of probability theory, and therefore the inference is mainly within the same probability distribution. On the other hand, NARS is designed to be a *logic*. As described previously, in NARS each belief is based on a separate body of evidence, so that the rules correspond to inference across different evidential basis, and the *coherence principle*, a cornerstone of Walley's theory and all other probability-based approaches, is not granted in NARS.

Now let us summarize the major similarity and difference between NARS and the other approaches in the definition and representation of evidence.

Binary-valued (and three-valued) logics: Like other logics, NARS represents beliefs and evidence using a formal language, though it uses a numerical truth-value to measure evidential support, which is inconclusive, and covers both positive evidence and negative evidence.

One-number Bayesian approach: The frequency measurement in NARS is intuitively related to probability, though it is defined on available evidence, and is accompanied by a confidence measurement. The truth values do not form a probability distribution in a closed statement space. All domain knowledge can be revised by new evidence.

Dempster-Shafer theory: NARS accepts the first three postulates of D-S theory (listed previously), though rejects Dempster's rule of evidence combination. Instead, the corresponding rule in NARS is directly implied by the additivity of the amount of evidence during combination.

Two-number Bayesian approaches: NARS uses two numbers to represent the degree of belief of a statement, like some approaches in the Bayesian tradition. However, it is still different from the others in the following aspects:

- The numbers are fully defined on available evidence, without assumption on their probability distribution or behavioral implication. For an AGI system opening to novel experience, its degree of belief does not necessarily converge to a limit. Some of its beliefs do not directly correspond to observable events or actions.
- The beliefs of a system are statements in a formal language (so their number is unlimited), rather than in a constant proposition/event set. For an AGI system opening to novel experience, new evidence may contain novel statements (even statements with novel words) for which no previous belief exists.
- Each belief is based on its own chunk of evidence, so there is no guaranteed coherence among the truth-values in the system. An AGI system should try to resolve incoherence among its beliefs, though conflicting evidence cannot always be formalized as conditional probabilities that are based on different conditions, nor can the system afford the resources to exclude all contradictions implied by its beliefs.

Given the above differences, NARS indeed proposes a novel formalization of evidence, though it is still similar to the existing approaches here or there.

The inference rules of NARS, which are summarized and explained in Wang (2006), are very different from those of the other theories. This paper cannot go into the details of the rules, so here we will only say that

- Different inference rules are unified in NARS by having similar formats and usages. They including *revision, choice, deduction, induction, abduction, comparison, analogy, compound-term composition and decomposition,* and so on.
- The inference rules are justified according to the experience-grounded semantics. For a given rule, the truth-value of its conclusion is determined only by the evidence provided by the premises.
- The truth-value functions included in the inference rules are designed using "T-norm" and "T-conorm" (Bonissone, 1987; Dubois and Prade, 1982). These functions cannot be derived from probability theory, partly because some of the values cannot be interpreted as probability, and the ones that can be taken as probability still do

not belong to the same probability distribution. Even so, there are situations where they produce similar results as probability-based approaches.

In summary, the formal treatment of evidence in NARS is designed according to the considerations of AGI research, and the result is consistent with our understanding of human intelligence. Furthermore, many related traditional problems are consistently handled.

7. Conclusion

The major conclusion of this article is the previously mentioned "Peirce-Keynes Thesis", which can be expressed, for our current purpose, as the following:

> *For a general-purpose system to base its beliefs on available evidence and to process novel evidence in real time, it is necessary to use two numbers to measure the system's degree of belief.*

These two numbers can be defined and used in different ways. For example, in NARS the same information can be represented as amount of evidence (w^+ and w), truth-value (f and c), or frequency-interval ($[l, u]$), and the system can switch among the three representations, plus some variants of them.

Though some results in this article are known to logicians or statisticians, the above conclusion is not trivial, because in most existing AI systems, degree of belief (or whatever it is called) is still either represented qualitatively, or measured using a single number, usually a probability. Though two-number approaches are common in fields like statistics, they are rare in AI, especially in implemented systems. This situation is to a large extent caused by the underestimation of the limitation of the binary-value and one-number approaches.

Binary logic, as exemplified by FOPL, can properly represent and handle beliefs with *conclusive* evidence, but cannot do that with *inconclusive* evidence. To introduce such evidence leads to counterintuitive results, as shown by Hempel's confirmation paradox and Wason's selection task.

The one-number Bayesian approach can represent inconclusive evidence in a simple and natural way, but has limitation in revising the current beliefs according to new evidence, because the ignorance of the system cannot be captured as conditional probability. As a result, it should be used only when a probability distribution function can be established on stable background knowledge.

To support revision in general, it is necessary to attach two numbers to each belief. Though there are more than one way to do it, the evidence combination rule should be consistent with the additivity of the amount of evidence. Furthermore, it is desired for the measurements to converge to probability as an extreme case. When the coherence of probability can be achieved, the imprecise probability theory or other two-number Bayesian approaches may work.

The representation and processing of evidence in NARS is developed specially for general-purpose intelligent systems, and is based on the assumption of insufficient knowledge and resources. Consequently, it is designed to work in realistic situations, and can carry out inference without making strong assumption on the environment, or requiring huge amount of resources to keep the coherence of the beliefs.

Acknowledgments

The author benefits from discussions with Ben Goertzel and Matthew Iklé on related issues. Jeremy Zucker made many concrete suggestions and English corrections to an early draft of the article. Paul Snow and Gabriele Kern-Isberner wrote detailed reviews of the submission and made major contributions to the revision of the article.

References

Achinstein, P., ed. 1983. *The Concept of Evidence.* Oxford: Oxford University Press.

Baroni, P., and Vicig, P. 2001. On the conceptual status of belief functions with respect to coherent lower probabilities. In Bishop, C., ed., *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty; Lecture Notes In Computer Science, Vol. 2143.* London: Springer-Verlag. 328–339.

Bonissone, P. P. 1987. Summarizing and propagating uncertain information with Triangular Norms. *International Journal of Approximate Reasoning* 1:71–101.

Carnap, R. 1950. *Logical Foundations of Probability.* Chicago: The University of Chicago Press.

Cheeseman, P. 1988. An inquiry into computer understanding. *Computational Intelligence* 4:58–66.

Clifford, W. K. 1877. The ethics of belief. *Contemporary Review.* Reprinted in The Ethics of Belief and Other Essays (Prometheus Books, 1999).

DeGroot, M. H. 1970. *Optimal Statistical Decisions.* New York: McGraw-Hill.

Dempster, A. P. 1967. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* 38:325–339.

Dubois, D., and Prade, H. 1982. A class of fuzzy measures based on triangular norms. *International Journal of General Systems* 8:43–61.

Dubois, D., and Prade, H. 1994. Conditional objects as nonmonotonic consequence relationships. *IEEE Transactions on Systems, Man, and Cybernetics* 24:1724–1740.

Fitelson, B., and Hawthorne, J. 2009. How Bayesian confirmation theory handles the Paradox of the Ravens. In Eells, E., and Fetzer, J., eds., *Probability in Science.* Chicago: Open Court. Forthcoming.

Good, I. J. 1950. *Probability and the Weighing of Evidence.* London: Griffin.

Good, I. J. 1985. Weight of evidence: a brief survey. In Bernardo, J.; DeGroot, M.; Lindley, D.; and Smith, A., eds., *Bayesian Statistics 2.* Amsterdam: North-Holland. 249–269.

Griggs, R. A., and Cox, J. R. 1982. The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology* 73:407–420.

Halpern, J. Y., and Pucella, R. 2006. A logic for reasoning about evidence. *Journal of Artificial Intelligence Research* 26:1–34.

Hempel, C. G. 1965. Studies in the logic of confirmation. In *Aspects of Scientific Explanation.* New York: The Free Press. 3–46. Reprinted in The Concept of Evidence, Achinstein, P. (Ed), Oxford University Press, pp. 11–43, 1983.

Hume, D. 1748. *An Enquiry Concerning Human Understanding.* London.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Berlin: Springer.

Keynes, J. M. 1921. *A Treatise on Probability.* London: Macmillan.

Kyburg, H. E. 1983a. Recent work in inductive logic. In Lucey, K., and Machan, T., eds., *Recent Work in Philosophy.* Totowa, NJ: Rowman and Allanfield. 89–150.

Kyburg, H. E. 1983b. The reference class. *Philosophy of Science* 50:374–397.

Kyburg, H. E. 1994. Believing on the basis of the evidence. *Computational Intelligence* 10:3–20.

McCarthy, J., and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., and Michie, D., eds., *Machine Intelligence 4.* Edinburgh: Edinburgh University Press. 463–502.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence* 3:151–160.

Milne, P. 1997. Bruno de Finetti and the logic of conditional events. *The British Journal for the Philosophy of Science* 48:195–232.

Oaksford, M., and Chater, N. 1994. A rational analysis of the selection task as optimal data selection. *Psychological Review* 101:608–631.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems.* San Mateo, California: Morgan Kaufmann Publishers.

Pearl, J. 1990. Jeffrey's rule, passage of experience, and Neo-Bayesianism. In Kyburg, H. E.; Loui, R. P.; and N., C. G., eds., *Knowledge Representation and Defeasible Reasoning.* Amsterdam: Kluwer Academic Publishers. 245–265.

Peirce, C. S. 1878. The probability of induction. *Popular Science Monthly* 12:705–718. Reprinted in *The Essential Peirce, Vol. 1*, N. Houser and C. Kloesel, eds., Bloomington, IN: Indiana University Press (1992), 155–169.

Popper, K. R. 1959. *The Logic of Scientific Discovery.* New York: Basic Books.

Reiter, R. 1987. Nonmonotonic reasoning. *Annual Review of Computer Science* 2:147–186.

Rescher, N. 1958. A theory of evidence. *Philosophy of Science* 25(1):83–94.

Shafer, G. 1976. *A Mathematical Theory of Evidence.* Princeton, New Jersey: Princeton University Press.

Smets, P., and Kennes, R. 1994. The transferable belief model. *Artificial Intelligence* 66:191–234.

Smets, P. 1991. The transferable belief model and other interpretations of Dempster-Shafer's model. In Bonissone, P. P.; Henrion, M.; Kanal, L. N.; and Lemmer, J. F., eds., *Uncertainty in Artificial Intelligence 6.* Amsterdam: North-Holland. 375–383.

Solomonoff, R. J. 1964. A formal theory of inductive inference. Part I and II. *Information and Control* 7(1-2):1–22,224–254.

Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185:1124–1131.

Tversky, A., and Kahneman, D. 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review* 90:293–315.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities.* London: Chapman and Hall.

Walley, P. 1996a. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B* 58:3–57.

Walley, P. 1996b. Measures of uncertainty in expert systems. *Artificial Intelligence* 83:1–58.

Wang, P. 1993. Belief revision in probability theory. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, 519–526. Morgan Kaufmann Publishers, San Mateo, California.

Wang, P. 1994a. A defect in Dempster-Shafer Theory. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 560–566. Morgan Kaufmann Publishers, San Mateo, California.

Wang, P. 1994b. From inheritance relation to nonaxiomatic logic. *International Journal of Approximate Reasoning* 11(4):281–319.

Wang, P. 1995a. *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence.* Ph.D. Dissertation, Indiana University.

Wang, P. 1995b. Reference classes and multiple inheritances. *International Journal of Uncertainty, Fuzziness and and Knowledge-based Systems* 3(1):79–91.

Wang, P. 1996a. Heuristics and normative models of judgment under uncertainty. *International Journal of Approximate Reasoning* 14(4):221–235.

Wang, P. 1996b. The interpretation of fuzziness. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 26(4):321–326.

Wang, P. 2001a. Abduction in non-axiomatic logic. In *Working Notes of the IJCAI workshop on Abductive Reasoning*, 56–63.

Wang, P. 2001b. Confidence as higher-order uncertainty. In *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, 352–361.

Wang, P. 2001c. Wason's cards: what is wrong? In *Proceedings of the Third International Conference on Cognitive Science*, 371–375.

Wang, P. 2004. The limitation of Bayesianism. *Artificial Intelligence* 158(1):97–106.

Wang, P. 2005. Experience-grounded semantics: a theory for intelligent systems. *Cognitive Systems Research* 6(4):282–302.

Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence.* Dordrecht: Springer.

Wason, P. C., and Johnson-Laird, P. N. 1972. *Psychology of Reasoning: Structure and Content.* Cambridge, Massachusetts: Harvard University Press.

Zadeh, L. A. 1975. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences* 8:199–249, 8:301–357, 9:43–80.