The Zoo Keeper's Paradox
A Decision Theoretic Analysis of Inconsistent Commonsense Beliefs

Kwok Hung Chan
Dept. of Mathematical Sciences
Memphis State University
Memphis, TN 38152
U.S.A.
chand@hermes.msci.memst.edu

Abstract

Default reasoning is a mode of commonsense reasoning which lets us jump to plausible conclusions when there is no contrary information. A crucial operation of default reasoning systems is the checking and maintaining of consistency. However, it has been argued that default reasoning is inconsistent: Any rational agent will believe that it has some false beliefs. By doing so, the agent guarantees itself an inconsistent belief set (Israel, 1980). Perlis (1986) develops Israel's argument into an argument for the inconsistency of recollective Socratic default reasoning systems. The Zoo Keeper's Paradox has been offered as a concrete example to demonstrate the inconsistency of commonsense beliefs.

In this paper, we show that Israel and Perlis' arguments are not well founded. A rational agent only needs to believe that some of its beliefs are possibly or probably false. This requirement does not imply that the beliefs of rational agents are necessarily inconsistent. Decision theory is used to show that concrete examples of seemingly inconsistent beliefs, such as the Zoo Keeper's Paradox, can be rational as well as consistent. These examples show that analyses of commonsense beliefs can be very misleading when utility is ignored. We also examine the justifications of the exploratory and incredulous approaches in default reasoning, decision theoretic considerations favor the exploratory approach.

Default Reasoning

The goal of artificial intelligence is to build electronic agents which can use knowledge to solve problems. A large part of what we know is commonsense knowledge consisting of general laws/rules which are almost always true, with a few exceptions (Reiter, 1980).

Example 1:

(R1) We can start a car by turning the key while stepping slowly on the gas pedal.

R1 is a general rule which is almost always true. There are exceptions to the rule, such as, "the gas tank is empty" and "the battery is low." However, usually we do not check to make sure that everything is normal. We just assume *by default* that the car is in working condition, *unless* there is information to the contrary. Say, if we notice that the ignition switch is lying on the floor with a loose wire, then we conclude that the car is out of order and it will not start. Such reasoning is not deductive, and has been called *default reasoning* in the literature.

Example 1 illustrates the *non-monotonicity* of default reasoning: A sentence A which is derivable from a theory T may not be derivable from a superset of T. Because of this property, formalizations of default reasoning have been called *non-monotonic logics*. In this paper we will use "non-monotonic logics" as a general term covering all formalizations of default reasoning with the property of non-monotonicity.[1]

The Consistency of Default Reasoning

A number of authors have worried about the integrity and consistency of commonsense/default reasoning. Israel (1980) claims that non-monotonic logics are not well motivated, because they rest on the confusion of proof-theoretic with epistemological issues. Israel also suggests that commonsense beliefs are very often inconsistent. Since most non-monotonic logics perform

[1] Ginsberg (1987b) contains original papers of major works before 1987. Besnard (1989) is a more recent introduction to non-monotonic logics.

a consistency test before making a default assumption, they would be paralyzed by inconsistent beliefs.

Perlis (1986) develops Israel's argument for the inconsistency of commonsense beliefs into an argument for the inconsistency of non-monotonic logics under some natural conditions. According to Perlis, ideal thinkers capable of appropriate commonsense reasoning must be able to reflect on their past errors. They must be aware of the fallibility of their use of defaults (Socratic) and able to recall what default assumptions they have made (recollective). However, recollective Socratic reasoning is inconsistent. Perlis also presents the Zoo Keeper's Paradox as a concrete example that illustrates the inconsistency of commonsense beliefs. The performance of the major formalizations, namely, Circumscription (McCarthy, 1980), Non-monotonic Logic (McDermott & Doyle, 1980) and Default Logic (Reiter, 1980), is compromised: they do not produce the intuitively correct commonsense default conclusions in cases like the Zoo Keeper's Paradox.

In this section, we will consider briefly the general arguments for the inconsistency of commonsense beliefs and default reasoning. A detailed analysis is presented in Chan (1992).

The Goal of Non-monotonic Logics

Israel's (1980) major complaint about non-monotonic logics is that the motivation behind non-monotonic logics is based on a confusion of proof-theoretic with epistemological issues. This has been misinterpreted by some authors as an issue of terminology: Logic is, by its very definition, monotonic, and the notion of "non-monotonic logic" is a contradiction in terms (Ginsberg, 1987a). Such misinterpretation misses the point of Israel's argument as well as the chance to show that Israel is mistaken.

Default reasoning makes a default assumption A only if there is there is no information to the effect that A is false. This requirement is implemented in non-monotonic logics as a consistency check. Before A is concluded by default, the system checks to see if A is consistent with the set of current beliefs. A is also required to be consistent with the justifications of default assumptions made previously. Hence, a default assumption will remain consistent with subsequent default beliefs. This consistency requirement is interpreted by Israel as follows:

> [To make an assumption that A is to believe that A is] both compatible with everything that a given agent believes at a given time and *remains so* when the agent's belief set undergoes certain kinds of changes *under the pressure of* both *new information* and further thought, and where those changes are the result of rational epistemic policies (Israel, 1980).

Based on this understanding Israel takes non-monotonic logics to be formal systems for *general belief fixation*. He then argues that there is no logic of belief fixation and scientific procedures are the only methods for belief fixation. Because there are no logics of belief fixation, non-monotonic logics can never achieve their goal.

Even if Israel is correct in claiming that there is no logic of belief fixation, his criticism of non-monotonic logics is not justified. This is because he has misinterpreted the consistency requirement and the aim of non-monotonic logics. Non-monotonic logics do not aim to be general logics for belief fixation. A default assumption A is *not* required to remain consistent when we add *new information*. Actually, the non-monotonic nature of default reasoning requires that A should be deleted when it is not consistent with new information! The correct intuitive understanding of the consistency requirement is: A new default assumption A should be consistent with current beliefs and should not falsify the justifications of default assumptions *previously* made. Hence, a default assumption is only guaranteed to remain consistent with *subsequent default beliefs*. Non-monotonic logics are not logics for making general hypotheses. That is the job of scientific procedures. Non-monotonic logics have a rather moderate aim. They are logics for the proper extensions of beliefs by, and *only by*, default assumptions supported by default rules such as "Typically P's are Q's."

The Consistency of Commonsense Beliefs

Israel also argues that the consistency requirement cannot be met in practice, because commonsense beliefs are mostly inconsistent. Any rational agent will believe that it has some false beliefs. By doing so, the agent guarantees itself an inconsistent belief set; there is no possible interpretation under which all of its beliefs are true (Israel, 1980).

If being rational requires our having an inconsistent set of beliefs, this notion of rationality is too strong, and should be replaced by a weaker notion. To be rational an agent does *not* need to believe that it *actually* has some false beliefs. It only needs to believe that some of its beliefs are *possibly or probably* false. Such belief sets may be consistent. Hence commonsense beliefs of a rational agent are not necessarily inconsistent.

Recollective Socratic Agents

Perlis (1986) develops Israel's argument for the inconsistency of commonsense beliefs into an argument for the inconsistency of default reasoning under some natural conditions.

According to Perlis, default reasoning consists of a sequence of steps involving, in its most general form, oracles, jumps, and fixes. Since consistency check is only semidecidable, we need to appeal to an *oracle* to tell us that a given default assumption is consistent with the current beliefs. Because default reasoning *jumps* to conclusions, it is error-prone and *fixes* are necessary to preserve (or re-establish) consistency. For rational agents to be capable of appropriate commonsense reasoning, they must be able to reflect on their past errors, and indeed, on their potential future errors. They must be aware of the fallibility of their use of defaults (Socratic) and able to recall what default assumptions they have made (recollective). Perlis (1986) shows that recollective Socratic agents are inconsistent.

As in the case of Israel's argument, Perlis' definition of Socratic thinkers is too strong. A rational agent does *not* need to believe each of its default assumptions and *simultaneously* believes that some of its default beliefs are in fact false. A rational agent only needs the weaker belief that some of its default beliefs are *possibly* false. Such recollective *weakly Socratic* agents are not necessarily inconsistent.

The Zoo Keeper's Paradox

In additional to a general argument for the inconsistency of rational agents, Perlis also offers the Zoo Keeper's Paradox as a concrete example of inconsistent commonsense beliefs.

Example 2: (the Zoo Keeper's Paradox)
Bob works as a zoo keeper and keeps a written record of the animals there. Ten American bare eagles have been recorded by Bob as in good health (and so able to fly). One day Bob receives a message from a laboratory saying that blood samples from the eagles show that some eagles in the zoo are infected by virus (and as a result cannot fly). However, the laboratory has mixed up the blood samples, so we cannot tell which eagle is infected. Bob still believes that each individual eagle at the zoo can fly, that he is highly unwilling to leave any of their cage doors open, and that he is also unwilling to call any one of them to the attention of the zoo veterinarian. Yet, he is also very concerned at the verterinarian's failure to arrive for work at the usual hour, because he also believes that some (unspecified) eagles in the zoo are sick (and cannot fly).[2]

Are Bob's beliefs consistent? What conclusions should a default reasoning system make? We may formalize the hard facts in this example as follows:

$$\text{eagle}(e_1) \wedge \cdots \wedge \text{eagle}(e_{10}) \quad \text{(C1)}$$

$$\text{sick}(e_1) \vee \cdots \vee \text{sick}(e_{10}) \quad \text{(C2)}$$

$$\text{eagle}(x) \wedge \text{sick}(x) \rightarrow \neg\, \text{fly}(x) \quad \text{(C3)}$$

We have the following default rule:

Eagles typically fly. (D)

Let us consider what default assumptions we should rationally make. For each eagle we would like to conclude that it can fly by default, because we do not have specific information about any individual eagle that it cannot fly. Applying this reasoning we conclude that the first nine eagles (in some arbitrary order) fly.

$$\text{fly}(e_1) \quad \text{(A1)}$$

$$\vdots$$

$$\text{fly}(e_9) \quad \text{(A9)}$$

There are three possibilities regarding the last eagle e_{10}:

1. Since we have no evidence to single out e_{10} from the rest, we may apply the same reasoning and conclude by default that e_{10} can fly. However, the addition of this last default conclusion results in a set of inconsistent beliefs. According to Perlis this is what Bob believes.
2. We may deduce from C1–C3 and A1–A9 that e_{10} does not fly. Since there are ten ways to pick this last eagle, there are ten possible extensions, each as good as the other. An *exploratory* system (Reiter, 1980) would pick an arbitrary extension.
3. An *incredulous* system (McCarthy, 1980; McDermott & Doyle, 1980) would consider as default conclusions only those shared by all extensions:
   Nine of the eagles fly. (G1)
   One of the eagles does not fly. (G2)

In the rest of this section we will consider the consistency of Bob's beliefs. The exploratory and incredulous approaches will be examined in the next section.

Are Bob's beliefs inconsistent? First of all, how do we know what Bob's beliefs are? Perlis proposes a behavioral criterion of *use-belief*.

Definition 1: (use-belief)
An agent believes a proposition p if it trusts and uses p in planning and acting, "as if it were true." The agent should be willing to recognize p as theorems and ignore the possibility that p may be false. If the agent also does something that is appropriate only if p is false, then the agent only believes that it is highly probable that the proposition is true (Perlis, 1986).

This definition tries to identify an agent's beliefs by its actions. However, actions are not determined only by beliefs. The rationality of an action also depends on its utility. In what follows, we will apply decision theory (Savage, 1972) to find out if Bob's belief/behavior is rational.

[2] This is a modified version of the Zoo Keeper's Paradox in Perlis, 1986. A similar paradox is the Lottery Paradox discussed in McDermott (1982), Shoham (1987) & Poole (1991).

Suppose Bob believes with probability p that an eagle e is sick. Being a zoo keeper, Bob is responsible for keeping e healthy as well as keeping e in the zoo. Let us represent the utility of different possibilities as follows:

Event	escape	stay	dead	sick	healthy
Utility	0	1	0	v	u

Table 1. Utility of events for Bob

We will consider the decision trees for closing/opening the cage door in two scenarios.

Scenario 1: Closing the cage door does not make the condition of a sick eagle worse.

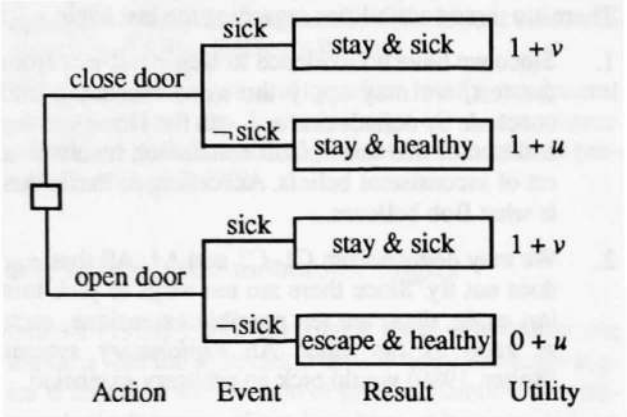

Figure 1. Decision tree for Scenario 1

The expected utility of closing the cage door is greater than the expected utility of opening the cage door by $(1-p)$ $(= p(1+v) + (1-p)(1+u) - p(1+v) - (1-p)u)$. As long as Bob is not absolutely certain that e is sick ($p < 1$), he better keeps the door closed. If $p = 1$, then it makes no difference if the door is closed or open. Hence, in Scenario 1 Bob should close the cage door no matter whether he believes that e is sick or not. Bob is probably in this situation in Example 2. This shows the possibility of interpreting Bob's behavior as rational without attributing an inconsistent set of beliefs to him. What he does is rational and is consistent with his belief that one of the eagle is sick.

Scenario 2: A sick eagle will die if the cage door is closed.

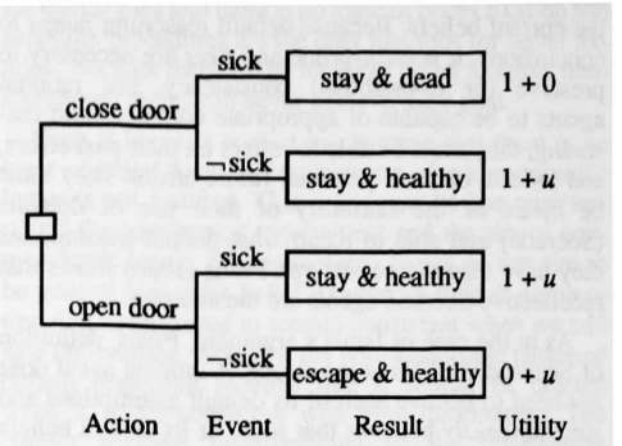

Figure 2. Decision tree for Scenario 2

The expected utility of closing the cage door is greater than the expected utility of opening the cage door by $(1-p) - pu$ $(= p + (1-p)(1+u) - p(1+u) - (1-p)u)$. The values of $EU(\text{close}) - EU(\text{open})$ for some representative values of p and u are shown in Table 2. In this scenario Bob has to make a choice between opening or closing the cage door. What he should do depends on p as well as u. If keeping the eagle in the zoo is as important as keeping the eagle healthy $(u \approx 1)$, then Bob should keep the cage door closed if he believes that probably the eagle is healthy, but he should open the cage door if he believes that probably the eagle is sick. However, if the eagle is an endangered species and it is very important to keep it healthy $(u \gg 1)$, he should open the cage door even if he believes that probably the eagles are not sick $(p \leq 0.5)$. Because the penalty for mistake is so high, it is rational for Bob to consider an unlikely proposition (sick(e)) to be true by default.

Scenario 2 is a special case in which two goals compete for an action. Keeping the cage door closed achieves the goal of keeping the eagle in the zoo, but violates the goal of keeping the eagle healthy. According to the policy of minimizing expected loss, it is rational to perform the action appropriate to the unlikely event if the penalty of overlooking the event is too

EU(close) $-$ EU(open)		u							
		.01	.11	.5	1	5	9	10	20
p	.1	.899	.889	.85	.8	.4	0	- 0.1	- 1.1
	.5	.495	.445	.25	0	- 2	- 4	- 4.5	- 9.5
	.9	.091	0	- 0.35	- 0.8	- 4.4	- 8	- 8.9	- 17.9

Table 2. The value of EU(close) – EU(open) for different p and u.

great. In other scenarios, the required remedy/preventive action for an unlikely event may not compete with the normal action appropriate to the more likely event. For example, the probability p of having a car collision is low. However, the penalty of not wearing a seat belt is very high if a car collision does occur. Fortunately, the preventive measure of wearing a seat belt can be performed simultaneously with other actions appropriate for the much more likely event of no car collision. In such cases we usually entertain two belief sets, belief set S1 is consistent with the occurrence of a normal event E, whereas belief set S2 is consistent with the unlikely possibility of $\neg E$. Actions appropriate to S1 are performed if they do not compete with remedy or preventive actions appropriate to $\neg E$.

The Zoo Keeper's Paradox illustrates that it is highly misleading to consider rational behavior without taking utility into account. Given a belief set, rational behavior is determined by utility/penalty. It is rational to make a default assumption only if the penalty for making a mistake is (very) low. If the penalty is great enough, even an unlikely proposition should be considered to be true by default: preventive measures or remedies are implemented as if the unlikely event will occur or has occurred.

Exploratory vs. Incredulous Approaches

Can we apply decision theory in the context of non-monotonic logics? Although probability and utility are *not* considered *within* non-monotonic logics, the *practice* of default reasoning is justified by decision theoretic considerations. Under the normal operating conditions of default reasoning, each default rule has a high probability of being true, it is desirable to draw the default conclusions and the penalty for drawing a false conclusion is low. In this section we will consider the decision theoretic justifications of the exploratory and incredulous approaches in default reasoning when these normal operating conditions are satisfied.

Let us take a detour and consider Bob's beliefs before he received the message from the laboratory. At that time Bob was presumably justified by the default rule D to conclude that all ten eagles can fly. Suppose that the actual world is a *maximally typical world* in which all individuals not known to be atypical are indeed typical, then all ten eagles can fly. Of course, the actual world is not maximally typical. However, because exceptions are rare, *most of our default conclusions are true*. The successful rate depends on how typical the world is. Although we may make mistakes from time to time, that is acceptable, because the penalty for such mistakes are low and in the long run true default assumptions out number mistaken default assumptions.

Now, consider again the original version of the Zoo Keeper's Paradox in which Bob knows that at least one of the eagles is sick. Since consistency check is an essential step in the normal operation of non-monotonic logics, the inconsistent belief set acknowledging ten healthy eagles cannot be tolerated. There are ten different consistent (maximal) extensions of the core beliefs. An incredulous non-monotonic logic does not commit itself to any one of the competing extensions. Only default conclusions shared by all consistent maximal extensions are made. Such shared conclusions are true in all maximally typical world with one sick eagle, and we can appeal to the same statistical justification for this conservative approach.

Should we commit ourselves to any one of the ten extensions? If we have some empirical evidence that makes one of eagle, say e_1, the prime suspect, then we should prefer an extension in which e_2 to e_{10} can fly. Otherwise, all ten extensions are equally justified and we have no reason to prefer one rather than another.

Suppose Bob knows that exactly one of the eagles is sick and he uses an exploratory non-monotonic logic to pick one of the ten extensions. Do we have any statistical justification for such a practice? There are ten extensions, so the chance of getting *all ten* default conclusions right is only 10%. This is a low percentage. However, let us compute the expected number of correct conclusions. Let p_i be the probability that extension i is correct and N_i be the number of correct conclusions if extension i is correct. The expected number of correct conclusions is

$$\sum_{i=1}^{10} N_i \times p_i = 10 \times 0.1 + 9 \times 8 \times 0.1 = 8.2.$$

The expected number of correct conclusions is summarized in Table 3.

Suppose Bob knows that at least one eagle is sick. Using an exploratory default logic, he would pick an extension with only one eagle being sick. If the world is a maximally typical world, then only one eagle is sick and the expected number of correct default conclusions is the same as the previous case (8.2). However, if each eagle has a 50% chance of being sick, then the expected number of correct default conclusions is reduced to only 4.6. In general if the default rule in question is very strong with a very high percentage of typical members,

No. of eagles known to be sick	1	2	3	4	5	6	7	8	9
Expected no. of correct conclusions	8.2	6.8	5.8	5.2	5	5.2	5.8	6.8	8.2

Table 3. Expected number of correct default conclusions

then the expected number of correct default conclusions would be very close to 8.2. Although there is no empirical reason to prefer one extension to another, *any one* of the extensions would serve just as well. Using an exploratory default logic, we can always backtrack and try another extension when we find out later that we have picked the wrong one.

From this example, we can see that an exploratory default logic may be justified even when we do not have any empirical evidence to prefer one extension over another. Moreover, under the normal operating conditions of non-monotonic logics, each default rule has a high probability of being true and incurs a very low penalty for false conclusions. Using an exploratory system we can make more default assumptions without incurring heavy penalty. In the long run the advantages of making more correct assumptions will outweigh the small penalty incurred by occasional false conclusions. On the other hand, we will miss the chance to make many useful default assumptions if we follow the incredulous approach.

In special cases where the normal operating conditions of default reasoning are not satisfied, neither the exploratory nor the incredulous approach would work as such. We need a more powerful mode of reasoning which can entertain competing belief sets and act on the basis of both sets. It is interesting to see how we can extend current non-monotonic logics or develop new systems to handle defaults with heavy penalty.

Conclusions

In this paper, we have proposed that a rational agent only needs to believe that some of its beliefs are possibly or probably false. This requirement does not imply that the beliefs of rational agents are necessarily inconsistent. Decision theory is used here to show that concrete examples of seemingly inconsistent beliefs can be rational as well as consistent. Such examples show that analysis of commonsense beliefs can be very misleading when utility is ignored. Justifications of the exploratory and incredulous approaches in default reasoning are examined and decision theoretical considerations favor the exploratory approach.

The use of decision theoretic analysis in default reasoning solves some old issues but also presents some new challenges. In particular, there are two types of default assumptions: (i) Some propositions are assumed to be true by default because they are probable. (ii) Other propositions are assumed to be true by default because it is too risky to assume that they are false. Default assumptions of the second type are numerous in practical applications and we need to extend existing systems or develop new systems to incorporate this overlooked type of default reasoning.

References

Besnard, P. 1989. *An Introduction to Default Logic*. Berlin: Springer-Verlag.

Chan, K. H. 1992. The Consistency of Commonsense and Default Reasoning. Unpublished.

Ginsberg, M. L. 1987a. Introduction to Readings in Nonmonotonic Reasoning. In Ginsberg, M. L. Ed. *Readings in Nonmonotonic Reasoning*, 1–23. Los Altos, CA: Morgan Kaufmann.

Ginsberg, M. L. Ed.. 1987b. *Readings in Nonmonotonic Reasoning*. Los Altos, CA: Morgan Kaufmann.

Israel, D. J. 1980. What's Wrong with Non-monotonic Logic? In *Proceedings of the First Annual National Conference on Artificial Intelligence*, 99–101. Los Altos, CA: Morgan Kaufmann.

McCarthy, J. 1980. Circumscription—A Form of Non-Monotonic Reasoning. *Artificial Intelligence*, 13:27–39.

McDermott, D. V. 1982. Nonmonotonic Logic II: Nonmonotonic Modal Theories. *JACM*, 29(1):33–57.

McDermott, D. & Doyle, J. 1980. Non-Monotonic Logic I. *Artificial Intelligence*, 13:41–72.

Perlis, D. 1986. On the Consistency of Commonsense Reasoning. *Computational Intelligence*, 2:180–190.

Poole, D. 1991. The Effect of Knowledge on Belief: Conditioning, Specificity and the Lottery Paradox in Default Reasoning. *Artificial Intelligence*, 49:281–307.

Reiter, R. 1980. A Logic for Default Reasoning. *Artificial Intelligence* , 13:81–132.

Savage, L. J. 1972. *The Foundations of Statistics, 2nd Ed.*. New York: Dover.

Shoham, Y. 1987. A Semantical Approach to Nonmontonic Logics. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 388–392. Los Altos, CA: Morgan Kaufmann.