Artificial Intelligence I

Matthew Huntbach, Dept of Computer Science, Queen Mary and Westfield College, London, UK E1 4NS. Email: mmh@dcs.qmw.ac.uk.

Notes on Reasoning with Uncertainty

So far we have dealt with knowledge representation where we know that something is either true or false. Stepping beyond this assumption leads to a large body of work in AI, which there is only time in this course to consider very briefly. Three different approaches, representing three different areas of uncertainty, are considered. The first two represent two different forms of quantitative uncertainty; that is where we attempt to give numerical values expressing the degree to which we are uncertain about pieces of knowledge. Probabilistic reasoning deals with cases where something is definitely true or false, but we do not know which and so we reason about the chances that something is true or false. Fuzzy logic deals with cases where something could be partially true; we can think of it as dealing with "shades of grey" rather than the "black or white" of classical logic.

The third approach, truth maintenance systems, is just one example of a non-monotonic reasoning system, that is one where adding new items of knowledge may cause conclusions we had previously drawn to become invalid. It is a form of common-sense reasoning, as mentioned in the first set of notes for this course, where we can start off by making assumptions which may not actually be correct, so that we can express any conclusion drawn as dependent on assumptions, and we can change assumptions as necessary, particularly when we find they lead to conflicting conclusions.

Probabilistic Reasoning

I mentioned briefly in the set of notes on predicate logic that one of the problems with the use of classical logic as a knowledge representation system is that it assumes that we can define all values as either completely true or completely false. In many cases we may wish to assign fractional values representing the fact that we are either uncertain as to whether something is true or false, but have some idea as to which is more likely, or that we wish to regard a value as only partially true.

On the uncertainty issue, the most long-established way of dealing with it is probability theory. There is not space in this course to go into much detail on probability theory (there are other courses taught in this college which do that) so I will just give some basics which fit into the approach to knowledge representation taken so far.

In the probabilistic approach, a logic sentence is labelled with a real number in the range 0 to 1, 0 meaning the sentence is completely false, and 1 meaning it is completely false. A label of 0.5 means there is equal chance that the sentence is true or false.

Simple probability is often illustrated by games of chance. If a coin is flipped, there is an equal chance of it landing on the heads side or the tails side, so if we say that H_1 stands for "on the first toss of the coin it lands heads-up", we can express this by $pr(H_1)=0.5$. Let us say that H_2 stands for "on the second toss of the coin it lands heads-up", then we know that $pr(H_2)$ is 0.5 as well. It is well known that the probability of two successive events occurring which do not interfere or depend on each other can be obtained by multiplying the probabilities of each occurring. We know that whether a coin lands heads or tails is in no way dependent in how it previously landed (assuming the coin is perfectly balanced), so that the probability of the first and second toss both landing on heads is 0.5*0.5=0.25. We can write this $pr(H_1 \wedge H_2)=0.25$, and in general for two independent events P and Q, $pr(P \wedge Q)=pr(P)*pr(Q)$.

In many cases, however, we cannot assume this independence. For example, let us suppose that in a college of 1000 students, 100 students take the logic course and 200 take the programming course. If P_F stands for "student Fred takes the programming course" and L_F stands for "Student Fred takes the logic course" where we know no more about Fred except that he is one of the 1000 students at the college, we can say that $pr(P_F)=0.2$ and $pr(L_F)=0.1$. It would be wrong to conclude though that $pr(P_F \wedge L_F)=0.2*0.1=0.02$, since we know that in fact the two subjects are related and many who take one would take the other. If however, we have another student, Wilma, and all we know about Wilma is that she is also a student at the same college, then we can correctly conclude

$pr(P_F \wedge L_W)=0.2*0.1=0.02$ where L_W stands for "Wilma takes logic" since what one arbitary student takes has no dependence on what another arbitary student takes. In all cases in probability theory if $pr(P)$ is the probability that P is true, then the probability that P is false is $1-pr(P)$, so we can say that $pr(P)=1-pr(\neg P)$.

If we know in fact that 60 of the 200 students taking the programming course also take the logic course, then we can say that the probability that Fred takes the logic course if we know that he takes the programming course is $\frac{60}{200}$ = 0.3. We write this $pr(L_F|P_F)=0.3$, where in general $pr(Q|P)$ means "the probability of Q when we know that P is true". So from this, if we don't know whether Q is true, but we know the probability of Q being true if P is true, we can calculate the probability of P and Q being true: $pr(P \wedge Q)=pr(P)*pr(Q|P)$. So $pr(P_F \wedge L_F)= 0.2*0.3=0.06$. If P and Q are independent then $pr(Q|P)=pr(Q)$.

Since $\wedge$ is a commutative operator (i.e., we can swap its arguments around without affecting its meaning) $pr(P \wedge Q)=pr(Q \wedge P)$, so $pr(P)*pr(Q|P)=pr(Q)*pr(P|Q)$. From this, it follows that:

$$pr(P|Q)=\frac{pr(P)*pr(Q|P)}{pr(Q)} \quad .$$

This is known as *Bayes' rule* after the philosopher who discovered it. In our example, it means that the probability that Fred takes the programming course if we know that he takes the logic course is $\frac{0.2*0.3}{0.1}$ =0.6. Typically this rule is used to find the probability of some hypothesis P being true given that we know some evidence Q.

Note that since $P \vee Q$ is equivalent to $\neg(\neg P \wedge \neg Q)$, then using the rule for $pr(\neg P)$ above we can obtain $pr(P \vee Q) = pr(P)+pr(Q)-pr(P \wedge Q)$. We can also obtain a value for $pr(P \rightarrow Q)$ in this manner, using the fact that $P \rightarrow Q$ is just another way of writing $\neg P \vee Q$. A distinction should be made between $pr(P \rightarrow Q)$ and $pr(Q|P)$, which are two separate things. $pr(P \rightarrow Q)$ is just the probability that $P \rightarrow Q$ holds for some arbitary x, recalling that $P \rightarrow Q$ is always true when P is false.

A more general statement of Bayes' theorem covers the case where we have some evidence E, and a range of hypotheses, $H_1, H_2, \ldots, H_n$ which could be drawn from it. The probability of any particular hypothesis is:

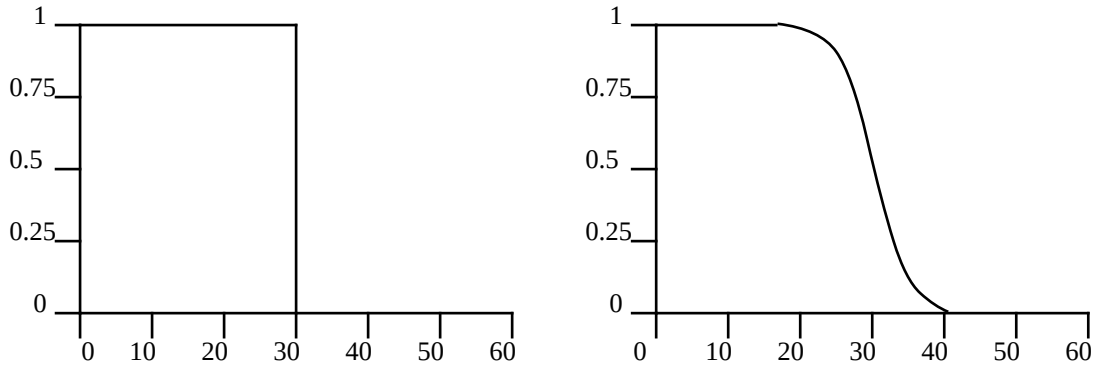
$$pr(H_i|E) = \frac{pr(E|H_i)*pr(H_i)}{\sum_{k=1}^{n}(pr(E|H_k)*pr(H_k))}$$

The problem with this is it makes the assumption that the statistical data on the relationships of the evidence with the hypotheses are known, and that all relationships between evidence and hypotheses $pr(E|H_k)$ are independent of each other. In general this indicates that as the number of factors we are considering increases it becomes increasingly difficult to be able to account for the possible interferences between them.

Fuzzy Logic

Fuzzy logic, developed by the computer scientist Lotfi Zadeh, is an alternative more intuitive approach to reasoning with imprecise knowledge. It does not claim to have the mathematical precision of probability theory, but rather to capture common-sense notions.

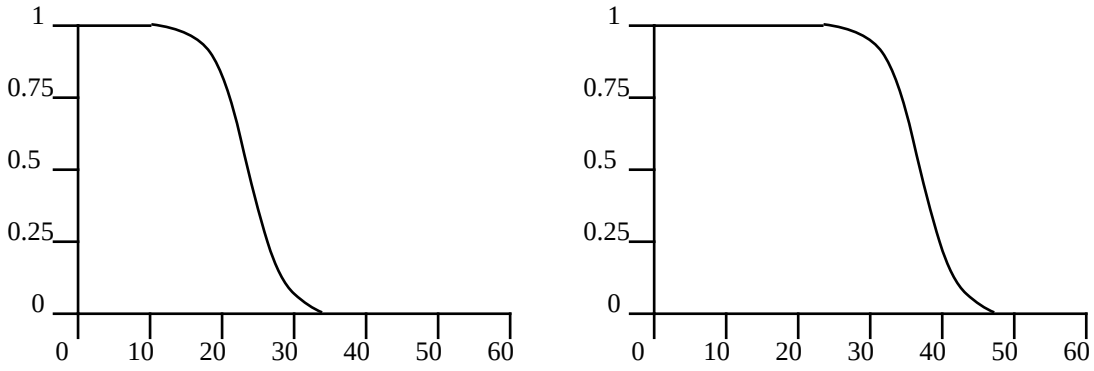
The background to the notion of fuzzy logic is the idea that a truth value may apply partially, and that human reasoning often uses this idea of partial truth. For example, the sentence "Fred is young" might be taken as definitely true if Fred is 19 years old, definitely false is Fred is 50, but applying partially if Fred is 28. In standard logic we might have a sharp cutoff point, say 30, and we would then say for any person p with age below 30, $young(p)$ is true, but for any person q above 30 $young(p)$ is false. In fuzzy logic a numerical value between 0 and 1 is given to express the degree to which a predicate applies. This can be expressed in graphical form, where the graph on the left represents the classical logic approach (the x-axis represents age, the y-axis degree of truthfulness):



This represents a situation where anyone under the age of 20 is reckoned as definitely young, anyone over the age of 40 is reckoned as definitely not young, but the predicate is partial or vague for those between the ages of 20 and 40. More precisely, where a predicate p is defined by the set of elements for each of which p(x) is true, in fuzzy logic a predicate p is defined by a set of pairs ‹x,μ_p(x)›, where μ_p(x) is the degree of membership of x in the set which defines p.

Note the difference between this and the probabilistic reasoning described above. When we said that pr(L_F) was 0.5 we did not mean that Fred was half taking the logic course, we meant that the probability that Fred was taking the course was 0.5. If in our example above the age of Fred is 30, if Fred is represented by f, and the fuzzy predicate young by y, then from the graph above μ_y(f) is 0.5, meaning that Fred is indeed "half young".

Fuzzy logic also captures the idea of qualifiers or *hedges* to predicates in natural language, such as "Fred is very young" or "Fred is fairly young". These might be represented in fuzzy logic by graphs which move the curbed part to the left and to the right of the scale respectively:



The degree of membership of an item x in the union of two fuzzy sets A and B is the maximum of its degree of membership in A or B, while the degree of membership in the intersection is the minimum. So the truth value of P∧Q for two sentences in fuzzy logic P and Q is the minimum of the truth values of P and Q, while the truth value of P∨Q is the maximum of the truth value for P and Q. The truth value for ¬P is the truth value for P subtracted from 1. So negation is dealt with in a similar way to probabilistic logic, but the connectives are not.

As an example, let us consider reasoning about the set of subjects in a hypothetical Philosophy department. This set, called Ω, is the "frame of discernment" (set of all possible values x for an attribute). In our example, this set is:

{Epistemology, Methodology, Aesthetics, Metaphysics, Logic, Politics, Theology}

Suppose we have a predicate Interesting, representing how interesting each subject is. In classical logic we would simply gave to divide the subjects up into those that are interesting and those that are not. If we suppose that Aesthetics, Politics and Theology are interesting, we could represent the predicate Interesting by the set

{Aesthetics, Politics, Theology}

However, the reality is that we will not have a straightforward division into interesting and non-interesting subjects, but rather rate them on a scale of interestingness (note also that, of course, what

one person finds interesting another may not, but for the sake of space we shall not explore this additional factor). Rather than impose an arbitary cut-off point on this scale, fuzzy logic would rate the subjects on a 0 to 1 scale. So perhaps `Aesthetics` has an interest factor of 0.9, `Politics` an interest factor of 0.7, `Theology` an interest factor of 0.8, `Metaphysics` and `Methodology` each an interest factor of 0.5, `Logic` an interest factor of 0.3, and `Epistemology` an interest factor of 0.1. We can represent this by the fuzzy set:

```
{<Epistemology,0.1>, <Methodology,0.5>, <Aesthetics,0.9>, <Metaphysics,0.5>,
                <Logic,0.3>, <Politics,0.7>, <Theology,0.8>}
```

which we shall call `I`. The function μ_I gives the interest factor of any particular subject, so μ_I`(Epistemology)` is 0.1, μ_I`(Methodology)` is 0.5, and so on. The cutoff point for the classical logic version was between 0.5 and 0.7. In fact there is a name for this: if A is a fuzzy set in Ω then the α-cut A_α is defined by

$$A_\alpha = \{\omega \in \Omega \mid \mu_A(\omega) \geq \alpha\}$$

The set we suggested for the classical logic set for `Interesting` above might perhaps be the $I_{0.6}$ cut. Similarly, $I_{0.4}$ is `{Methodology, Aesthetics, Metaphysics, Politics, Theology}` (note that though we are restricting values of μ_I`(x)` to single decimal places for convenience, this is not essential).

Dealing with negation, `Interesting(x)` is μ_I`(x)`, so `¬Interesting(x)` is `1-`μ_I`(x)`. For example, `¬Interesting(Politics)` is `0.3`. Indeed, we can define a set `¬I` representing the fuzzy set of uninteresting things simply by replacing all the μ_I`(x)` in `I` with $1-\mu_I$`(x)`:

```
{<Epistemology,0.9>, <Methodology,0.5>, <Aesthetics,0.1>, <Metaphysics,0.5>,
                <Logic,0.7>, <Politics,0.3>, <Theology,0.2>}
```

Now let us suppose there is another fuzzy set, `H`, represent how hard each subject is:

```
{<Epistemology,0.7>, <Methodology,0.6>, <Aesthetics,0.4>, <Metaphysics,0.9>,
                <Logic,0.5>, <Politics,0.2>, <Theology,0.8>}
```

Applying the rules of fuzzy set theory, the set `H`$\cap$`I` is:

```
{<Epistemology,0.1>, <Methodology,0.5>, <Aesthetics,0.4>, <Metaphysics,0.5>,
                <Logic,0.3>, <Politics,0.2>, <Theology,0.8>}
```

while the set `H`$\cup$`I` is:

```
{<Epistemology,0.7>, <Methodology,0.6>, <Aesthetics,0.9>, <Metaphysics,0.9>,
                <Logic,0.5>, <Politics,0.7>, <Theology,0.8>}
```

So we now have a measure of the degree to which a subject fits into the class of "Hard and interesting things"represented by the set `H`$\cap$`I`, and a measure of the extent to which it falls into the class of "Hard or interesting things" represented by the set `H`$\cup$`I`.

Note that whereas in classical logic, `A`$\wedge$`¬A` is always false, so the set representing `A`$\wedge$`¬A` is always the empty set, this is not so in fuzzy logic. The set `I`$\wedge$`¬I`, representing the class of things which are both interesting and not interesting is:

```
{<Epistemology,0.1>, <Methodology,0.5>, <Aesthetics,0.1>, <Metaphysics,0.5>,
                <Logic,0.3>, <Politics,0.3>, <Theology,0.2>}
```

so it can be seen that the subjects which have the highest degree of membership of this class are those that are in the middle of the interesting–not-interesting scale. Clearly, the maximum degree to which any item can have a membership of the set `A`$\wedge$`¬A` is 0.5.

Most of the expected properties of a boolean algebra apply, for example, de Morgan's rules: `¬(A`$\wedge$`B)` is equal to `¬A`$\vee$`¬B`. We could define a $\rightarrow$ symbol where as in classical logic `A`$\rightarrow$`B` is equivalent to `¬A`$\vee$`B`, so for example, the set representing `Hard`$\rightarrow$`Interesting` is:

```
{<Epistemology,0.3>, <Methodology,0.5>, <Aesthetics,0.9>, <Metaphysics,0.5>,
                <Logic,0.5>, <Politics,0.8>, <Theology,0.8>}
```

which could be seen as a measure to which the various subjects fit the notion that the hard subjects are the interesting one. In fact classical predicate logic could be considered just the special case of fuzzy logic where $\mu(\omega)$ is restricted to be 0 or 1.

Recalling the suggestion that $\forall x A(x)$ might be considered as equivalent to $A(n_1) \wedge A(n_2) \wedge \ldots A(n_m)$ for $n_1 \ldots n_m$ ranging over the objects in the universe, $\forall x S$ for S a sentence in fuzzy logic is simply the minimum μ value for any pair in the set representing S, while similarly since $\exists x A(x)$ was considered as representing $A(n_1) \vee A(n_2) \vee \ldots A(n_m)$, has the value of the maximum μ value for any pair in the set representing S. So $\forall x \mathtt{Interesting}(x)$ has the value 0.1 and $\exists x \mathtt{Interesting}(x)$ has the value 0.9.

Truth Maintenance

Truth Maintenance Systems (TMSs) are methods of keeping track of reasoning that may change with time as new items of knowledge are added. They are by their nature non-monotonic, as they are based on the idea that some "facts" may only be assumptions which are believed until disproved by further knowledge being learnt. Truth Maintenance Systems keep a record showing which items of knowledge are currently believed or disbelieved along with links to other items of knowledge on which this belief or disbelief is based. Because of this, if a new fact is learnt or an assumption changed the ripple effect it has on other beliefs based on the previous assumptions can be managed by tracing along the links. Although the name "Truth Maintenance System" has stuck as it was used to describe early versions of this sort of system, a better name that has been proposed and is also used is *Reason Maintenance System*, as these systems are more about belief and disbelief and managing the reaosn for belief or disbelief than about truthhood or falsehood.

Truth maintenance systems are another form of knowledge representation which is best visualised in terms of graphs. The idea is that each item of knowledge which a system may believe is represented by a node in the graph. These nodes are labelled either *in* (for a node which represents something which is currently believed) or *out* (representing something which is not currently believed). Note the distinction between in and out representing beliefs, and true and false representing absolute values.

There are several forms of truth maintenance system. For illustration we shall start off with a simple one. The links in the graph are directed and represent *justifications*, and are of two sorts, positive and negative. A positive link from node A to node B represents the idea that believing in A is a reason for believing in B, or A is a justification for believing in B. It is equivalent to the logic $A \rightarrow B$. A negative link from A to B represents the idea that believing in A is a reason for not believing in B, the equivalent to logic $A \rightarrow \neg B$.

For example, let us suppose we are representing the reasoning over whether we should put a roof on a pen in which we are going to keep some animal called Fred. We might then reason

"If Fred can fly, the pen needs a roof" or $\mathtt{Fly} \rightarrow \mathtt{Roof}$

"If Fred is a bird, Fred can fly" or $\mathtt{Bird} \rightarrow \mathtt{Fly}$

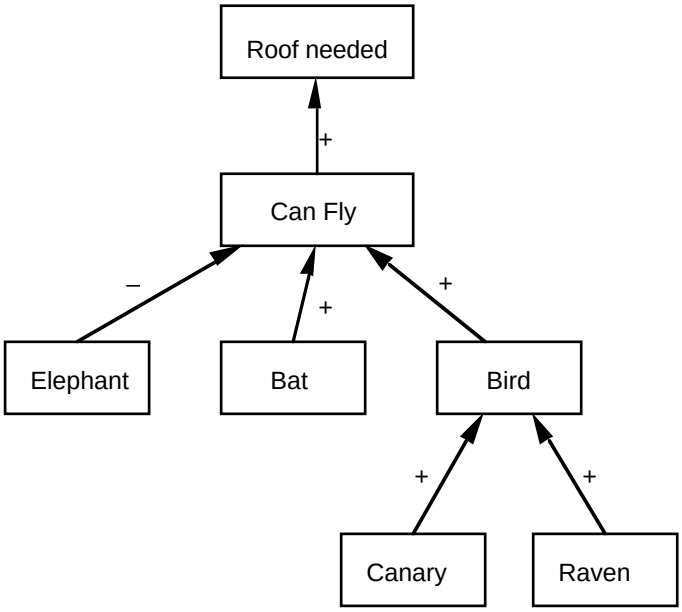
"If Fred is an elephant, Fred cannot fly" or $\mathtt{Elephant} \rightarrow \neg \mathtt{Fly}$.

"If Fred is a bat, Fred can fly" or $\mathtt{Bat} \rightarrow \mathtt{Fly}$

"If Fred is a canary, Fred is a bird" or $\mathtt{Canary} \rightarrow \mathtt{Bird}$.

"If Fred is a raven, Fred is a bird" or $\mathtt{Raven} \rightarrow \mathtt{Bird}$.

This gives us the graph:



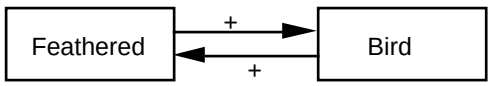
Reasoning works by labelling the leaf nodes (i.e. those without any incoming arcs) as *in* or *out*. This represents our *assumptions*. If we assume Fred is a canary, and Fred is not an Elephant, a Bat or a Raven, we label the node `Canary` as *in*, and the nodes `Elephant`, `Bat` and `Raven` as *out*. Any internal node which has at least one positive arc leading from an *in* node, and no negative arc from an *in* node is then labelled as *in*, any internal node which has at least one negative arc from an *in* node, and no positive arc from an *in* node is labelled as *out*. This process continues until all nodes with at least one arc coming from an *in* node are labelled as *in* or *out*. In the above example, this will result in `Can Fly` and `Roof needed` being labelled as in.

Note that if a node is *out*, it is not used for any justification of any sort, so there is no negation-as-failure effect of assuming that if we can't find a justification for something it must necessarily be false. For example, let us suppose we assume Fred is neither an elephant, a bat, a canary or a raven. This means there are no justifications for the node `Bird`. In Prolog, with negation as failure we would then assume that Fred is not a bird. However in the TMS if we want to assume that Fred is a bird of a sort we don't yet know we would label `Bird` as *in*, and in general any nodes without justifications may be labelled as *in* or *out*, representing further assumptions.

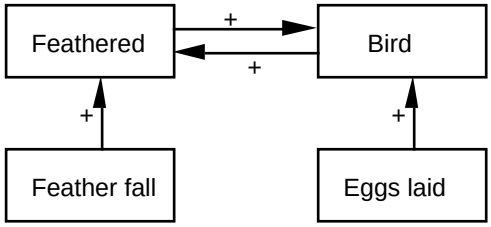
We can change our assumptions as we like, so any node labelled *in* or *out* without justifications can be changed to *out* or *in*. Any such change of assumptions means that other nodes which have arcs leading from these assumptions may have to be changed, and the change will ripple through the system. Suppose in the the above example we changed our assumptions and wanted to assume that Fred is an elephant, and not a bat, canary or raven. We would then label `Elephant` as *in*, and `Bat`, `Canary` and `Raven` as *out*. This would mean that `Can Fly` would have to have its label changed to *out*. `Roof needed`, though, would not necessarily have to have its label changed. It would have no *in* justifications necessitating it being labelled either *in* or *out*, and could be labelled in either way representing an assumption. This makes sense since, though we have ruled out Fred having the ability to fly being an argument for needing a roof, we still need to account for the possibility that there is some other reason not yet known why a roof might be needed.

Now suppose that we started off with our original assumptions that `Canary` is *in*, while `Elephant`, `Bat` and `Raven` are *out*, and then decided to change our minds and assume that Fred is an elephant without also dropping the assumption that Fred is a canary. This would cause `Can Fly` to have both a negative justification *in* and a positive justification *in*. This situation is termed a contradiction: we have made assumptions that lead to a contradictory conclusion, both that Fred can fly and Fred cannot fly. When this happens, the TMS goes through a process known as *dependency-directed backtracking*, which means that it traces down the arcs from *in* nodes leading to the contradictory node to find all the assumptions on which the contradictory justifications depend i.e. all nodes reached by tracing down the links and getting to the point where a node has no further *in* links leading to it. One of the assumptions would then have to be changed; in this case it would note that the assumptions are `Elephant` and `Canary`, and one of these would have to be made *out* to remove the contradiction.

Problems arise in this truth maintenance system if there are circular chains of justifications. For example, we might argue that "Fred has feathers" is a justification for assuming Fred is a bird, while if we know Fred is a bird we can assume Fred has feathers. This would lead to the structure:

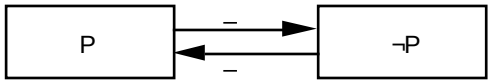


If this occurred as part of the justification graph, on what is said above we could say that neither `Bird` nor `Feathered` is just an assumption since both have justifications supporting them. However we have here what is known popularly as a "circular argument" – if we have no other support for it (that is a positive justification pointing inwards to one of the nodes from a node which is not part of the circle) the entire argument is just an assumption, and the same would apply for a circular argument of three or more nodes. We might decide that if we have observed Fred laying eggs that is a justification for Fred being a bird, or if we have observed a feather fall from Fred that is a justification for Fred being feathered:

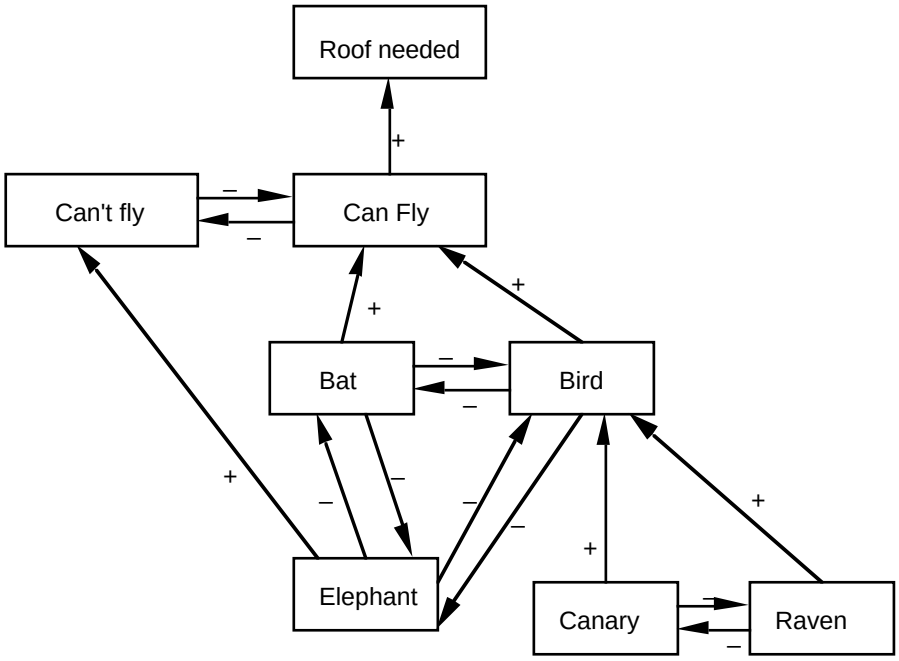


In this case if either `Feather fall` or `Eggs laid` becomes in, it is a support for the whole circular argument.

Note that mutual negative justifications are acceptable, and in fact give the effect of negation:



Here it is not possible for both `P` and `¬P` to be *`in`* without it causing a contradiction. It is possible for both to be *`out`*, representing the case where `P` is unknown. This can be generalised so that if there is a set of options out of which at most one can be true, the node for each option is linked to the rest by mutual negative links. If we want to show that something can only be one out of a bat, a bird and an elephant, and also only one out of a raven and a canary, we could modify our diagram to that below. Note how we have also introduced a separate node for the negation of `Can Fly`. Note that this graph does not rule out the possibility of a bird which is neither a raven nor a canary. In this case, `Bird` would be *`in`*, but both `Raven` and `Canary` would be *`out`*.

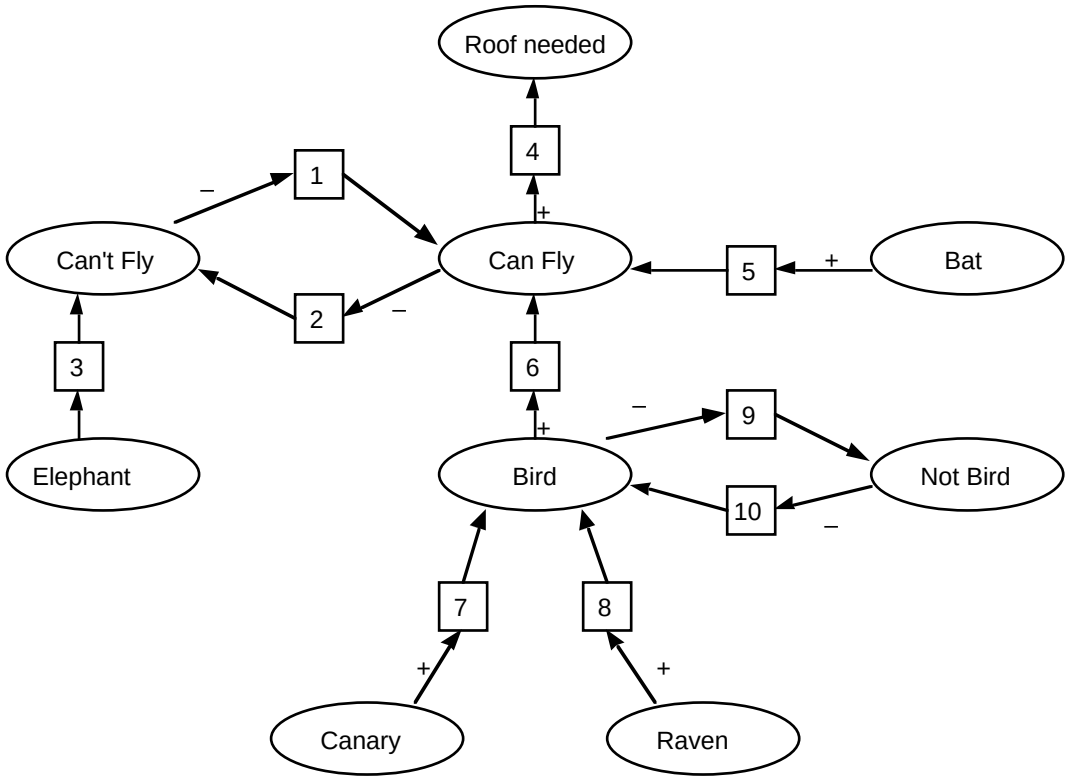


The truth-maintenance system described above was a deliberately simplistic one given to demonstrate some of the ideas behind truth maintenance. It suffers in particular from the fact that the *out*ness of some node cannot be used as a justification for the *in*ness of another. This means, for example, that we cannot have a full negation effect, since we cannot use ¬P becoming *out* as a justification for P coming *in*, rather both P and ¬P stay *out* (representing P is unknown) unless a change of assumption is made.

Doyle's Truth Maintenace System

The original Truth Maintenance System, as proposed by Jon Doyle, had a rather more complex and hence more powerful system of assumptions and justifications. In Doyle's TMS a justification is not a single belief, but a set of beliefs, or rather two sets of beliefs. For a justification to be *valid*, all of its first set of beliefs must be *in*, and all of its second set *out*. For this reason, these two lists are referred to as the *inlist* and *outlist* respectively. A belief is *in* if at least one of its justifications is valid, but unlike the system described above there is no concept of negative justifications forcing a belief to be *out*. A belief is an *assumption* if it relies on a justification with a non-empty outlist, that is it is *in* because of the outness of some other beliefs. An assumption may be justified by its own negation being *out*, similar to the circular belief structure described above. A justification with both inlist and outlist empty is always valid, so any belief it justifies will always be *in*. Such a justification is referred to as a *premise* justification.

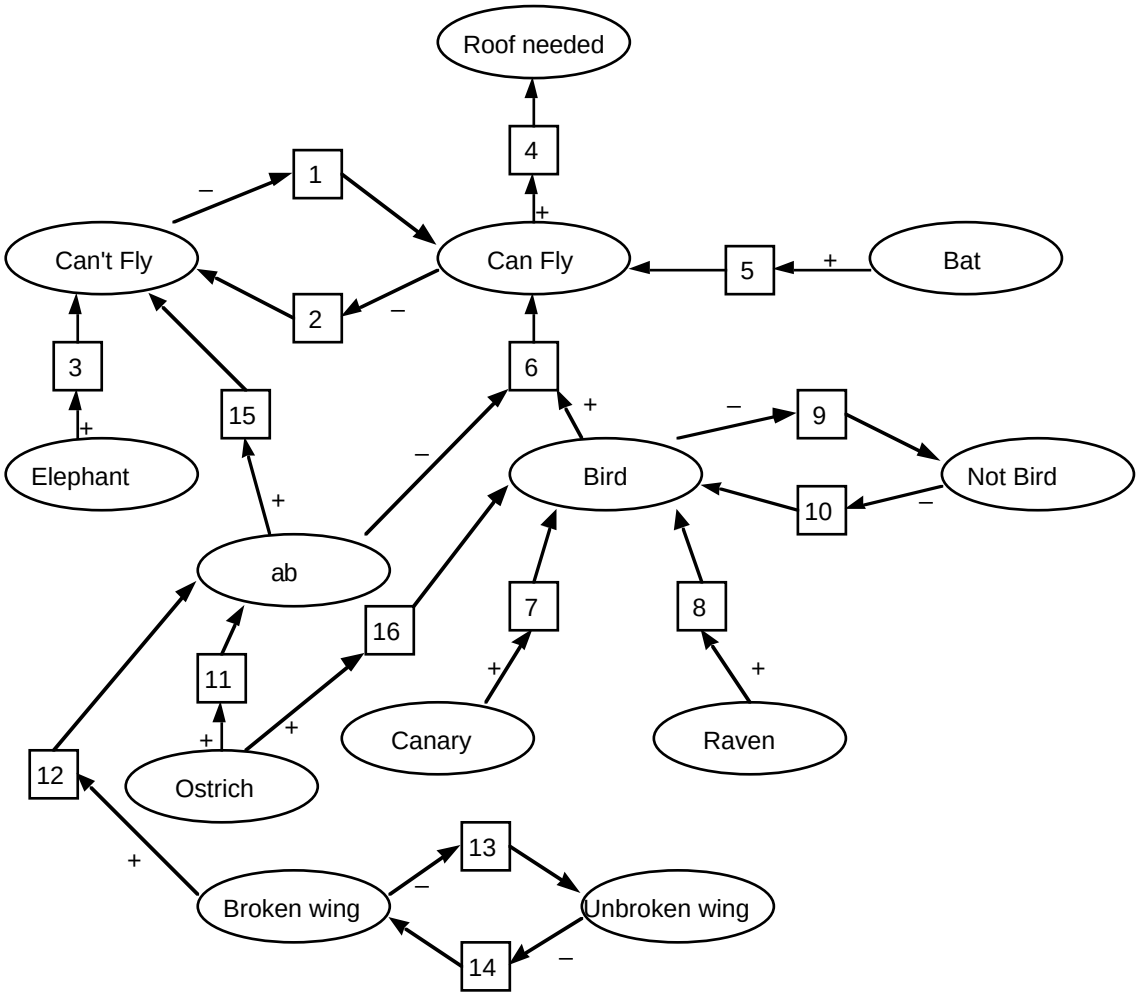
A graphical representation of the set of beliefs and justifications requires a graph with two different sorts of nodes. Such a graph, illustrating the situation described above is shown below. Beliefs are represented by elliptical nodes, and justifications by square nodes. For convenience in discussion, the justifications are numbered.



The resulting graph is similar to that shown before as in this case each justification has a single belief. In this graph, however, a justification supported by a belief being *out* can be used to make another belief *in*, for example making `Can't Fly` *out* makes justification 1 valid, and therefore the belief `Can Fly` becomes *in*. `Can Fly` is justified by any of justifications 1, 5 or 6 being *in*. In practice there would have to be some further beliefs and justifications not shown, for example `Elephant` would have to be justified by the outness of a node `Not Elephant` and vice versa.

Unlike our previous system, this one can correctly deal with default reasoning. Previously we noted that we would like to say that we will assume that a bird can fly providing it is not an abnormal non-flying bird, such as an ostrich, or one whose wing is broken. Adding this to our system creates a

justification for `Can Fly` which depends both on the inness of `Bird` and the outness of `ab` (meaning "Fred is a non-flying bird"). Such a set of beliefs is shown below:

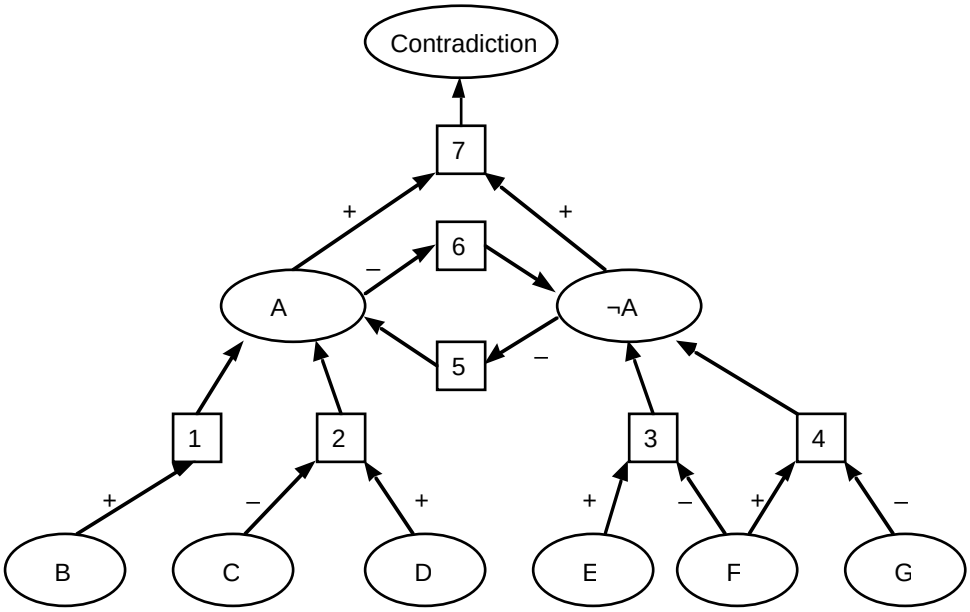


Here, we can start off just by assuming Fred is a bird by having `Bird` *`in`*, supported by `Not Bird` being *`out`*, while if `Canary`, `Raven` and `Ostrich` are also *`out`* we are not making any assumptions as to the sort of bird Fred might be. We will also assume that Fred does not have a broken wing, so `Broken Wing` is *`out`*. This means there is no justification for `ab`, so we can make `ab` *`out`*, making justification 6 valid, and hence `Can Fly` is *`in`*. We have successfully shown that from the assumption that Fred is a bird and without any further assumptions, we can draw the further assumption that Fred can fly.

If, however, we then add the assumption that Fred is an ostrich or Fred has a broken wing by making either `Ostrich` or `Broken Wing` *`in`*, we have a valid justification for `ab`. This means that 15 becomes a valid justification supported by `ab`, while 6 becomes invalid since one of its outlist is *`in`*. `Can't fly` becomes *`in`*, supported by justification 15, while `Can fly` goes *`out`* because 6 is invalid, and 1 too is invalid since 1's outlist now also has an *`in`* in it; we assume that `Bat` is also *`out`*, so the remaining possible justification for `Can Fly`, 5, is invalid.

Note that, correctly, the simple assertion that Fred is an ostrich will lead to justifying both that Fred is a bird, and that Fred cannot fly, while the assertion that Fred is a canary or Fred is a raven will lead to justifying both that Fred is a bird and that Fred can fly.

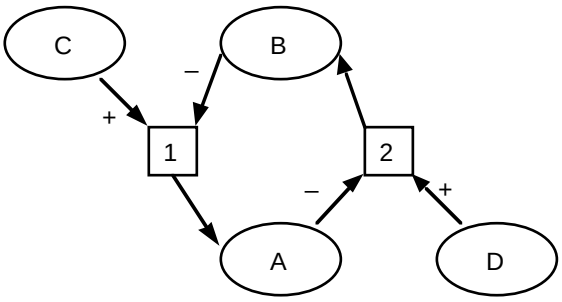
As described so far, this truth maintenance system does not have any way of removing assumptions that lead to contradictions, or indeed of recognising contradictions. Two facts which are contradictory may both be in so long as they both have valid assumptions. The way contradictions re dealt with is to have explicit contradiction nodes. If a contradiction node becomes *`in`*, it is noted and at least one of the assumptions that led to it becoming *`in`* has to be retracted. As an abstract example, consider the following diagram:



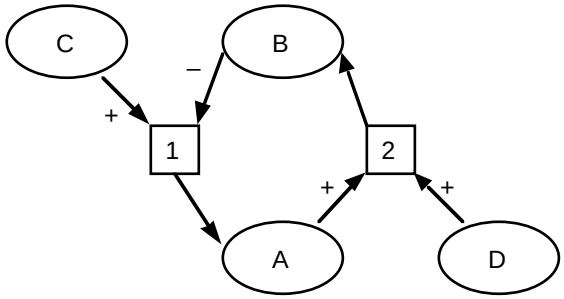
in a situation where B, D and E are assumed *in*, while C, F and G are assumed *out*. This makes 1, 2 and3 valid justifications, so both A and ¬A are *in*, since a belief is *in* so long as any of its justifications are valid. 4 is not a valid justification, since it requires F to be *in* whereas it is *out*. The fact that both 5 and 6 are invalid does not make A or ¬A *out*, since they both have other valid justifications. However, a contradiction node is added with justification 7 which becomes valid only when A and ¬A are *in*. The contradiction being *in* requires sufficient assumptions being changed to make it *out*.

The assumptions that are changed must be sufficient either to make A *out* or to make ¬A *out*. This means that both 1 and 2 must be made invalid or 3 must be made invalid. To make 1 and 2 invalid, the assumption for B must be changed to make it *out* and also C must be made *in* and D *out*. To make ¬A *out*, 3 must be made invalid, but note this must be done in a way that does not make 4 valid. So if F were made *in* that would not be sufficient since it would then make make 4 valid so ¬A would still be supported, though by 4 rather than 3. So to make ¬A *out* either E must be made *out*, or both F and G must be made *in*.

Note the difficulty that cycles of justifications may cause in this Truth Maintenance system. If there is an even number of non-monotonic justifications (i.e. links labelled -) in a cycle, the result is that there is more than one way in which the beliefs in the cycle can be *in* or *out*. In the case below if C and D are *in* and there are no more justifications for A and B, then if A is *in*, 2 is not a valid justification so B is *out* making 1 a valid justification supporting A's *in*ness. If A is *out* then 2 is a valid justification supporting B's *in*ness and making 1 invalid so there is no support for A so it stays *out*. A *in* and B *out*, or A *out* and B *in* are two different sets of beliefs one could infer from assuming C and D are *in*.



If there is an odd number of non-monotonic justifications in a cycle the problem is there may be no possible way of making the beliefs *in* or *out*, for example in:



if C and D are *in* and there are no further justifications for A and B, with A *in*, 2 is a valid justification, making B *in*, but then 1 is invalid making A *out*. But with A *out*, 2 is invalid so B is out, which makes 1 valid putting A *in* and so on.

Further Reading

Books on probability are more likely to be found in the Mathematics section than in the Artificial Intelligence section of the library. Two books which are particularly on the application of probability theory to artificial intelligence are:

F.Bacchus *Representing and Reasoning with Probabilistic Knowledge* MIT Press 1990.

R.E.Neapolitan *Probabilistic Reasoning in Expert Systems* Wiley 1990.

There has been a recent growth in interest in fuzzy logic, and it has found a variety of applications. Three books discussing it are:

A.Kandel *Fuzzy Techniques in Pattern Recognition* Wiley 1982.

C.V.Negoita *Expert Systems and Fuzzy Systems* Benjamin/Cummings 1985.

B.Souč ek *Fuzzy Holographic and Parallel Intelligence* Wiley 1992.

Doyle's original paper on his Truth Maintenance System, as well as many other key papers on issues in non-monotonic reasoning are found in the collection

M.L.Ginsberg (ed) *Readings in Nonmonotonic Reasoning* Morgan Kaufmann 1987.

Another book on non-monotonic reasoning is:

W.Lukaszewicz *Non-monotonic Reasoning* Ellis Horwood 1990.

A book on truth maintenance systems is:

B.Smith and G.Kelleher (eds) *Reason Maintenance Systems and their Applications* Ellis Horwood 1988.