COSC462 Lectures 5 & 6
Forming defeasible beliefs

Willem Labuschagne
University of Otago

Abstract

We explore the use of heuristics in reasoning. More precisely, we look at ways to extend the classical entailment relation to a defeasible entailment relation. Lecture 5 looks at numerical approaches (especially probability), and Lecture 6 at non-numerical approaches (nonmonotonic logic).

1 Introduction

The information an agent obtains about a system is usually limited. A doctor in the emergency room of a hospital must treat a patient according to the visible symptoms without knowing their history and without awaiting the results of laboratory tests. Indeed, even if there were time to wait for test results, it is possible that none of the tests would reveal the cause of the symptoms. Under such circumstances, the agent will simply have to do the best she can, which may mean going beyond her limited definite information and using rules of thumb to form a conjecture such as 'This patient has had a stroke.'

When the agent has to use such a conjecture as a basis for action, then the agent does not merely toy with it as an abstract speculation but adopts it as a belief. Of course, the agent should remain aware that the belief goes beyond her definite information and thus that there is an element of uncertainty about it.

If by *defeasible* beliefs we understand beliefs that are formed on the basis of information that contains an element of uncertainty — *default* information — then there is an important question to be asked: Under what circumstances would it be rational to base action on defeasible beliefs? Short answer: when the risks are not too great.

1.1 Emotions and default information

There is a tradition dating back to Ancient Greece according to which people in everyday life are erratic, emotion-driven, and found wanting when measured against the ideal of cold intellectual rationality. This tradition had an effect on psychology, which for decades paid more attention to negative forms of behaviour than to sanity and happiness. But from the 1950s onward, cognitive psychologists began to interact increasingly with those in artificial intelligence who were attempting to design artificial agents capable of coping in normal everyday situations, e.g. tidying up a room. This interaction gave rise to the broader field known as cognitive science, and one of the insights to emerge from it is that our weaknesses and strengths are two sides of the same coin. An agent is like a well-designed data structure — some operations may be rendered less efficient as the price paid to make other operations more efficient. Our ideas of rationality are changing as we improve our understanding of the design decisions that go into agents. For example, emotions are not unfortunate aberrations that undermine rationality — they are ways of organising responses from a modular brain, and help us to make decisions. What is more, there is an interesting connection between emotions and the adoption of defeasible beliefs.

The psychologist Alice Isen showed that people who are feeling mildly happy make more use of default information (i.e. heuristics) than people who are feeling unhappy. And their use of default information allowed the happy people to reason more efficiently, solving problems in far less time[1]. While Isen's experiments showed that both happy and unhappy people when confronted with a problem-solving situation came to the same eventual decision, the difference was that the happy people got there quickly by taking bold shortcuts while the unhappy people defensively tried to minimise the risk of being wrong.

Why should our emotions have a connection with the way we reason? It appears from the work of cognitive psychologists like Isen and Barbara Fredrickson that positive and negative emotional states evolved in order to mediate between the environment and the agent's momentary thought-action repertoire. When the agent is in a preferred environment, an exploratory attitude is appropriate, whereas a threatening or unsafe environment is best met by a defensive, risk-minimising attitude. Positive affect is the internal reflection of a preferred environment, negative affect is the internal reflection of an unsafe environment. Positive affect is accompanied by a broadened thought-action repertoire — the mind is opened to a wider array of thoughts and actions, which facil-

[1] References are given at the end.

itates exploration, learning, creativity. Negative affect is accompanied by a narrowing of the thought-action repertoire to result in what have been called 'specific action tendencies': fear is linked with an urge to flee, anger with the urge to attack, disgust with the urge to expel, and so on. Clearly, in a life-threatening situation, a narrowed thought-action repertoire promotes quick and decisive action that carries direct and immediate benefit, because the specific action tendencies called forth by negative emotions represent the sort of actions that worked best to save the lives of our ancestors. A positive emotional state allows the mind to entertain alternatives; a negative emotional state suppresses such distractions.

Consider a concrete example — the agent in the control room, monitoring a control panel and trying to decide what the actual state is, or at least whether it is a safe state. The control panel gives limited information, so that there are several candidates for being the actual state of the system. If the system (i.e. environment) appears to be safe, the agent may freely use rules of thumb as basis for her beliefs about the state of the system. But should there be some indication that the environment is unsafe, say an alarm going off or a red light blinking on, then we would expect the agent to react by taking the specific action of pulling the red lever in the corner, thereby shutting down the system and summoning the supervisor from home.

As a second example, consider the doctor in the emergency room. If the social environment is one characterised by sanity and trust, one in which malpractice litigation is not instituted against doctors without strong evidence of carelessness or incompetence, then the doctor will feel free to apply all the rules of thumb she has acquired over the years to decide creatively on her diagnosis and treatment. But if there is a fashion for frivolous lawsuits and the doctor is afraid she may be sued for anything that goes wrong, then it would be both natural and rational for the doctor to play it safe by limiting herself to standard procedure, even if the rules of thumb reflecting her experience suggest that this may not be the best thing for the patient. The point is that proceeding defensively is best for the doctor, since her environment signals danger and rules of thumb don't absolutely guarantee success.

These observations not only help us to understand human behaviour but inform our design of artificial agents. If we design a robot to explore the surface of Mars, it would make sense to equip it with some analog of a positive emotional state, in which it follows up anything that looks interesting, is open to possibilities for new knowledge, and is willing to make a plausible guess that the colour of that rock over there might indicate the presence of a rare mineral, because the price of guessing

wrong is not too heavy. Similarly, it would make sense to equip the robot with some analog of a negative emotional state, so that when its battery gets low, it will become 'worried' and narrow its thought-action repertoire to concentrate on getting back to one of the locations where spare batteries have been stored, without being distracted by otherwise intriguing features of the landscape.

1.2 Kinds of reasoning

In logic, the cautious, risk-minimising type of reasoning is represented by classical entailment. Suppose that $\alpha \models \beta$. An agent who knows that α is the case also knows that β is the case, because β is guaranteed to hold under all circumstances that make α true. The conclusion β is fully justified by the knowledge α, even if the agent doesn't really know α and is just asking 'what if?'. Classical logic is the study of $\models$ and algorithms for simulating it in a sound and complete way. However, classical logic is not the whole of logic. We are now going to move on to a newer part of logic that tries to model common-sense 'plausible' reasoning, in which the agent uses default information (heuristics, rules of thumb, statistical data) to form defeasible conclusions, conclusions that are partly justified but not wholly guaranteed by the agent's knowledge.

The difference between classical logic and the new logic lies in the entailment relation. In classical logic we focus just on the hypothesis-conclusion pairs that belong to $\models$. In a logic that formalises defeasible reasoning, we want to regard a larger variety of hypothesis-conclusion pairs as being acceptable. How should we enlarge $\models$ to a defeasible entailment relation $\mid\!\sim$ containing hypothesis-conclusion pairs in which the hypothesis may support the conclusion partially rather than fully? There are two main approaches: numerical and non-numerical.

Recall that $\alpha \models \beta$ iff $\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$. One can interpret the relation $\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$ numerically by, say, counting the number of models of α that failed to be models of β, and it then becomes possible to arrive at a defeasible consequence relation by relaxing this constraint. For example, if $\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$, then 0 of α's models fail to be models of β, and we could imagine relaxing this constraint by allowing some proportion, say a quarter of α's models, to not satisfy β. The numerical approach involves arithmetic and leads to *probabilities* or *fuzzy sets*.

In contrast, a non-numerical approach might instead relax the constraint $\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$ by working with a subset of $\mathcal{M}(\alpha)$ or with a superset of $\mathcal{M}(\beta)$. In other words, we might form a defeasible consequence relation by requiring either that $X \subseteq \mathcal{M}(\beta)$, where X contains only some of the models of α, or by requiring that $\mathcal{M}(\alpha) \subseteq Y$, where Y contains all the models in $\mathcal{M}(\beta)$ plus more. The non-numerical approach

involves set theory rather than arithmetic, and leads to *nonmonotonic logic*.

In this and the next lecture, we shall briefly outline how probabilistic logic and fuzzy logic work, point out the kind of problem that the numerical approach may suffer from, and describe the kind of nonmonotonic logic based on preferential semantics. It will turn out that preferential semantics defines a defeasible entailment relation $\mid\!\sim$ such that $\alpha \mid\!\sim \beta$ iff $X \subseteq \mathcal{M}(\beta)$, where X is a particular subset of $\mathcal{M}(\alpha)$ consisting of the most preferred models.

2 Probabilistic logic

Default information may be statistical. Statistics is based on probability theory. Let us look at a simple way to use probabilities in logic.

Probabilities arise from some experiment, i.e. from some activity having an observable outcome. Probabilities are based on the idea of proportion — given some collection of equally likely outcomes, the proportion of those that are thus-and-so constitutes the probability of being thus-and-so. Suppose a bucket contains 100 balls of which 70 are white and 30 black. Consider the experiment of drawing a ball 'at random' from the bucket and noting its colour. The probability that the ball is white should be 70 out of 100, or 0.7, since that is the proportion of white balls. Similarly the probability of drawing a black ball is 30/100, or 0.3. So probabilities are fractions ranging from 0 to 1, and the values satisfy certain obvious constraints — for example, whatever the value $\Pr(white)$ associated with drawing a white ball may be, the probability of drawing a ball that is not white must be $1 - \Pr(white)$.

Definition 1 *(**Probabilities**) A **probability space** is a triple $(S, \mathcal{B}, \Pr)$ where S is a nonempty set called the **sample space**, $\mathcal{B}$ is a collection of subsets of S that is a field of subsets of S, and $\Pr : \mathcal{B} \longrightarrow [0,1]$ is a function from $\mathcal{B}$ to the unit interval of real numbers $[0,1]$ such that $\Pr$ is a probability measure on $\mathcal{B}$.*

*Members of the sample space S are called **outcomes**, and the subsets of S that are in $\mathcal{B}$ are called **events**.*

*$\mathcal{B}$ is a **field of subsets** of S iff*

- *$S \in \mathcal{B}$*
- *whenever $X \in \mathcal{B}$, then also the complement $\overline{X} \in \mathcal{B}$*
- *whenever $X, Y \in \mathcal{B}$, then also $X \cup Y \in \mathcal{B}$*
- *whenever $X, Y \in \mathcal{B}$, then also $X \cap Y \in \mathcal{B}$.*

$\Pr$ *is a* ***probability measure*** *on* $\mathcal{B}$ *iff*

- $\Pr(X) \geqslant 0$ *for all* $X \in \mathcal{B}$
- $\Pr(S) = 1$
- *whenever* $X \cap Y = \varnothing$*, then* $\Pr(X \cup Y) = \Pr(X) + \Pr(Y)$.

Example 2 *Let's consider the experiment of flipping a coin. What is the associated probability space? Let us stipulate that the coin eventually comes down, and ignore the case where the coin comes to rest on its edge.*

For our sample space we take $S = \{H, T\}$*, where* H *and* T *denote the outcome of a flip being a head or a tail respectively. Clearly* S *is a sensible sample space, because the outcome of flipping a coin has to be one, and only one, of* H *and* T*.*

Next, $\mathcal{B}$ *must be a collection of subsets of* S *representing the events we want to be able to talk about. Suppose we want to be able to talk about the event that heads is the outcome, the event that tails is the outcome, the event that either heads or tails is the outcome, and the event that neither is the outcome. Then we should have the subsets* $\{H\}$*,* $\{T\}$*,* $\{H, T\} = S$*, and* $\varnothing$ *in* $\mathcal{B}$*. (The event that neither heads nor tails comes up, corresponding to the subset of outcomes* $\varnothing$*, is impossible but we still want to be able to talk about it, for example so that we can say it has zero probability.) Now it is easy to check that* $\mathcal{B} = \{\varnothing, \{H\}, \{T\}, S\}$ *is a field of subsets of* S*.*

Intuitively, the events $\{H\}$ *and* $\{T\}$ *are special because we can regard them as basic building blocks from which all the remaining events in* $\mathcal{B}$ *are built up:* $S = \{H\} \cup \{T\}$ *and* $\varnothing = \{H\} \cap \{T\}$*. We call such building blocks the 'elementary' events. If the elementary events are chosen, as in this case, to correspond exactly to the outcomes in* S*, then* $\mathcal{B}$ *will end up being the collection of all subsets of* S*.*

Finally, we can define a probability measure on $\mathcal{B}$ *by making the assignments* $\Pr(\varnothing) = 0$*,* $\Pr(\{H\}) = \frac{1}{2}$*,* $\Pr(\{T\}) = \frac{1}{2}$*, and* $\Pr(S) = 1$*. It is easy to check that* $\Pr$ *is a probability measure on* $\mathcal{B}$*. We could in fact have defined* $\Pr$ *by merely making the assignments* $\Pr(\{H\}) = \frac{1}{2}$ *and* $\Pr(\{T\}) = \frac{1}{2}$*. The probability* $\Pr(S) = 1$ *now follows from the facts that* $S = \{H\} \cup \{T\}$ *and* $\{H\} \cap \{T\} = \varnothing$*, so that* $\Pr(S) = \Pr(\{H\}) + \Pr(\{T\})$*. That* $\Pr(\varnothing) = 0$ *follows from the fact that* $\{H\} \cup \varnothing = \{H\}$ *so that* $\Pr(\{H\} \cup \varnothing) = \Pr(\{H\})$*.*

Note that this is not the only possible probability measure we could define on the given $\mathcal{B}$*. If we suspected that the coin was not fair but that, say, the likelihood of heads was three times that of tails, then we could define a different measure by stipulating* $\Pr(\{H\}) = \frac{3}{4}$ *and* $\Pr(\{T\}) = \frac{1}{4}$*.*

In the above example, the set $\mathcal{B}$ of events was taken to be the collection of all subsets of the sample space $S = \{H, T\}$. While this is often convenient, it is not obligatory. We may decide that the elementary events should be something different from the sets that contain single outcomes, like $\{H\}$. In the following example, $\mathcal{B}$ consists of all sets built up by union, intersection, and complement from the elementary events W and B, where both W and B are large subsets of the sample space S.

Example 3 *Consider a second experiment, that of drawing a ball from an urn. Suppose we have a big bucket (traditionally called an urn in probability theory) and that this bucket contains 70 white balls, numbered from 1 to 70, and 30 black balls, numbered from 71 to 100. All the balls are the same size and shape and texture and weight, so that if we close our eyes there is no way we can tell the difference between one ball and another. Imagine that we mix the balls thoroughly, without looking at them, and then stick in a hand and blindly (randomly) draw out a ball. Then we look at the ball that was drawn.*

For our sample space we take $S = \{1, 2, \ldots, 100\}$*, since these are the identification numbers of the balls.*

Assume that we are interested only in the colour of the ball we draw, not in which ball it is. For $\mathcal{B}$ *we take the subsets* $\varnothing$*,* B*,* W*, and* S*. Here* $\varnothing$ *correponds to the event that the outcome was neither a black ball nor a white ball (impossible, but we still want to be able to talk about it),* $B = \{71, \ldots, 100\}$ *corresponds to the event that a black ball was drawn,* $W = \{1, \ldots, 70\}$ *corresponds to the event a white ball was drawn, and* S *is the event that we drew either a black ball or a white ball. Is it clear that* $\mathcal{B}$ *is a field of sets?* $B = \overline{W}$*,* $S = B \cup W$*, and* $\varnothing = B \cap W$*.*

A probability measure reflecting the proportion of white balls in the urn could specify that $\Pr(W) = \frac{7}{10}$*, whence it would follow that* $\Pr(B) = \frac{3}{10}$ *because* $B \cup W = S$ *while* $B \cap W = \varnothing$ *so that* $\Pr(B) + \Pr(W) = \Pr(S)$ *and we know that we must have* $\Pr(S) = 1$ *so that* $\Pr(B) = 1 - \Pr(W)$*.*

Exercise 4

1. *Consider the urn with 100 numbered balls, those numbered 1 to 70 being white and the remainder black. Suppose we draw a ball at random and note the number of the ball, not merely the colour. What would be an appropriate probability space for this experiment? (Hint: The elementary events correspond closely to outcomes.)*

2. *Describe the probability space for the experiment of rolling a 6-sided die and looking to see which face is uppermost when it stops rolling. What is the probability of the event that the number showing on the uppermost face of the die is even?*

3. *Consider any experiment. One of the events is* $\varnothing$*. Prove that* $\Pr(\varnothing) = 0$*, always.*

4. *Suppose the Land Transport Safety Authority measures the traffic on an Auckland highway during the hour from 11:00 to 12:00 on 300 consecutive days. The number of cars, and the frequency with which that number was observed, is given by the following table.*

Number of cars	*Frequency*
≤ 1000	*30*
$1001 - 3000$	*45*
$3001 - 5000$	*135*
$5001 - 7000$	*75*
> 7000	*15*

Give the probability space that represents the data gathered by this experiment.

2.1 Probabilities and sentences

How shall we connect the idea of a probability space with logic?

Consider the propositional language L_A with $A = \{p, q\}$ and $S = W_A = \{11, 10, 01, 00\}$. Let the Light-Fan System be the system of interest. We want, in some sensible manner, to associate with every sentence α a numerical value $\Pr(\alpha)$. The key to doing this is to identify sentences with their sets of models.

Let the sample space be $S = W_A = \{11, 10, 01, 00\}$ and let $\mathcal{B}$ be the collection of all 16 subsets of S. The outcomes in S are valuations representing the four possible states of the Light-Fan System and thus the four possible answers to the question 'What is the state of the system?' that might be given by an agent with a god's-eye view of the system (i.e. an agent capable of acquiring complete information by observing the system). An agent with more limited information about the system would need to be able to give a wider variety of answers reflecting the extent to which she was able to narrow down the possibilities by ruling out some of the states. The 16 events in $\mathcal{B}$ provide this variety, from the answer $\{11, 10, 01, 00\}$, when the agent is unable to rule out any states, to the opposite extreme $\varnothing$, when the agent (mistakenly) rules out all the states.

Before we go on to consider probability measures on $\mathcal{B}$, note that we can associate the events in $\mathcal{B}$ with particular object-language sentences: every $X \in \mathcal{B}$ corresponds with any sentence α such that $\mathcal{M}(\alpha) = X$, and since L_A is a finitely generated language we have an algorithm for finding α and can even demand that α be in SDNF. By virtue of the correspondence between events in $\mathcal{B}$ and sentences in L_A, any probability

measure $\Pr : \mathcal{B} \longrightarrow [0,1]$ may be regarded as assigning probabilities to sentences of L_A.

To describe the probability space associated with L_A we need not only a sample space S and a set of events $\mathcal{B}$ but also a probability measure Pr. How do we get a probability measure on sets of valuations?

It is enough to decide, somehow, what probabilities we want to assign to the outcomes 11, 10, 01, and 00, because the set $\mathcal{B}$ of events is the collection of all sets that can be built by union, intersection and complement from the elementary events $\{11\}$, $\{10\}$, $\{01\}$, and $\{00\}$. We might, for instance, have reason to believe that the outcomes are equally likely — perhaps because we have observed the system for 113 years and found that it spends an equal amount of time in each of the four states. In this case we would give each outcome the probability $\frac{1}{4}$. (In terms of sentences, we may think of the probabilities as being assigned to the state descriptions $p \wedge q$, $p \wedge \neg q$, $\neg p \wedge q$, and $\neg p \wedge \neg q$.)

After probabilities have been assigned to the elementary events, every event automatically gets its own probability and so every sentence of L_A gets its own probability. Since each elementary event consists of a single outcome, to find the probability of a sentence α, go to $\mathcal{M}(\alpha)$ and add up the probabilities of the outcomes (states) in the set $\mathcal{M}(\alpha)$. (Or if you like, think of it instead as going to the SDNF of α and adding up the probabilities of the state descriptions in it.)

So, for example, the contradiction $p \wedge \neg p$ has probability 0 because $\mathcal{M}(p \wedge \neg p) = \emptyset$. (Or if you prefer, the contradiction has no state descriptions in its SDNF, and so the sum of the probabilities of the state descriptions in its SDNF is 0.)

We call the probability measure constructed for the sentences of L_A in this way an *initial* probability measure for L_A. The construction of such an initial probability space can be generalised to any language L_A having any ontology (S, V), with the reservation that when A becomes infinite we put into $\mathcal{B}$ only the sets X of states for which we can find sentences α such that $\mathcal{M}(\alpha) = X$. Thus for finitely generated languages, we usually take $\mathcal{B}$ to be the collection of all subsets of the sample space S, but for infinitely generated languages, we take $\mathcal{B}$ to consist only of subsets that can be described in the object language, and leave out the ineffable subsets of S. Clearly it is much simpler to work with finitely generated languages.

Exercise 5 *1. Consider the language with $A = \{p, q\}$. Suppose that, as above, the four outcomes in $W_A = \{11, 10, 01, 00\}$ are equally likely. Work out the initial probability for each of the following sentences:*

- $q \vee \neg q$
- $p \vee q$
- $p \vee \neg q$
- $p \rightarrow q$
- $p \rightarrow \neg q$
- p
- q
- $p \leftrightarrow q$
- $\neg(p \leftrightarrow q)$
- $\neg p$
- $\neg q$
- $p \wedge q$
- $p \wedge \neg q$
- $\neg p \wedge q$
- $\neg p \wedge \neg q$
- $q \wedge \neg q$

2. *Again consider the language with* $A = \{p, q\}$. *Assume we again take* $\mathcal{B}$ *to be the collection of all sets that can be built by union, intersection, and complement from the elementary events* $\{11\}$, $\{10\}$, $\{01\}$, *and* $\{00\}$.

   Having observed the system for a long time, we know that 11 *is the state* $\frac{1}{3}$ *of the time,* 01 *is the state* $\frac{1}{3}$ *of the time, and* 00 *is the state* $\frac{1}{3}$ *of the time, while state* 10 *never arises. For each of the sixteen sentences above, give the initial probability.*

3. *Consider the 3 Card System. An honest dealer shuffles a pack of three cards coloured red, green and blue, and then deals the top card to player 1, the next card to player 2, and the last card to player 3. By a state of the system we understand a deal. Suppose the knowledge representation language had 9 atoms* r_1, r_2, r_3, g_1, g_2, g_3, b_1, b_2, b_3 *where* r_1 *stands for 'Player 1 has the red card' and so on.*

   - *Describe the six valuations that represent states (i.e. deals). (You may write, say,* rgb *for the valuation making the atoms* r_1, g_2, *and* b_3 *true and all other atoms false.)*

- *Assume we take the set S of deals as the sample space, and take the set $\mathcal{B}$ of events to consist of all subsets of S. Write down the probability of each outcome (i.e. of each elementary event $\{s\}$ where s is a deal).*
- *What is the probability that a deal gives player 2 the red card?*

4. *In exercise 3 above we encountered an interesting sort of system, in which the obvious knowledge representation language has more valuations than are needed for representing the physically realisable states of the system. This happens quite often in knowledge representation. Let's examine another example.*

   Consider the experiment of throwing a 6-sided die and looking to see which face lies uppermost. As knowledge representation language take L_A where $A = \{p_1, \ldots, p_6\}$ and think of p_i as saying 'The number i is uppermost.'

   - *Describe the six valuations that represent outcomes of the experiment (i.e. states of the system).*
   - *Assuming that each outcome is equally likely, what is the probability that the number showing is even? What is the probability that it is > 3? What is the probability that it is ≤ 2 or ≥ 5?*

5. *Assume that A is finite. Let α and $\beta \in L_A$. Prove that if $\alpha \models \beta$ then $\Pr(\alpha) \leq \Pr(\beta)$.*

2.2 Conditional probabilities

Suppose we have built the initial probability measure for a language L_A having ontology (S, V). Can we now replace the classical entailment relation $\models$ by a useful defeasible entailment relation $\mid\!\sim$? Indeed we can, by means of *conditional* probabilities. For what does it mean, probabilistically speaking, if $\alpha \models \beta$? To verify that $\alpha \models \beta$, we go to the subset $\mathcal{M}(\alpha)$ of S and check that all of the models in this subset satisfy β. In other words, we check that the proportion of models in $\mathcal{M}(\alpha)$ that satisfy β is 100%, or 100/100, or 1. Loosely speaking, we check that if α *conditions* our sample space by cutting it down to $\mathcal{M}(\alpha)$, then the new 'conditional' probability of β is 1, although this idea is not yet precise, because we have not said how changing the sample space changes the probability measure. Assuming there is a sensible way to get the new probability measure, it becomes obvious how to define a defeasible entailment relation $\mid\!\sim$. Choose some number t less than 1, and define that

$\alpha \mid\!\sim \beta$ if the new probability of β relative to the new smaller sample space $\mathcal{M}(\alpha)$ is at least t.

To begin with, we have an initial probability measure for L_A, based on the sample space S. If we cut the sample space down from S to $\mathcal{M}(\alpha)$, then the probability space needs to change.

Definition 6 ***(Conditional probabilities)*** *Suppose we have a probability space $(S, \mathcal{B}, \Pr)$ and let some event $E \in \mathcal{B}$ be given, where $E \neq \varnothing$. The conditional probability space relative to E is $(E, \mathcal{B}', \Pr')$ where $\mathcal{B}' = \{X \cap E \mid X \in \mathcal{B}\}$ and where, for every $Y \in \mathcal{B}'$, $\Pr'(Y) = \frac{\Pr(Y)}{\Pr(E)}$.*

Thus the conditional probability space restricts the sample space to an event E, and uses the old probability measure to define the new by, in effect, looking at the relative proportion of Y-things inside E. If the event E is the set of models $\mathcal{M}(\alpha)$, then for every sentence β the 'conditioned' set of models we are going to look at will be $\mathcal{M}(\alpha) \cap \mathcal{M}(\beta)$, in other words we look only at models of β that live inside the new space $\mathcal{M}(\alpha)$.

For example, consider the initial probability space for L_A with $A = \{p, q\}$, and suppose we want to take $\alpha = p$ and restrict consideration to $\mathcal{M}(p) = \{11, 10\}$. To build the new conditional probability space, we take as events all the subsets of our new sample space $\mathcal{M}(p) = \{11, 10\}$, and on this set $\mathcal{B}'$ of events we define the probability measure $\Pr'$ which assigns to the two elementary events that correspond to outcomes in $\mathcal{M}(p)$ the probabilities

- $\Pr'(\{11\}) = \frac{1}{4}/\frac{1}{2} = \frac{1}{2}$ where I have used the fact that the original probabilities were $\Pr(\{11\}) = \frac{1}{4}$ and $\Pr(\{11, 10\}) = \frac{1}{2}$
- $\Pr'(\{10\}) = \frac{1}{2}$ similarly.

There are only two other events in $\mathcal{B}'$ and their probabilities may be calculated either from the formula in the definition or from the properties of probability measures: $\Pr'(\{11, 10\}) = 1$ and $\Pr'(\varnothing) = 0$. In general, we may choose either to calculate conditional probabilities by using the formula or to exploit the additivity of probabilities to derive the conditional probability from previously calculated conditional probabilities.

To illustrate, the conditional probability of β relative to α may be found either by calculating $\Pr(\mathcal{M}(\alpha) \cap \mathcal{M}(\beta)) = c$ and $\Pr(\mathcal{M}(\alpha)) = d$ and then $\frac{c}{d}$, or else by adding up the conditional probabilities of the outcomes in $\mathcal{M}(\alpha)$ that satisfy β, if we already know these. Using either method, the conditional probability of, say, $p \wedge \neg q$ relative to $\alpha = p$ is $\frac{1}{2}$,

where you will recall that $\mathcal{M}(p \wedge \neg q) = \{10\}$ and we previously worked out $\Pr'(\{10\})$. Similarly, the conditional probability of $\neg p$ relative to p is now 0, because in the new sample space $\mathcal{M}(\alpha) = \{11, 10\}$, the sentence $\neg p$ has no models.

Notation 7 *Suppose we have a probability space $(S, \mathcal{B}, \Pr)$ and a conditional probability space relative to some event E, namely $(E, \mathcal{B}', \Pr')$. When there is no danger of ambiguity, we may write $\Pr(Y \mid E)$ for $\Pr'(Y)$. This notation has the virtue of reminding us that the conditional probability is relative to event E, but the danger of letting us forget that we are working with a new probability measure, not the old.*

In a useful extension of this notation, we may write for any event $X \in \mathcal{B}$ that the conditional probability of X relative to some event E is $\Pr(X \mid E) = \frac{\Pr(X \cap E)}{\Pr(E)}$.

Once we have the idea of getting conditional probability measures out of the initial probability measure, we are practically there. All that remains to be done, in order to build a defeasible entailment relation based on probabilities, is to choose a threshold value in the unit interval $[0, 1]$, say $t = \frac{1}{2}$. Now we may define $\mid\!\sim$ by requiring that $\alpha \mid\!\sim \beta$ iff $\Pr(\beta \mid \alpha) \geqslant \frac{1}{2}$. Thus for example $p \mid\!\sim p \wedge \neg q$, whereas $p \not\models p \wedge \neg q$, showing that the defeasible consequence relation $\mid\!\sim$ is different from the classical $\models$. In fact, $\mid\!\sim$ is an enlargement of $\models$, because whenever $\alpha \models \beta$, then $\alpha \mid\!\sim \beta$. This is easy to show in a manner unaffected by the choice of initial probability and unaffected by the choice of threshold: if $\alpha \models \beta$ then β is true at every model of α and so the conditional probability of β relative to α is 1. (Well, almost unaffected by the initial probability — bear in mind that we can do this only for α such that $\Pr(\alpha) > 0$.) Let's formalise this.

Definition 8 ***(Probabilistic defeasible entailment)*** *Let $t \in [0, 1]$. For any sentence α such that $\Pr(\alpha) > 0$, we say that α defeasibly entails β (probabilistically), written $\alpha \mid\!\sim \beta$, iff $\Pr(\beta \mid \alpha) \geqslant t$.*

Conditional probabilities are useful for more than defining defeasible entailment relations. They are convenient for defining the product rule and also the importance notion of independence.

Example 9 *Consider an urn containing 70 white balls (numbered 1 to 70) and 30 black balls (numbered 71 to 100). Let the experiment consist of drawing a ball at random, noting only its colour, replacing the ball in the urn, stirring the balls thoroughly, and then randomly drawing a ball for the second time and noting only its colour. Assume we want to remember which colour we saw first and which second.*

Previously, the sample space for a single draw was $S = \{1, \ldots, 100\}$. The new sample space is $S' = S \times S$, the set of ordered pairs (i, j) where i and j are the numbers of the first and second ball drawn, respectively. But since we are interested only in colours, not the identity of the balls, what should we take our set of events $\mathcal{B}$ to be?

Let's take our elementary events to be

- $\{WW\}$*, the event that the first ball is white and the second ball is white*
- $\{WB\}$*, the event that the first ball is white and the second is black*
- $\{BW\}$*, the event that the first ball is black and the second white, and*
- $\{BB\}$*, the event that both the first and the secopnd balls are black.*

Actually, $\{WW\}$ is just a convenient name for a particular subset of S', namely all those pairs (i, j) such that $i \leq 70$ and $j \leq 70$. Similarly $\{WB\}$ is our name for the set of (i, j) such that $i \leq 70$ and $71 \leq j \leq 100$, and so on.

The remaining events in $\mathcal{B}$ are built up from the elementary events by forming unions, intersections, and complements. If we can work out the probabilities of these elementary events, we will be able to calculate the probabilities of all the other events. But it is not obvious what the probabilities of the elementary events are. One way to work them out is with the help of the product rule.

We know that for any event $X \in \mathcal{B}$, the conditional probability of X relative to an event E is $\Pr(X \mid E) = \frac{\Pr(X \cap E)}{\Pr(E)}$. Multiply both sides by $\Pr(E)$ to get:

Definition 10 *(Product Rule)* $\Pr(X \cap E) = \Pr(E) \cdot \Pr(X \mid E)$.

Example 11 *Going back to our urn, what do we know? Well, we know that there are 70 white balls and 30 black balls at the start. Let E be the event that the first ball is white, so that $E = \{WW, WB\}$. Clearly $\Pr(E) = 0.7$. Similarly, if F is the event that the first ball is black, then $F = \{BB, BW\}$ and $\Pr(F) = 0.3$.*

These are not the only easy probabilities. Recall that balls are replaced after the first draw, so that before the second draw the urn again contains 70 white and 30 black balls. Let X be the event that the second ball is black, so that $X = \{WB, BB\}$. Clearly $\Pr(X) = 0.3$. Let Y be the event that the second ball is white, then $Y = \{WW, BW\}$ and $\Pr(Y) = 0.7$.

Now consider the elementary event $\{WB\}$. *Note that* $\{WB\} = X \cap E$. *By the product rule, if we know* $\Pr(E)$ *and* $\Pr(X \mid E)$ *then we can simply multiply them together to get* $\Pr(X \cap E)$. *We already have* $\Pr(E)$. *It is just as easy to see that* $\Pr(X \mid E) = 0.3$. *After all, the first ball is replaced after being drawn and the urn restored to its full tally of 70 white and 30 black balls before the second ball is drawn, so that* $\Pr(X \mid E) = \Pr(X)$. *Thus by the product rule* $\Pr(X \cap E) = 0.21$ *and this is also the probability of the outcome* WB.

In a similar fashion we find that

- $\Pr(\{WW\}) = 0.49$, *because* $\{WW\} = Y \cap E$ *and* $\Pr(Y \mid E) = \Pr(Y)$
- $\Pr(\{BW\}) = 0.21$, *because* $\{BW\} = Y \cap F$ *and* $\Pr(Y \mid F) = \Pr(Y)$
- $\Pr(\{BB\}) = 0.09$, *because* $\{BB\} = X \cap F$ *and* $\Pr(X \mid F) = \Pr(X)$.

In the example of drawing two balls in succession from the urn, with replacement of the first ball before the second is drawn, it is intuitively clear that events such as

- E : "the first ball is white"
- X : "the second ball is black"

are independent of one another. After all, the effect of the first draw is wiped out by replacing the ball.

On the other hand, suppose we change the experiment so that we draw two balls in succession *without* replacement. In other words, we draw the first ball, note its colour, throw it away instead of returning it to the urn, and then draw a second ball and note its colour. Now the events E and X are no longer independent — the effect of the first draw is to change the proportion of black balls in the urn, making the chance of drawing a black ball greater.

Definition 12 *(**Independence**) Let* E *and* X *be events. Then* E *and* X *are* ***independent*** *iff* $\Pr(X \mid E) = \Pr(X)$ *and* $\Pr(E \mid X) = \Pr(E)$.

Corollary 13 E *and* X *are independent iff* $\Pr(X \cap E) = \Pr(E) \cdot \Pr(X)$.

Example 14 *Consider the urn with 70 white and 30 black balls. This time our experiment consists of drawing a ball, noting its colour, and then, without replacing the first ball, drawing a second ball and noting its colour. What probability space is associated with this experiment?*

As before, the sample space is $S' = S \times S$ *and the events in* $\mathcal{B}$ *are all the subsets of* S' *that can be built from* $\{WW\}$, $\{WB\}$, $\{BW\}$, *and* $\{BB\}$ *by union, intersection and complement. To fully specify the probability measure, we need to determine the probabilities of the elementary events.*

The probabilities of some events are easy to determine. Let E *be the event that the first ball is white, i.e.* $E = \{WW, WB\}$. *Since there are 70 white and 30 black balls when the first ball is drawn,* $\Pr(E) = 0.7$. *Let* F *be the event that the first ball is black, then* $F = \{BB, BW\}$ *and* $\Pr(F) = 0.3$ *because there are 30 black balls in the urn when the first ball is drawn.*

Let X *be the event that the second ball is black, i.e.* $X = \{WB, BB\}$. *The probability of* X *is not quite so easy to determine. The likelihood of the second ball being black is affected by whether we first drew a white ball or a black ball. On the other hand, the conditional probabilities are straightforward.*

Suppose the first ball is white, i.e. the event E *occurs. Now there remain 69 white balls and 30 black balls in the urn, so that the probability of the second ball being black is* $\Pr(X \mid E) = \frac{30}{99} = 0.303$ *(roughly). Or suppose the first ball was black, i.e. that* F *occurred. Then there would remain 70 white balls and 29 black balls in the urn, so that the probability of the second ball being black would be* $\Pr(X \mid F) = \frac{29}{100} = 0.290$.

By the product rule, $\Pr(X \cap E) = \Pr(E) \cdot \Pr(X \mid E) = \frac{7}{10} \cdot \frac{30}{99} = 0.212$ *(roughly). But* $X \cap E = \{WB\}$ *so we know that elementary event* $\{WB\}$ *has probability* 0.212. *Similarly* $\Pr(X \cap F) = \Pr(F) \cdot \Pr(X \mid F) = \frac{3}{10} \cdot \frac{29}{100} = 0.087$. *Since* $X \cap F = \{BB\}$, *we know that* $\{BB\}$ *has probability* 0.087.

Since $X = \{WB, BB\}$ *and we know the probabilities of the individual outcomes in* X, *we get that* $\Pr(X) = \Pr(\{WB\}) + \Pr(\{BB\}) = 0.212 + 0.087 = 0.299$. *We see that events* X *and* E *are not independent, since* $\Pr(X) \neq \Pr(X \mid E)$, *and similarly for* X *and* F.

While independence is an important notion, it is relevant for us only because it can help to build the probability space needed for a defeasible entailment relation.

Exercise 15 *1. Consider the language with* $A = \{p, q\}$ *and ontology* $S = W_A$. *Suppose the four outcomes in* $S = \{11, 10, 01, 00\}$ *are*

equally likely as in the first of the previous set of exercises. Keep in mind the 16 sentences listed there.

- *Take $\alpha = p \vee q$ and threshold $t = \frac{1}{2}$. For each of the 16 sentences, work out the conditional probability of each sentence β and hence decide whether $\alpha \mathrel{|\!\sim} \beta$.*
- *Take $\alpha = p$ and threshold $t = \frac{9}{10}$. For each of the 16 sentences, work out the conditional probability of each sentence β and hence decide whether $\alpha \mathrel{|\!\sim} \beta$.*

2. *Again consider the language with $A = \{p, q\}$ and $S = W_A$. Suppose that 11 is the actual state $\frac{1}{3}$ of the time, 01 is it $\frac{1}{3}$ of the time, and 00 is it $\frac{1}{3}$ of the time, while state 10 never arises. By calculating conditional probabilities and setting the threshold at $\frac{3}{4}$, determine whether $p \mathrel{|\!\sim} q$ and whether $\neg p \mathrel{|\!\sim} \neg q$.*

3. *Consider the 3 Card System in which each of three players get one of three cards coloured red, green or blue. Suppose the knowledge representation language had 9 atoms r_1, r_2, r_3, g_1, g_2, g_3, b_1, b_2, b_3 where r_1 stands for 'Player 1 has the red card' and so on.*

   The cards are dealt to players by a new method. The dealer rolls a 6-sided die, in which each side is equally likely to be uppermost. If the uppermost face shows a 1, 2, 3, or 4, the dealer gives the red card to player 1. If the uppermost face is 5, player 1 gets the green card, and if it is 6 he gets the blue card. Next the dealer shuffles the remaining two cards and gives one to player 2 and the last to player 3.

   - *Suppose we take deals as outcomes and take the set $\mathcal{B}$ of events to be the collection of all subsets of the set S of outcomes. Thus the elementary events are the subsets of form $\{s\}$ where s is a deal. Write down the probability of each elementary event.*
   - *What is the probability that a deal gives player 2 the red card?*
   - *Suppose the deal has not given player 2 the red card. Player 2 knows this. Player 2 wonders whether player 1 has the red card. Setting the threshold at $\frac{3}{4}$, and calculating conditional probabilities, work out whether $\neg r_2 \mathrel{|\!\sim} r_1$.*

4. *Suppose we have a probability space $(S, \mathcal{B}, \Pr)$ for L_A. Prove that there are no α and β such that $\Pr(\alpha \wedge \beta) > \Pr(\alpha)$, in other words that $\Pr(\alpha \wedge \beta)$ is always $\leq \Pr(\alpha)$.*

3 Fuzzy logic

A fuzzy set is a function from some set X into the unit interval $[0, 1]$. Thus every probability measure is a fuzzy set, but of course a fuzzy set need not satisfy the constraints that characterise probability measures. Functions into the unit interval are familiar, but it was Lotfi Zadeh who gave them a special name and suggested some interesting applications in 1965. The original definition states: "A fuzzy set (class) A in X (a universe of discourse) is characterised by a membership function $f_A(x)$ which associates with each point in X a real number in the interval $[0, 1]$, with the value of $f_A(x)$ the 'grade (degree) of membership' of x in A". Although this definition obscurely speaks of a metaphysical entity A as well as the set-theoretic object f_A, we may identify A with the function f_A, which is all there really is.

The basic idea is that many categories are not conveniently modelled merely by ordinary 'crisp' sets. For example, in the universe X of people, one way to separate out a category of young people is to decide on a cut-off age, say 29, and to take the set of all people with ages not exceeding 29 to constitute the category of young people. However, this distorts the gradual transition from being unquestionably young at age, say, 8 to being fairly young at age 18, youngish at 28, and not far from young at age 30. Instead of a gradual movement from typical members of the category to less typical members that cling to the edge of belonging, all members of the crisp collection are equally valid members and there is a sudden sharp transition from being a young 29 year old to being a no-longer-young 30 year old. Using a fuzzy set is one way in which to introduce gradual movement and avoid the sharp transition. This might be achieved as follows.

Let the fuzzy set *young* have a membership function assigning to the integers from 1 to 100 non-increasing values in the interval $[0, 1]$, say such that the integers (ages) up to 10 are assigned the value 1, the integers from 11 to 20 are assigned the value $\frac{3}{4}$, those from 21 to 25 the value $\frac{2}{3}$, those from 26 to 29 the value $\frac{1}{2}$, those from 30 to 35 the value $\frac{5}{12}$, those from 36 to 40 the value $\frac{1}{3}$, and those above 40 the value 0. Admittedly these choices are somewhat arbitrary, but at least we have achieved our goal — there is no longer a skin sharply dividing the inside of the category from the outside.

Now fuzzy logic becomes possible. Consider a statement such as 'Ali is young'. We can express this by some atom, say p. Usually, a valuation would assign to the atom a truth value that is either 1 or 0, which we may interpret (crisply) as saying either that Ali fully belongs to the category of young people or that Ali completely fails to belong to the category of young people. In fuzzy logic we would take the truth value of

p to be given by the grade of membership of Ali in the fuzzy set *young*. If, for instance, Ali's age is 27 and the fuzzy set *young* is defined as above, then the truth value of p would be $\frac{1}{2}$. We thus arrive at a many-valued logic having infinitely many truth values from 0 (absolutely false) to 1 (absolutely true). (Historical point: although we used fuzzy sets, invented by Zadeh in 1965, to arrive at this infinite-valued logic, the logic itself was first studied by Lukasiewicz in the 1920s.)

Many-valued logics are respectable things. But there are many of them. They differ firstly in the set of truth values chosen. Having fixed the set of truth values, say $[0, 1]$, they may differ in the way that truth values are percolated upward from the atoms. For example, we may be happy to decree that if the truth value $val(\beta)$ is already known then $val(\neg\beta) = 1 - val(\beta)$, or that if $val(\beta)$ and $val(\gamma)$ are both known then $val(\beta \vee \gamma) = \max(val(\beta), val(\gamma))$ and $val(\beta \wedge \gamma) = \min(val(\beta), val(\gamma))$. But there is no general agreement on how the truth values of conditional sentences are to be calculated:

- $val(\beta \rightarrow \gamma) = \min(1, 1 - val(\beta) + val(\gamma))$ is a proposal due to Lukasiewicz and adopted by Zadeh. But notice that if $val(\beta) = \frac{1}{2}$ and $val(\gamma) = 0$, then $val(\beta \rightarrow \gamma) = \frac{1}{2}$. This seems wildly counter-intuitive.

- $val(\beta \rightarrow \gamma) = 1$ iff $val(\beta) \leqslant val(\gamma)$, else 0, is suggested by Lakoff in an influential early paper. But notice that conditional sentences now cannot have intermediate truth values. This seems very strange.

- $val(\beta \rightarrow \gamma) = \max(1 - val(\beta), val(\gamma))$ is suggested by Gaines. This reduces to the usual definition if the only truth values are 0 and 1. But it means that $val(\beta \rightarrow \gamma) = 1$ only if $val(\beta) = 0$ or $val(\gamma) = 1$. This conflicts with the intuition apparent in each of the previous two definitions, namely that if $val(\beta) \leqslant val(\gamma)$ then $val(\beta \rightarrow \gamma)$ should be 1. And it means that sentences of the form $\beta \rightarrow \beta$ need no longer be tautologies (just let $val(\beta) = \frac{1}{2}$).

- $val(\beta \rightarrow \gamma) = 1$ if $val(\beta) \leqslant val(\gamma)$, else $val(\gamma)$, is an alternative also suggested by Gaines. This idea has peculiarities of its own. A small change in the truth value of β or γ can produce a large change in the truth value of $\beta \rightarrow \gamma$. For instance, if $val(\beta) = \frac{1}{10}$ and $val(\gamma) = 0$ then $val(\beta \rightarrow \gamma) = 0$, but if $val(\beta) = val(\gamma) = 0$ then suddenly $val(\beta \rightarrow \gamma) = 1$. Another oddity is that it is impossible to have $val(\beta) \leqslant val(\beta \rightarrow \gamma) < 1$.

- $val(\beta \rightarrow \gamma) = \min(1, val(\gamma)/val(\beta))$ was suggested by Goguen. He felt that the validity of a chain of nearly valid deductions should decrease as the length of the chain increases.

Just as there is no general agreement on how to calculate the truth values of conditional sentences, so there is no general agreement on how to define entailment. One obvious possibility is to use a threshold value, just as we did in the probabilistic case. For instance, we might define that $\alpha \mid\!\sim \beta$ iff $val(\alpha \rightarrow \beta) \geqslant t$ for some threshold $t \in [0, 1]$. With this approach, the defeasible entailment relation will depend strongly on the treatment of conditional sentences.

4 Problems with numerical approaches

There are a couple of reasons why we might want some alternative to probability or fuzzy sets. Numerical approaches such as these are fine-grained, and there are often patterns that are only visible at a more coarse-grained level. For example, predicting the financial markets is a very difficult problem for which the first 'successful' strategy (approximately 55% success rate) was discovered in 2003 by physicists at Oxford (see *New Scientist* 10 April 2004 p34). The basic idea is that the history of a dynamical system can be thought of as a sequence of states. At any point of the sequence, predicting the future involves looking at the part of the sequence that has already occurred, i.e. the past. The various possible histories of even very complicated systems can often be sorted into a limited number of categories in such a way that all the histories in a given category give rise to the same short-range future. When several histories are thrown into the same category, many details of the individual histories become irrelevant, so that one is in effect tackling the problem at a coarse-grained level. The move up from the fine-grained level of detail to a coarser-grained level often involves changing from a quantitative to a qualitative model.

Even when a quantitative model is desired, it may be hard to come by. The major problem faced by every application of probability or fuzzy logic is: Where do the numbers come from?

Sometimes we are in a well-defined situation and it is clear how to get the numbers. At other times it is much less clear where the numbers should come from. Suppose we have an urn containing 100 balls of which 70 are white and 30 are black. The population is clearly demarcated, and we can make a definite judgment about relative frequency: 70 out of 100 balls are white, so the probability that a randomly drawn ball is white is $\frac{7}{10}$. On the other hand, suppose there is no urn, and balls of various colours are scattered about on the lawn. We know that we want

to talk about the white balls, and we can see a cluster of 70 white balls lying nearby, but around them are another 30 black balls, some red balls, some yellow balls, and further away there are more black balls, as well as balls of various other colours. Where do we draw a line of demarcation to get the population? Should we draw a circle around only the 70 white balls and 30 black balls nearest to us? Why should we not include some red balls, or some of the black balls lying a bit further away? In this example, we have no basis for judging relative frequency except some subjective judgment about which balls are 'relevant'.

Although the example of the balls is artificial, it points to a real problem. In applications of probability (i.e. in statistics) a great deal of effort should go into experimental design, in order to ensure that the numbers are meaningful. Often this aspect is neglected, because people think it is enough to go through some mechanical process of generating numbers and sticking these into formulas. As a result, most probabilities you encounter, even in scientific papers, simply cannot be taken seriously.

There is another (subtle) problem related with the way probability treats conjunctions. If E and X are independent events then $\Pr(E \cap X) = \Pr(E) \cdot \Pr(X)$, so that if if $E = \mathcal{M}(\alpha)$ and $X = \mathcal{M}(\beta)$, then $\Pr(\alpha \wedge \beta) = \Pr(\alpha) \cdot \Pr(\beta)$. Now this makes $\alpha \wedge \beta$ much less likely than either α or β, unless one of these has probability 1. The philosopher John Pollock has a nice example to show that this is counter-intuitive:

> For instance, consider an engineer who is designing a bridge. She will combine a vast amount of information about material strength, weather conditions, maximum load, costs of various construction techniques, and so forth, to compute what the size of a particular girder must be. These various bits of information are, presumably, independent of one another, so if the engineer combines 100 pieces of information, each with a probability of 0.99, the conjunction of that information has a probability of only $(0.99)^{100}$, which is approximately 0.366. According to the probabilist, she would be precluded from using all of this information simultaneously in an inference — but then it would be impossible to build bridges.

We shall see another example of the clash between probability and our intuitions about conjunctions in the next section.

Earlier we asked where the numbers come from. Perhaps because of this difficulty, humans have evolved alternatives to arithmetic calculation. In many circumstances our sense of likelihoods is qualitative rather than quantitative. We base our lives on default rules such as 'Cars

typically stop at red lights', but none of us calculate precise probabilities based on an assignment of precise numerical values to the possible outcomes. Clearly, humans use some alternative to probability, at least some of the time. The question is, what?

It can't be fuzzy logic. Although there are many equally attractive (or equally odd!) ways in which to specify the behaviour of the connectives in fuzzy logic, the main problem is that the truth value assigned to an atom has to be a precise number, even when we are representing a vague or inexact concept. After all, a fuzzy set is a function which assigns to each input a specific real number. How to choose that specific real value is often unclear. The general response of proponents has been to choose functions according to their graphical shape and to assume a kind of robustness: 'any function that looks more or less like this will do'. What is lacking is, of course, a compelling proof that the assumption of robustness is justified. Despite the use of numbers and arithmetical operations, everything still seems to be based, eventually, on a subjective judgement which often is merely the qualitative judgement 'this is a more typical member of the category than that'. Perhaps it would be easier to see what's going on if we strip away the numerical aspects and simply make the subjective judgements about typicality as obvious as possible.

5 Typicality

A basic idea of fuzzy logic is that a *category* (such as the category of young people) can be modeled mathematically by a fuzzy set, and then the fuzzy set can be used to give truth values to atoms in a multivalued logic, which in turn may provide a basis for defeasible beliefs. In this section, we summarise what cognitive psychology has to tell us about categories. We shall see that the idea of a 'typical member' of a category is important. In the next section we shall explore a new way of representing 'typicality' or 'normality' by means of order relations, and these order relations will provide the kind of basis for defeasible beliefs called nonmonotonic logic.

There are at least two reasons why humans evolved to understand the world by classifying things into categories (and why it would make sense to design artificial agents to do likewise).

One reason has to do with memory. As you know, retrieval by linear (sequential) search is very inefficient. Lumping together things that belong together allows more efficient retrieval. The second reason for categorisation is that it allows an agent to cope with situations never before encountered but bearing a family resemblance to previous situations (reducing new problems to old, reasoning by analogy, recognising

shapes, and so forth). Once bitten by a bull terrier, we show caution when meeting future bull terriers or other animals that bear a family resemblance to bull terriers.

How do people lump things together into categories? One influential view (sometimes called the Whorfian hypothesis) is that the world around us is an unstructured continuum of stimuli which children are taught to break into chunks in a manner that reflects the culture they grow up in and especially the language they learn to speak. If one believes this, it is conceivable that the animal kingdom might be categorised in the manner attributed to an ancient Chinese encyclopedia called the *Celestial Emporium of Benevolent Knowledge*:

> On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance (Borges, *Other Inquisitions*, Washington Square Press 1966).

But the interesting thing about this classification system is that it exists in no real culture. The classifications used in everyday life are much less arbitrary, as the psychologist Eleanor Rosch showed in a remarkable series of experiments from which three insights emerged.

Firstly, categorisation exploits *correlational structure*:

"Creatures with feathers are more likely to have wings than creatures with fur, and objects with the visual appearance of chairs are more likely to have functional sit-on-ableness than objects with the appearance of cats" as Rosch puts it. Categorisation is not the product of historical accident or culture or language, or at least not wholly. When experiments are devised that allow us to peer beneath the superficial differences of vocabulary, people of different cultures use roughly the same categories, at least at a basic level. The reason is that categorisation depends on the structure that exists in the world. The world does not consist of an unstructured continuum of stimuli; the world contains 'intrinsically separate things'. Objects have properties that persist (at least for a while), and these properties do not occur randomly, with equal likelihood, independently of the other properties of the object. Some combinations of properties tend to occur together, others don't. Since properties clump

together in the world, it makes sense to reflect this in our mental representation of the world. So categories consist of objects that share a family resemblance given by some clump of properties.

Secondly, categorisation has a *preferred level of granularity*:

Humans reason more efficiently about the categories 'chair' or 'table' than they reason about the more general (superordinate) category 'furniture' or the more specific (subordinate) categories 'kitchen chair' or 'armchair'. For example, we are faster when deciding whether two objects belong to the same category if the category is 'chair' rather than 'furniture' or 'kitchen chair'. Why is this? An agent's architecture determines that there is a particular level (or grainsize) at which categorisation is most efficient. Call this the basic level for that kind of agent. (For humans, 'chair' and 'table' are examples of basic level categories.) The basic level categories represent a compromise between having lots of small categories or having a few big ones. On the one hand, "it would appear to the organism's advantage to have as many properties as possible predictable from knowing any one property (which, for humans, includes the important property of the category name), a principle which would lead to formation of large numbers of categories with the finest possible discriminations between categories". On the other hand, "one purpose of categorization is to reduce the infinite differences among stimuli to behaviorally and cognitively usable proportions. It is to the organism's advantage not to differentiate one stimulus from others when that differentiation is irrelevant for the purposes at hand." The basic level of categorisation is the level at which objects belonging to the same category have the most properties in common while sharing the fewest properties with objects from other categories. It is also the most general level at which the same sequence of body and muscular movements can be used when interacting with objects belonging to the same category, and the most general level at which objects look sufficiently alike for the average shape of objects in the category to be readily recognized. We prefer to use basic categories because reasoning with them is easier.

Thirdly, categorisation makes use of *prototypes*:

Categories are not sets of things that are all equivalent. Some members of a category tend to be regarded as more typical than others. (This is particularly true of basic categories, though also often true of categories at other levels of granularity.) In the category of birds, for instance, robins and sparrows are the most typical members with parrots and hawks less typical, ducks and chickens still less typical, and ostriches and penguins very untypical. Interestingly, people "overwhelmingly agree in their judgments of how good an example or clear a case

members are of a category". The more typical of a category an object is judged to be, the more properties it has in common with other members of the category and the fewer properties it shares with members of contrasting categories. Flying is a property that many members of the category of birds share, and typical birds can fly. Untypical birds like ostriches and penguins share with members of other categories like cows and chairs the attribute of being unable to fly. Not only are some members of a category more typical than others, but we tend to use the typical members (the prototypes of the category) as 'cognitive reference points' when reasoning. For example, a line at an angle of 85 degrees to the horizontal is regarded as 'almost vertical', a vertical line is not regarded as 'almost 85 degrees'.

The asymmetry in the structure of categories, caused by the fact that some members are more typical than others, profoundly influences the way we think. It provides default rules that affect both memory and coping with new situations, and can sometimes lead us into error. Here are two examples:

Example 16 *Retrieval from memory may sometimes be inaccurate precisely because the most typical members of a category are generated automatically. A procedure developed by Deese and refined by Roediger & McDermott works as follows. The subject is presented with the list of associated words 'pin, eye, sewing, sharp, point, prick, thimble, haystack, thorn, hurt, injection, syringe, cloth, knitting' and then with another list, say 'bed, awake, rest, tired, dream, snooze, blanket, doze, slumber, snore, nap, peace, yawn, drowsy' and after the elapse of time subjects are asked whether each of the following words had been read aloud as part of the lists: 'sewing, door, needle, sleep, candy, awake'. Most people correctly remember that they had heard 'sewing' and 'awake', and correctly remember that they had not heard 'door' and 'candy', but MISTAKENLY claim that they had indeed heard 'needle' and 'sleep'. Interestingly, PET scans of brain activity during the process of recall show general similarity during both accurate and false recollection, but nonetheless a slight difference — a part of the frontal lobe showed more activity during false than during true recognition. In any case, it would seem that the falsely remembered items were so typical of their categories that they fooled the retrieval system.*

Example 17 *Consider making predictions about an unfamiliar object if you know only the category to which it belongs. For instance, a friend tells you that he has a new pet called Tweety, and that Tweety is a bird. It is sound common sense for you to assume that Tweety can fly. Of course, you may subsequently learn that Tweety is in fact a penguin, in*

which case your prediction will have been mistaken. But at the time, it was a plausible conjecture, because typical birds can fly.

Common-sense reasoning is full of 'category-based induction', where we reason from the knowledge that x is a member of category X to the conclusion that x has the properties that are typical of category X. Is this probabilistic reasoning in disguise? It certainly is not explicit probabilistic reasoning, since we don't perform arithmetical calculations on precise numerical values. But could it be unconscious probabilistic reasoning? The following example was invented by the psychologists Tversky and Kahneman to contrast typicality-based reasoning with probabilistic reasoning.

Example 18 *A personality sketch of a fictitious person is given: Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.*

Next, a list of statements is given, and subjects are asked to rank them by their likelihood:

1. *Bill is a physician who plays poker for a hobby.*
2. *Bill is an architect.*
3. *Bill is an accountant.*
4. *Bill plays jazz for a hobby.*
5. *Bill surfs for a hobby.*
6. *Bill is a reporter.*
7. *Bill is an accountant who plays jazz for a hobby.*
8. *Bill climbs mountains for a hobby.*

When this test was administered to a group of 88 undergraduates at the University of British Columbia in Canada, something interesting was observed. The great majority ranked item 7 (Bill is an accountant who plays jazz for a hobby) as more likely than item 4 (Bill plays jazz for a hobby).

From a probabilistic perspective, this doesn't make sense. Item 4 is an atomic sentence q and item 7 is a conjunction of atomic sentences $p \wedge q$. The probability of the conjunction can never be greater than the probability of one of its conjuncts, i.e. we must always have $\Pr(p \wedge q) \leq$

$\Pr(q)$. And so the reasoning of the experimental subjects would appear to violate the constraints of probability theory (which is how Kahneman and Tversky interpreted their results).

There is however a different way to look at the matter. The personality sketch was carefully constructed to make Bill seem a typical accountant. Item 4 doesn't mention Bill being an accountant, and so people tend to read item 4 as equivalent to (Bill is not an accountant and plays jazz for a hobby). Such a reading is reasonable, because we expect communications to supply all (and only) relevant information — we would be equally dissatisfied if someone withheld relevant facts and if they flooded us with irrelevant information that obscures the point. If the experimental subjects were applying their usual presuppositions about what constitutes good manners in conversation, then item 7 is certainly more likely than item 4 because the subjects were interpreting item 4 not as an atom q but as a conjunction $\neg p \wedge q$. Under this interpretation, their reasoning was probabilistically correct. (This interpretation was not investigated by Kahneman and Tversky.)

Finally, note that although Kahneman and Tversky suppose that a probabilistic approach is the only correct one, it is by no means clear what the sample space should be. Instead of having balls of two distinct colours in an urn, we have balls of many different colours strewn on the lawn. Does the sample space include the entire population of humans? Only the males? Only the males in a certain age group? Only the males in a certain geographical area? Only the males who have certain personality traits? Over what period of history? There are no clear boundaries marking the population of interest the way an urn separates the balls of interest from the rest of the world. Some sort of subjective choice would need to be made before a probability space could be set up.

It would seem that the relationship between probabilistic and typicality reasoning deserves further investigation. There is a distinct possibility that typicality reasoning is a quick and dirty approximation to probabilistic reasoning, which humans have evolved to use because it is faster than arithmetic and also because it can be applied in situations where probabilistic reasoning cannot (because there is no clear way to get the numbers that probability would require).

We now turn to a way to represent categories that is different from fuzzy sets — a very general approach adequate not only for categories but for all *default rules*. We show how logic can be adapted to make use of these representations.

6 Nonmonotonic logic

Suppose we have an agent who has been monitoring the Light-Fan System for a long time. The agent has noted that:

- The light is on most of the time. (Thus 11 and 10 are the most normal states.)
- When the light is on, the fan spends about as much time being on as off. (Thus 11 and 10 are equally normal.)
- When the light is off, then the fan is usually on and very seldom off. (Thus 01 is more normal than 00, though less normal than 11 and 10.)

This heuristic information helps the agent to reason about the system, for if observation reveals only that the light is off (so that the fan cannot be seen in the darkness), the agent may plausibly conclude that the fan is on. (Of course, very occasionally this conclusion is mistaken, and the agent obliged to revise her thinking.)

The agent's reasoning could of course be based on probabilities. If the agent had meticulously recorded, over the past ten years, exactly how much time the system spent in each state, then the proportion of time spent in state v could be regarded as the probability of state v. Once probabilities have been assigned to the states (outcomes), then as we saw before, every set of states (and hence every sentence) receives a probability too.

But what if the agent's heuristic information were qualitative rather than statistical? Suppose the agent has not kept precise records, and instead has merely acquired the firm impression that 11 and 10 were the most normal states, with 01 somewhat less normal and 00 quite abnormal? Common-sense reasoning would still support the idea that when the light is off, the system is typically in state 01, so that it is reasonable to expect the fan to be on. The purpose of nonmonotonic logic is to formalise this kind of common-sense reasoning in a way that can cope with both quantitative and qualitative heuristics..

The nonmonotonic logic we shall describe evolved out of research in artificial intelligence during the 1970s. John McCarthy invented an approach to common-sense reasoning called 'circumscription', which eventually led to a much more elegant approach first described by Shoham. Shoham's approach was then elaborated by Kraus, Lehmann, and Magidor. The approach is called 'preferential semantics' or 'minimal model semantics'. The idea is the following.

Let us refer to the agent's heuristic information as a *default rule*. Examples of default rules, expressed in our metalanguage, are 'If something is a bird, then normally it can fly' and 'If the light is off, then normally the fan is on'. Such default rules are not expressible by sentences in the sort of propositional object language with which we have been working. However, default rules can be represented semantically, by taking the set of states of the system and ordering these from most normal to least normal. Since the ordering indicates the states that are most normal, the agent can test sentences to see whether they are true in the most normal states.

6.1 Ranked interpretations

We start by examining order relations.

Definition 19 *(**Orderings**) A relation $\preccurlyeq$ on a set S is a **preorder** on S iff $\preccurlyeq$ is reflexive on S and is transitive.*

*A preorder $\preccurlyeq$ on S is **total** if any two elements of S are comparable, i.e. if for all $x, y \in S$, either $x \preccurlyeq y$ or $y \preccurlyeq x$.*

We write $x \prec y$ to indicate that $x \preccurlyeq y$ but not $y \preccurlyeq x$.

*For every $X \subseteq S$, an element y is **minimal** in X iff $y \in X$ and there is no $x \in X$ such that $x \prec y$.*

For example, let $S = \{1, 2, 3, 4\}$ and $\precsim$ be the relation $\{(1, 1), (2, 2), (3, 3), (4, 4), (1, 2), (1, 3), (2, 3)\}$. Then $\precsim$ is a preorder, but not a total preorder since the elements 1 and 4 are incomparable, as are the elements 2 and 4.

On the other hand, if we let $\preceq$ be the relation $\{(1, 1), (2, 2), (3, 3), (4, 4), (1, 2), (2, 1), (1, 3), (2, 3), (1, 4), (2, 4), (3, 4), (4, 3)\}$ on S, then $\preceq$ is a total preorder. If $X = \{1, 2, 3\}$ then both 1 and 2 are minimal elements of X. However, 3 is not minimal in S because there is an element, namely 1, such that $1 \preceq 3$ while it is not the case that $3 \preceq 1$. Minimality is defined relative to subsets X. Even though 3 is not minimal in S, 3 is minimal in, say, the subset $X = \{3, 4\}$.

Total preorders are important because (for finite sets S) they tell us how to arrange elements into levels. In the case of the total preorder $\preceq$, the bottom level is occupied by the elements 1 and 2, and this level lies below a second, higher, level occupied by the elements 3 and 4. To see this, read a pair (x, y) in $\preceq$ as asserting that x lies below or at the same level as y. It follows that whenever both $x \preceq y$ and $y \preceq x$, then x and y occupy the same level, whereas when $x \preceq y$ but $y \not\preceq x$ then x lies on a level strictly below that of y.

For small sets S, it is often convenient to depict a preorder $\preccurlyeq$ visually, and if the preorder is total then it is very easy to visualise. One may

use a diagram I call a *filing-cabinet*. We place the members of S into drawers of a filing cabinet, and we place x into a lower drawer than y to indicate that $x \prec y$. In case we have $x \preccurlyeq y$ and $y \preccurlyeq x$ we put x and y in the same drawer. Here is a diagram of $\preceq$.

3	4
1	2

When we use a total preorder $\preccurlyeq$ to arrange states from most normal to least normal, the convention is to put the most normal states at the bottom and the most abnormal states at the top. Thus we would read $x \preccurlyeq y$ as 'x is at least as normal as y' or as 'y is at least as abnormal as x'.

Definition 20 *(**Interpretations**) Let L_A be any propositional language. By a **finite ranked interpretation** of L_A we understand a triple $\mathcal{I} = (S, \preccurlyeq, V)$ such that the set S of states is finite and nonempty, $V : S \longrightarrow W_A$ is a labelling function, and $\preccurlyeq$ is a total preorder on S.*

The simplest sort of ranked interpretation of L_A would of course make $S = W_A$ and V the identity function given by $V(s) = s$ for every $s \in S$. A finite ranked interpretation is our way of representing a system together with an agent's default rule Here is an example.

Example 21 *Consider the Light-Fan System. Suppose we have the default rule expressed in the metalanguage by "The light is on most of the time; when the light is on, the fan spends about as much time being on as off; and when the light is off then the fan is usually kept on". Then we can express the system plus default rule semantically by taking $S = \{11, 10, 01, 00\} = W_A$, V to be the identity function on S, and $\preccurlyeq$ to be the total preorder depicted below.*

00	
01	
11	10

What if the system has an infinite set of states? It is still possible to represent default rules by total preorders on the set S of states, but care must be taken to ensure that the ordering works properly. For instance, one should ensure that there is no infinite descending chain of states $s \succcurlyeq s' \succcurlyeq \ldots$ because then we would lack any concept of 'lowest down' or 'minimal'. The technical precautions are spelled out in the references.

For most purposes, a finite set of states will suffice, and then the formal details are simpler.

Suppose we associate with the language L_A some ranked interpretation $\mathcal{I} = (S, \preccurlyeq, V)$. What does it mean to say that a state $s \in S$ satisfies (i.e. makes true) a sentence α? Exactly what we've always understood. We use V to get to the valuation associated with s in order to decide whether atoms are true, and then percolate truth values upward according to the way connectives behave. The order relation $\preccurlyeq$ plays no role in satisfaction, and is only used when we want to pick out the models of a sentence that are *minimal*, i.e. as low down in the ordering as possible.

Exercise 22 *1. Give filing-cabinet diagrams depicting all the total preorders on $S = \{a, b, c\}$. Do the same for $S' = \{a, b, c, d\}$.*

2. Consider the Light-Fan System, and think of it as a metaphor for a helicopter. The rotor (fan) provides lift and is augmented by a jet (since the jet uses heat to produce expansion of air, we model it by the light). Construct a finite ranked interpretation of the language with $A = \{p, q\}$ in which the ordering portrays the following rules of thumb. It is very unusual for the jet to be on while the rotor is off. It is normally the case that the helicopter is on the ground with everything switched off. It is less normal but not very unusual for the helicopter to be flying, in which case it is equally likely to have just the rotor on as it is to have both rotor and jet on.

Consider the models of $\neg p$. What are the minimal elements of the set $\mathcal{M}(\neg p)$?

3. Suppose discovery of a cheap non-polluting source of fuel means that the helicopter can spend most of its time in the air, and that the rotor is needed only for taking off and landing. (Thus the jet alone is used in the air, while the jet and rotor together are used for take-offs and landings. The rotor is never used alone.) Assume that the helicopter spends about as much time on the ground undergoing maintenance as it does doing take-offs and landings. Construct a new interpretation that fits this scenario.

Consider the models of q. What are the minimal elements of the set $\mathcal{M}(q)$?

4. Consider the 3 Card System in which each of three players get one of three cards coloured red, green or blue. Suppose the knowledge representation language has 9 atoms r_1, r_2, r_3, g_1, g_2, g_3, b_1, b_2, b_3 where r_1 stands for 'Player 1 has the red card' and so on. The

cards are dealt to players by a new method. The dealer rolls a 6-sided die, in which each side is equally likely to be uppermost. If the uppermost face shows a 1, 2, 3, or 4, the dealer gives the red card to player 1. If the uppermost face is 5, player 1 gets the green card, and if it is 6 he gets the blue card. Next the dealer shuffles the remaining two cards and gives one to player 2 and the last to player 3.

Construct a finite ranked interpretation to represent this system together with the default rule which regards the most probable state as the most normal and less probable states as less normal.

Consider the models of g_2. What are the minimal elements of the set $\mathcal{M}(g_2)$?

6.2 Rational entailment

Every finite ranked interpretation $\mathcal{I}$ determines its own defeasible entailment relation on L_A.

Definition 23 *($\mid\!\sim$) Let $\mathcal{I} = (S, \preccurlyeq, V)$ be a finite ranked interpretation of the propositional language L_A. For any sentence α, let $\mathcal{M}in(\alpha)$ be the set of minimal models of α. The defeasible entailment relation induced by $\mathcal{I}$ is the relation $\mid\!\sim$ given by*

$$\alpha \mid\!\sim \beta \text{ iff } \beta \text{ is satisfied at all the minimal models of } \alpha.$$

In other words, $\alpha \mid\!\sim \beta$ iff $\mathcal{M}in(\alpha) \subseteq \mathcal{M}(\beta)$.

First, an example. Suppose the agent observing the Light-Fan System has the default rule represented by the total preorder below.

00
01
11 10

The obvious finite ranked interpretation has $S = \{11, 10, 01, 00\} = W_A$, V the identity function, and $\preccurlyeq$ the total preorder representing the default information. The defeasible entailment relation $\mid\!\sim$ induced by the interpretation is such that, for instance, $\neg p \mid\!\sim q$. To see this, look at the models of $\neg p$, namely 01 and 00. The lower of the two is 01, so $\mathcal{M}in(\neg p) = \{01\}$. But q is satisfied by 01. So $\mathcal{M}in(\neg p) \subseteq \mathcal{M}(q)$.

This is in stark contrast with classical entailment, for it is certainly not the case that $\neg p \models q$.

Let us compare the properties of $\mid\!\sim$ with those of the classical entailment relation $\models$. Recall that $\alpha \models \beta$ iff $\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$. Recall that

semantic equivalence is given by $\alpha \equiv \beta$ iff $\mathcal{M}(\alpha) = \mathcal{M}(\beta)$. The following basic properties of $\models$ are easy to verify: $\models$ is reflexive on L_A (*i.e.* for all α, $\alpha \models \alpha$); $\models$ is transitive (*i.e.* if $\alpha \models \beta$ and $\beta \models \gamma$ then $\alpha \models \gamma$); and $\models$ is monotonic (*i.e.* if $\alpha \models \beta$ and if γ is arbitrary, then $\alpha \wedge \gamma \models \beta$. Monotonicity is a kind of cautiousness property - it says that if α entails β then β is so trustworthy it cannot be undermined by any additional knowledge γ. But of course defeasible reasoning is deliberately more bold than that, and can lead to conclusions that may need to be retracted in the light of new information; thus we would expect defeasible entailment not to be monotonic in general. And for this reason the use of defeasible entailment relations is called nonmonotonic logic.

Theorem 24 *The defeasible entailment relation $\mid\!\sim$ induced by a finite ranked interpretation $\mathcal{I} = (S, \preccurlyeq, V)$ has the following properties:*

- $\mid\!\sim$ *is well-behaved relative to semantic equivalence: if $\alpha \equiv \alpha'$ and $\beta \equiv \beta'$ and $\alpha \mid\!\sim \beta$ then $\alpha' \mid\!\sim \beta'$*
- $\mid\!\sim$ *is supraclassical: if $\alpha \models \beta$ then $\alpha \mid\!\sim \beta$*
- $\mid\!\sim$ *is reflexive: for every sentence α, $\alpha \mid\!\sim \alpha$*
- $\mid\!\sim$ *is not necessarily transitive: it is possible to find a finite ranked interpretation whose induced $\mid\!\sim$ is such that for some sentences α, β, γ it is the case that $\alpha \mid\!\sim \beta$ and $\beta \mid\!\sim \gamma$ but not the case that $\alpha \mid\!\sim \gamma$*
- $\mid\!\sim$ *is not necessarily monotonic: it is possible to find a finite ranked interpretation whose induced $\mid\!\sim$ is such that for some sentences α, β, γ it is the case that $\alpha \mid\!\sim \beta$ but not the case that $\alpha \wedge \gamma \mid\!\sim \beta$*

Proof. *Suppose $\mathcal{I}$ is a finite ranked interpretation.*

- *(Well-behavedness) Suppose $\alpha \equiv \alpha'$ and $\beta \equiv \beta'$ and $\alpha \mid\!\sim \beta$. Thus $\mathcal{M}in(\alpha) \subseteq \mathcal{M}(\beta)$. But $\mathcal{M}(\beta) = \mathcal{M}(\beta')$ and since $\mathcal{M}(\alpha) = \mathcal{M}(\alpha')$ it follows that $\mathcal{M}in(\alpha) = \mathcal{M}in(\alpha')$. Thus $\alpha' \mid\!\sim \beta'$.*
- *(Supraclassicality) Suppose $\alpha \models \beta$. Thus $\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$. Since $\mathcal{M}in(\alpha) \subseteq \mathcal{M}(\alpha)$, it follows that $\alpha \mid\!\sim \beta$.*
- *(Reflexivity) Exercise for the reader.*
- *(Failure of transitivity) Consider the language with $A = \{p, q\}$ and let $\mathcal{I}$ be the finite ranked interpretation with $S = W_A$ and V the identity function and $\preccurlyeq$ any suitable ordering such that $10 \preccurlyeq 11$.*

Now $p \wedge q \mid\!\sim p$ because p is true in 11, which is the only (minimal) model of $p \wedge q$. And in turn $p \mid\!\sim p \wedge \neg q$ because $p \wedge \neg q$ is true in 10, which is the minimal model of p. However, it is not the case that $p \wedge q \mid\!\sim p \wedge \neg q$ because $p \wedge \neg q$ is not satisfied by 11, which is the minimal model of $p \wedge q$.

- *(Failure of monotonicity) Let $\mathcal{I}$ be as in the preceding, with $\preccurlyeq$ such that 11 is alone in the bottom level. Now $p \mid\!\sim q$ but since the only model of $p \wedge \neg q$ is 10, it is not the case that $p \wedge \neg q \mid\!\sim q$.*

■

Theorem 25 *The defeasible entailment relation induced by a finite ranked interpretation is **cumulative**, by which we mean that $\mid\!\sim$ has the following properties:*

- *(Reflexivity)*
- *(Left Equivalence) If $\alpha \mid\!\sim \gamma$ and $\alpha \equiv \beta$ then $\beta \mid\!\sim \gamma$*
- *(Right Weakening) If $\alpha \mid\!\sim \beta$ and $\beta \models \gamma$ then $\alpha \mid\!\sim \gamma$*
- *(Cut) If $\alpha \wedge \beta \mid\!\sim \gamma$ and $\alpha \mid\!\sim \beta$ then $\alpha \mid\!\sim \gamma$*
- *(Cautious Monotonicity) If $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$ then $\alpha \wedge \gamma \mid\!\sim \beta$*

Proof. *Reflexivity has been established already.*

Left Equivalence follows because if γ is true at the minimal models of α, and β has exactly the same models as α, then γ is true at the minimal models of β as well.

Right Weakening follows because if $\mathcal{M}in(\alpha) \subseteq \mathcal{M}(\beta)$ and $\mathcal{M}(\beta) \subseteq \mathcal{M}(\gamma)$ then $\mathcal{M}in(\alpha) \subseteq \mathcal{M}(\gamma)$.

To see that Cut follows, suppose γ is true at all those states that are models of both α and β and are minimal in this regard (i.e. as low down in the ordering as possible). If $\alpha \mid\!\sim \beta$, then β is true at all the minimal models of α. Now we claim that $\mathcal{M}in(\alpha) \subseteq \mathcal{M}in(\alpha \wedge \beta)$. For if $s \in \mathcal{M}in(\alpha)$, then s is a model of β, so s is a model of $\alpha \wedge \beta$ and the only question is whether it is as low down as possible in the set $\mathcal{M}(\alpha \wedge \beta)$. And indeed it must be, because $\mathcal{M}(\alpha \wedge \beta) \subseteq \mathcal{M}(\alpha)$ and nothing in $\mathcal{M}(\alpha)$ could be found that lives below s. Thus $\mathcal{M}in(\alpha) \subseteq \mathcal{M}in(\alpha \wedge \beta)$. And so $\alpha \mid\!\sim \gamma$.

For Cautious Monotonicity, we proceed as follows. Suppose $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$. We want to show that $\alpha \wedge \gamma \mid\!\sim \beta$. Pick any $s \in \mathcal{M}(\alpha \wedge \gamma)$ which is minimal. Is $s \in \mathcal{M}(\beta)$? Well, if we could show that $s \in \mathcal{M}in(\alpha)$, then the required result would follow. But s must indeed be a

minimal model of α, for suppose this were not the case. Then there is some minimal model s' of α below s, for otherwise there would have to be an infinite descending chain of models of α. Since $\alpha \mathrel{|\!\sim} \gamma$, γ is true at s' and so $s' \in \mathcal{M}(\alpha \wedge \gamma)$. But this contradicts the choice of s as a minimal element of this set, because s' is below s. ■

A defeasible entailment relation cannot be expected to be monotonic in the way classical logic is monotonic. That is, it cannot be expected that for all γ it will be the case that if $\alpha \mathrel{|\!\sim} \beta$ then $\alpha \wedge \gamma \mathrel{|\!\sim} \beta$. The property of Cautious Monotonicity expresses the idea that learning a new fact, the truth of which could earlier have been defeasibly concluded, should not invalidate previous defeasible conclusions. Thus Cautious Monotonicity singles out certain 'safe' γ for which monotonicity does still hold.

Theorem 26 *The defeasible entailment relation induced by a finite ranked interpretation is **preferential**, by which we mean that $\mathrel{|\!\sim}$ has the following properties:*

- *(Cumulativity)*
- *(And) If $\alpha \mathrel{|\!\sim} \beta$ and $\alpha \mathrel{|\!\sim} \gamma$ then $\alpha \mathrel{|\!\sim} \beta \wedge \gamma$*
- *(Or) If $\alpha \mathrel{|\!\sim} \gamma$ and $\beta \mathrel{|\!\sim} \gamma$ then $\alpha \vee \beta \mathrel{|\!\sim} \gamma$*

Proof. *Cumulativity has been established already.*

The property And is easy to derive directly. But it is also interesting to note that it follows from Cumulativity. For suppose that $\alpha \mathrel{|\!\sim} \beta$ and $\alpha \mathrel{|\!\sim} \gamma$. By Cautious Monotonicity, $\alpha \wedge \beta \mathrel{|\!\sim} \gamma$. Now notice that, classically, $\alpha \wedge \beta \wedge \gamma \models \beta \wedge \gamma$ and so by Supraclassicality we get $\alpha \wedge \beta \wedge \gamma \mathrel{|\!\sim} \beta \wedge \gamma$. By Cut we may now get first that $\alpha \wedge \beta \mathrel{|\!\sim} \beta \wedge \gamma$ and by another application of Cut that $\alpha \mathrel{|\!\sim} \beta \wedge \gamma$. (We could also prove And directly, using the ordering.)

To see that the property Or holds, suppose that $\alpha \mathrel{|\!\sim} \gamma$ and $\beta \mathrel{|\!\sim} \gamma$. We want to show that $\mathcal{M}in(\alpha \vee \beta) \subseteq \mathcal{M}(\gamma)$. Pick any state s that is minimal in $\mathcal{M}(\alpha \vee \beta)$. Since $\mathcal{M}(\alpha \vee \beta) = \mathcal{M}(\alpha) \cup \mathcal{M}(\beta)$, s belongs to $\mathcal{M}(\alpha)$ or $\mathcal{M}(\beta)$. Suppose the former. There was no state in the larger $\mathcal{M}(\alpha \vee \beta)$ which lived below s in the ordering, and so there certainly cannot be any state in the smaller $\mathcal{M}(\alpha)$ which lives below s. Thus s is minimal in $\mathcal{M}(\alpha)$. Since $\alpha \mathrel{|\!\sim} \gamma$, it follows that s makes γ true. The argument is similar if s belongs to $\mathcal{M}(\beta)$ rather than to $\mathcal{M}(\alpha)$. Thus every minimal model of $\alpha \vee \beta$ satisfies γ. ■

Theorem 27 *The defeasible entailment relation induced by a finite ranked interpretation is* ***rational****, by which we mean that* $\mid\!\sim$ *is preferential and satisfies*

- *(Rational Monotonicity) If* $\alpha \mid\!\sim \beta$ *then either* $\alpha \wedge \gamma \mid\!\sim \beta$ *or* $\alpha \mid\!\sim \neg\gamma$

Proof. *It has been established that* $\mid\!\sim$ *is preferential.*

To establish that Rational Monotonicity holds, suppose that $\alpha \mid\!\sim \beta$*. There are two possibilities: either* $\alpha \mid\!\sim \neg\gamma$ *or this is not the case. Let us assume it is not the case that* $\alpha \mid\!\sim \neg\gamma$ *and try to show that* $\alpha \wedge \gamma \mid\!\sim \beta$*.*

Since $\neg\gamma$ *fails to be satisfied by all the minimal models of* α*, there is some minimal model of* α *(say* t*) which satisfies* γ*. Let* s *be a minimal model of* $\alpha \wedge \gamma$*. Now notice that* t *is a model of* $\alpha \wedge \gamma$*. So* t *cannot live strictly below* s*, for* s *is minimal in* $\mathcal{M}(\alpha \wedge \gamma)$*. But this means that* s *is a minimal model of* α*. Why? Well,* $s \in \mathcal{M}(\alpha)$ *and, since* $\preccurlyeq$ *is a total preorder, it must be the case that* $s \preccurlyeq t$ *or* $t \preccurlyeq s$ *or both. We have excluded above the possibility that* $t \preccurlyeq s$ *unless it is also the case that* $s \preccurlyeq t$*. Hence it must certainly be the case that* $s \preccurlyeq t$*. And this establishes that* $s \in \mathcal{M}in(\alpha)$*, for if* $s' \in \mathcal{M}(\alpha)$ *and* $s' \preccurlyeq s$ *without* $s \preccurlyeq s'$ *in return, then* $s' \preccurlyeq t$ *and it cannot hold that* $t \preccurlyeq s'$*, contradicting the minimality of* t *in* $\mathcal{M}(\alpha)$*. What this means is that we have taken an arbitrary minimal model* s *of* $\alpha \wedge \gamma$ *and shewn it to be a minimal model of* α*. By the assumption that* $\alpha \mid\!\sim \beta$*, it now follows that* β *is satisfied by* s*. But then* $\alpha \wedge \gamma \mid\!\sim \beta$*.* ■

Recall that Cautious Monotonicity demonstrated that some γ were safe increments of the information in α, in the sense that if $\alpha \mid\!\sim \beta$ then $\alpha \wedge \gamma \mid\!\sim \beta$, basically because γ was something that could have been defeasibly inferred from α anyway. The property of Rational Monotonicity strengthens Cautious Monotonicity by asserting that it is only if the additional information γ is a surprise, $\neg\gamma$ having been expected, that this new information can require us to withdraw previous conclusions.

Exercise 28 *1. Consider again the Light-Fan System as a metaphor for a helicopter. Recall the finite ranked interpretation of the language with* $A = \{p, q\}$ *in which the ordering portrays the following rules of thumb. It is very unusual for the jet to be on while the rotor is off. It is normally the case that the helicopter is on the ground with everything switched off. It is less normal but not very unusual for the helicopter to be flying, in which case it is equally likely to have just the rotor on as it is to have both rotor and jet on.*

Let $\mid\!\sim$ be the defeasible entailment relation induced by this interpretation. Recall that p expresses that the light (jet) is on and q that the fan (rotor) is on. Is it the case that

- $p \mid\!\sim p \wedge q$*?*
- $q \mid\!\sim p \wedge q$*?*
- $p \vee q \mid\!\sim p$*?*
- $p \vee q \mid\!\sim q$*?*
- $p \vee \neg p \mid\!\sim p$*?*
- $p \vee \neg p \mid\!\sim \neg p$*?*

2. *Recall that discovery of a cheap non-polluting source of fuel changes things so that that the helicopter can spend most of its time in the air, and that the rotor is needed only for taking off and landing. Assume that the helicopter spends about as much time on the ground undergoing maintenance as it does doing take-offs or landings. If $\mid\!\sim'$ is the defeasible entailment relation induced by the new ranked interpretation, verify whether*

   - $p \mid\!\sim' p \wedge q$*?*
   - $q \mid\!\sim' p \wedge q$*?*
   - $p \vee q \mid\!\sim' p$*?*
   - $p \vee q \mid\!\sim' q$*?*
   - $p \vee \neg p \mid\!\sim' p$*?*
   - $p \vee \neg p \mid\!\sim' \neg p$*?*

3. *Recall the 3 Card System in which each of three players get one of three cards coloured red, green or blue. Suppose the knowledge representation language has 9 atoms r_1, r_2, r_3, g_1, g_2, g_3, b_1, b_2, b_3 where r_1 stands for 'Player 1 has the red card' and so on. The cards are dealt to players by a new method. The dealer rolls a 6-sided die, in which each side is equally likely to be uppermost. If the uppermost face shows a 1, 2, 3, or 4, the dealer gives the red card to player 1. If the uppermost face is 5, player 1 gets the green card, and if it is 6 he gets the blue card. Next the dealer shuffles the remaining two cards and gives one to player 2 and the last to player 3.*

   Take the obvious finite ranked interpretation which represents this system together with the default rule which regards the most probable state as the most normal and less probable states as less normal.

Let $\mid\!\sim$ be the defeasible entailment relation induced by this interpretation. Verify whether

- $g_2 \mid\!\sim r_1$*?*
- $g_2 \vee g_3 \mid\!\sim r_1$*?*
- $r_1 \mid\!\sim g_2$*?*
- $r_1 \mid\!\sim g_2 \vee b_2$*?*

4. *Construct a finite ranked interpretation whose induced defeasible entailment relation is just classical entailment $\models$.*

5. *Show that classical entailment has the following properties:*

   - *(Contraposition) If $\alpha \models \beta$, then $\neg\beta \models \neg\alpha$.*
   - *(Conditional Monotonicity) If $\alpha \rightarrow \beta$ is a tautology and $\beta \models \gamma$, then $\alpha \models \gamma$.*
   - *(EHD) If $\alpha \models \beta \rightarrow \gamma$ then $\alpha \wedge \beta \models \gamma$.*
   - *(S) If $\alpha \wedge \beta \models \gamma$ then $\alpha \models \beta \rightarrow \gamma$.*
   - *(Merged or) If $\alpha \models \gamma$ and $\beta \models \delta$ then $\alpha \vee \beta \models \gamma \vee \delta$.*

   Suppose $\mid\!\sim$ is the defeasible entailment relation induced by a finite ranked interpretation. Is it necessarily the case that

   - *if $\alpha \mid\!\sim \beta$ then $\neg\beta \mid\!\sim \neg\alpha$?*
   - *if $\alpha \rightarrow \beta$ is a tautology and $\beta \mid\!\sim \gamma$, then $\alpha \mid\!\sim \gamma$?*
   - *if $\alpha \mid\!\sim \beta \rightarrow \gamma$ then $\alpha \wedge \beta \mid\!\sim \gamma$?*
   - *if $\alpha \wedge \beta \mid\!\sim \gamma$ then $\alpha \mid\!\sim \beta \rightarrow \gamma$?*
   - *if $\alpha \mid\!\sim \gamma$ and $\beta \mid\!\sim \delta$ then $\alpha \vee \beta \mid\!\sim \gamma \vee \delta$?*

   Give reasons for your answers.

6. *Show that if $\mid\!\sim$ is induced by a finite ranked interpretation then $\mid\!\sim$ has the property:*

   - *(Disjunctive Rationality) if $\alpha \vee \beta \mid\!\sim \gamma$ then $\alpha \mid\!\sim \gamma$ or $\beta \mid\!\sim \gamma$.*

7. *Explain the difference between cautious monotonicity and rational monotonicity by describing a situation in which cautious monotonicity would not entitle us to conclude that $\alpha \wedge \gamma \mid\!\sim \beta$ but rational monotonicity would.*

7 Review of lectures 5 and 6

In these two lectures we saw that while classical entailment formalises cautious reasoning, defeasible entailment relations are needed to formalise common-sense reasoning. In humans, emotions influence whether cautious reasoning or defeasible reasoning involving heuristic information is used:

- Isen AM and Means B: The influence of positive affect on decision-making strategy. *Social Cognition* **2**:18-31 1983.
- Isen AM, Daubman KA, and Nowicki G: Positive affect facilitates creative problem-solving. *Journal of Personality and Social Psychology* **52**:1122-1131 1987.
- Fredrickson B: What good are positive emotions. *Review of General Psychology* **2**:300-319 1998.
- Fredrickson B: The role of positive emotions in Positive Psychology: The broaden-and-build theory of positive emotions. *American Psychologist* **56**:218-226 2001.

Defeasible reasoning is characterised by the use of heuristic (indefinite) information. We saw how probabilities can be employed to represent heuristic information, provided one has enough information about the system to come up with a probability measure in the first place. My own favourite references on probability are:

- Carnap, R: *Logical Foundations of Probability*, University of Chicago Press 1950 (a classic).
- Hamming, RW: *Art of Probability for Scientists and Engineers*, Addison-Wesley 1991.
- Howson, C and Urbach, P: *Scientific Reasoning: The Bayesian Approach* (2nd ed), Open Court 1993.

We briefly examined the use of fuzzy sets. References are:

- Zadeh, L: Fuzzy Sets. *Information and Control* **8**:338-353 1965.
- Zadeh L: Fuzzy logic and approximate reasoning. *Synthese* **30**:407-428 1975.
- Lakoff G: Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* **2**:458-508 1965.

- Gaines BR: Foundations of fuzzy reasoning. *International Journal of Man-Machine Studies* **6**:623-668 1976.

- Goguen JA: The logic of inexact concepts. *Synthese* **19**:325-373 1968/69.

Where less information is available than would be needed to support a numerical approach, total preorders can be used to represent default rules in a qualitative rather than quantitative way. This representation of default rules leads to the idea of representing a system by a ranked interpretation. Logic whose semantics is provided by ranked interpretations and their induced defeasible entailment relations is called nonmonotonic logic, because the defeasible entailment relations cannot be counted on to satisfy more than limited forms of monotonicity.

The approach to nonmonotonic logic that we sketched in this lecture is called by various names, including 'preferential model semantics' or 'minimal model semantics'. It should be noted that this is not the only approach to nonmonotonic logic that has been developed. At the time McCarthy was taking the first steps towards minimal model semantics with circumscription, Ray Reiter was inventing an approach called default logic and McDermott, Doyle, and Moore were exploring modal approaches that evolved into auto-epistemic logic. However, preferential semantics is in a sense the most general approach, because (as Shoham showed) the alternative approaches can be reconstructed within minimal model semantics. Key references on preferential semantics are:

- John McCarthy: Epistemological Problems of Artificial Intelligence. *Proceedings of the IJCAI 1977*, reprinted in M Ginsberg (ed), *Readings in nonmonotonic reasoning*. Morgan Kaufmann 1987 pp46-52. (Of historical interest.)

- Yoav Shoham: *Reasoning about change*. MIT Press 1988. (Very readable.)

- Kraus S, Lehmann D, and Magidor M: Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence* **44**:167-207 1990. (The most famous reference.)

- Lehmann D and Magidor M: What does a conditional knowledge base entail? *Artificial Intelligence* **55**:1-60 1992. (Talks about modular partial orders, which do the same as total preorders.)

There is another reason why preferential semantics has a claim to be the dominant approach to nonmonotonic logic. Defeasible reasoning

is about forming conclusions which, while plausible, may under exceptional circumstances be mistaken. If an agent has formed a defeasible conclusion and subsequently learns it to be mistaken, what then? Somehow the agent must revise its beliefs. Research into belief change found natural connections with preferential semantics, as we shall see in the next lecture. This increases one's confidence that preferential semantics is the right way to go.

Finally, this lecture briefly sketched research on categorisation and typicality that offers insights into human defeasible reasoning. Relevant references are, alphabetically by author:

- Deese: On the prediction of occurrence of particular verbal intrusions in immediate recall: *Journal of Experimental Psychology* **58**:17-22 1959.

- Osherson, Smith, Wilkie, Lopez and Shafir: Category-based induction. *Psychological Review* **97**:185-200 1990.

- Rips L: Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior* **14**:665-681 1975.

- Roediger & McDermott: Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**:803-814 1995.

- Rosch E: Natural categories. *Cognitive Psychology* **4**:328-350 1973.

- Rosch E: Cognitive reference points. *Cognitive Psychology* **7**:532-547 1975.

- Rosch E and Mervis C: Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* **7**:573-605 1975.

- Rosch, Mervis, Gray, Johnson and Boyes-Braem: Basic objects in natural categories. *Cognitive Psychology* **8**:382-439 1976.

- Rosch E: Principles of categorization. In Collins A & Smith EE (editors): *Readings in Cognitive Science*. Morgan Kaufmann 1988. (Both Rosch's paper and the broader collection of readings are highly recommended.)

- Tversky A and Kahneman D: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* **90**(4):293-314 1983.

Glossary

- **default rule** — heuristic information, represented in preferential semantics by an order relation on the set of states.
- **interpretation** — what some of us call the semantic stucture that we associate with an object language L_A in order to talk about satisfaction and models; the preferential semantics of nonmonotonic logic involves finite ranked interpretations.
- **ranked interpretation** — an ontology (S, V) together with an order relation $\preccurlyeq$ on S which is a total preorder.
- **total preorder** — an important new kind of order relation, defined as a binary relation that is reflexive, transitive, and total; total preorders arrange things into levels.