COSC462 Applied Logic
Lecture 1: Introduction

Willem Labuschagne
University of Otago

Abstract

In this lecture we introduce the idea that applied logic is *agent-oriented* and includes *commonsense reasoning*. Applied logic is motivated by a particular kind of situation, the *basic epistemic scenario*. After describing some examples realising this scenario, we develop a small but valuable knowledge representation language.

1 What agents want to know

An agent observes a system in order to find out what the state of the system is. The reason the agent wants to know the state is that this will influence decisions about the actions to be taken in pursuit of goals. Unfortunately, in most cases the agent is unable to verify by observation more than a fraction of the facts that characterise the state. Given limited information, what is it reasonable for the agent to believe about the state of the system?

Let's make the scenario more concrete.

Example 1 *Imagine an agent sitting in the control room of some complicated system. In front of the agent is the control panel, providing information about the components of the system. Unfortunately, the control panel shows the readings of a comparatively small number of sensors, and therefore does not give a complete picture of the state of the system. This is a potential problem, because some states of the system are safe and others are dangerous.*

In the corner of the room there is a big red lever. If the control room agent believes that the system is entering a dangerous state, then he must pull the lever. Several things will happen: the system will shut down, which is a very expensive business, and the supervisor will be

summoned, perhaps from the comfort of her home. If the control room agent pulls the lever unnecessarily, he may well lose his job; at the very least, his supervisor will be annoyed.

Clearly the agent wants to know the answer to the question: "What is the state of the system?" And clearly the agent does not want to make random guesses. Under ideal circumstances, the readings on the control panel would identify the state of the system unambiguously. But circumstances are often less than ideal, and then the control panel does not provide enough information to identify the state exactly. What then?

Common sense can try to make a difference. When the agent's information is limited, so that as far as he can tell the system may be in any of several different states, then he is sometimes able to use a default rule (a rule of thumb, a heuristic). For example, he may have noticed during years of lonely vigil that when valve X is open, it almost always means that cooling system Y is operating (although it could theoretically also mean that there is a leak in transmogrifier Z). When he now sees that valve X is open, he feels justified in believing, at least tentatively, that cooling system Y is operating. The agent could be wrong — under exceptional circumstances it may be that valve X opened because of the leak in Z. But he has taken no wild guesses and has simply tried to bring all his information to bear — the definite information provided by the control panel as well as the less definite, more uncertain, commonsense information contained in the default rule.

Even with the aid of a default rule, the agent may be unable to identify the state of the system unambiguously. But at least the agent will have narrowed down the possibilities as far as possible, and if the dangerous states are among those that have been eliminated, the agent will feel justified in not pulling the big red lever just yet.

This basic epistemic scenario arises in a wide variety of different contexts. The agent may be human or non-human, made of meat or of metal. Observation of the system may be carried out by living senses or artificial sensors, and the information involved may be represented by means of sentences or in some other way (*e.g.* by maps or other diagrams). The system may be a large natural system such as the Milky Way galaxy or a small artificial system such as a photocopying machine. The motivation for the agent to discover the state of the system may vary from abstract philosophical curiosity to a programmed 'desire' to help users control a photocopier by providing details of whether and where the paper is jammed.

What is characteristic of the basic epistemic scenario is that the agent's knowledge is usually limited and can usefully be supplemented

by information that is more tentative and reflects 'common sense'. Applied logic differs from traditional logic in accommodating such tentative commonsense information within the logical framework of syntax and semantics. In other words, applied logic = traditional logic + commonsense information.

Let us explore a second concrete example.

Example 2 *A man (the agent) emerges from his residence one evening and sees a motor car (the system) parked some distance away with its headlights on. The man is too far away to tell whether the engine is running. (Thus the agent's observation provides a limited amount of definite information about the state of the system, too limited to reveal unambiguously what the state of the system is.) The man feels a neighbourly concern at the possibility that the lights may drain the battery, leading to inconvenience for the car's owner. Thus the agent is motivated to discover whether the system is in the state characterised by having both the engine and the lights on, or in the state having the engine off but the lights on. The former state is safe, the latter 'dangerous'.*

Clearly all the elements of the basic epistemic scenario are present — the agent, the system, the incomplete information, the motivation to identify the state of the system. What happens next? How might this agent extract more information from the system?

There may be several courses of action available, for example to approach more closely and listen for the sound of a running engine. However, the agent is usually subject to constraints in the form of costs associated with each course of action. The man may, for example, be hurrying to an appointment, and will need to weigh up the advantages of delaying to investigate against the disadvantages of being late. Thus the agent may have to be content with a limited amount of definite information. But again, the definite information provided by sight may be supplemented by indefinite information provided by a rule of thumb. For example, the agent may be able, from the shape of the headlights, to identify the car as an Edison. Let us suppose that the most popular model of the Edison is known to have the feature that when the engine is switched off, the lights are automatically switched off too. It may not be clear, from a distance and in the dark, which model the car is, and it is possible that this particular car is one of the less usual Edison models which lack the safety feature. Nevertheless, in the absence of any indication that the car is unusual it seems sensible to assume that the car is of the popular type, from which it follows that the engine is running, and hence that the owner is in no danger of a flat battery. Having resolved his dilemma by commonsense reasoning based on a default

rule, the agent gratefully departs without feeling obliged to make a more thorough investigation.

Commonsense reasoning leads to conclusions that are *defeasible* (capable of being defeated by exceptional circumstances). Should the agent on returning at midnight see the same car in the same place being jump-started with the aid of a spare battery, his tentative beliefs about the car would turn out to have been mistaken and would need to be revised.

2 Representations

Every instance of the basic epistemic scenario involves an agent extracting information from a system. What is an agent? Consider the difference between a rabbit and a rock. Each is a persistent structure in a changing world, but they attain their persistence in different ways. A rock persists passively, by being hard — hard to eat, hard to dissolve, hard to bash into smithereens. An agent survives by interacting with its environment (the system) and changing it. A rabbit is an example of an agent, and we know how actively rabbits can change their environments. Agents may be biological or artificial. Living agents need not be animals — the mighty redwood is a particularly charming example of a non-animal agent. Artificial agents may be software (a spell-checker) or hardware (a thermostat) or both (a robot).

When an agent interacts with its environment, its behaviour has a pattern. A rabbit seeks food, builds shelter, has babies. It doesn't thrash about randomly unless something is very wrong with it, because the actions taken by an agent need to bear some relation to its environment. A thermostat changes its readings in accordance with temperature changes in the surrounding environment, and thereby causes things to happen (air to be cooled or heated). The agent's behaviour is adaptive because it responds to the environment.

What allows the agent to be responsive to the environment is that the agent has an internal *representation* of the environment. This representation is a substitute for the real system — a surrogate that can be consulted and manipulated by the agent. Representations vary along a spectrum from iconic (also called analog, or continuous) to symbolic (or digital, or discrete). The simplest agents have only iconic representations. More complex agents have both iconic and symbolic representations.

An *iconic representation* is one that somehow directly resembles or mimics the external system. For instance, a moth camouflaged to resemble the bark of the tree on which it rests bears on its surface an iconic representation of its environment. The face in a photo is an iconic representation of the person concerned. A landscape painting is an iconic

representation of what it depicts. Also iconic is the model airplane used by an aerodynamic engineer in a wind-tunnel — the model faithfully resembles the real thing in those aspects that are considered relevant. Sometimes the resemblance is not quite as direct as visual similarity but arises from a causal connection that leads the iconic representation to mimic what it represents. For example, the image of an object on the retina of an eye is a pattern of excitation caused by the object via reflected light, and constitutes an iconic representation of that object. Another example is a thermostat containing a bar made of two metals that expand at different rates when heated, so that as the temperature rises the bar echoes this by curving, and as the temperature falls the bar straightens out. The curvature of the bar is the thermostat's iconic representation of the relevant aspect of the environment.

A *symbolic representation* is one in which the resemblance between representation and environment is indirect, abstract, a matter of convention — for example, when we use letters of the alphabet to stand for components of a system, or give a description of the environment using the sentences of some language. Whereas the image of a friend on our retina is an iconic representation, the name that pops into mind after our brain has processed the image is a symbolic representation of the friend. It would be a bit of an over-simplification to claim that symbolic representations are just those involving language. A wedding ring is a symbol representing a mutual agreement. Languages themselves may be relatively more iconic or more symbolic. The change from Egyptian hieroglyphs, where a wavy line represented water, to words built up from an alphabet in which letters were associated with sounds by convention, was a change from more iconic to more symbolic representations. A formal language like mathematics is all the way over on the symbolic side of the spectrum. But, roughly, we may think of symbolic representations as descriptions in some language.

Every agent has at least an iconic representation of its environment, but the agents with which applied logic is concerned are those with symbolic representations as well, i.e. those with access to language. Why is language important? The obvious and crowning achievement of symbolic representation is that, in a world containing more than one agent, it serves to communicate information from one agent to another and thus to facilitate co-operative solutions to problems. However, communication is not the only thing language can accomplish. The crucial benefit of language is that it is discrete — it is built up of basic building blocks that may conveniently be rearranged, manipulated, and stored. Writing things down helps our rather vague memories and helps us to understand continuous processes by focusing our attention on key as-

pects. One may learn a lot more from hearing a tennis coach describe the key steps in hitting a topspin backhand than from simply looking at players hitting backhands (which is why coaches exist). Similarly, a set of instructions for folding a sheet of paper to produce a paper airplane may be much easier to understand than the folded paper airplane itself. Moreover, the following example, taken from Steven Pinker's book *The Blank Slate*, shows that language can be useful even in the absence of between-agent communication, audible speech, or writing:

> The mind makes use of a "phonological loop": a silent articulation of words or numbers that persists for a few seconds and can be sensed by the mind's ear. The loop acts as a "slave system" at the service of a "central executive". By describing things to ourselves using snatches of language, we can temporarily store the results of a mental computation or retrieve chunks of data stored as verbal expressions. Mental arithmetic involving large numbers, for example, may be carried out by retrieving verbal formulas such as "Seven times eight is fifty-six."

Although symbolic representations serve important purposes, we do not at all wish to suggest that iconic representations are useless. A blueprint may be very useful indeed to an engineer attempting to repair a system, and such a blueprint lies closer to the iconic side of the spectrum than a description in words or a more abstract diagram would. In general diagrams may be either more iconic, if they resemble the system fairly directly (like a topographical map that faithfully records mountains and valleys) or more symbolic (like the map of the London Underground). The change from iconic to symbolic representation can be regarded in some sense as a change from concrete to abstract. Often the concrete picture is just as important as an abstract symbolic version. Whenever you hear the word *semantics*, it will be about the relationship between a concrete iconic representation and an abstract symbolic representation — the iconic representation provides the semantics for the symbolic representation. Iconic representations are primary, formed by the processes governing input to the agent (perception). Symbolic representations are then (in a sufficiently high-level agent) constructed out of the iconic representations or out of other symbolic representations. Terrible problems arise when a symbolic representation has no iconic representation underlying it (i.e. has no semantics). In applied logic we are more concerned with the opposite (and less worrying) problem, which arises when an agent has an iconic representation of something

but no symbolic representation by which to communicate the information contained in the iconic representation. In a later chapter we call this the fundamental problem of applied logic. (The idea of calling this the 'fundamental' problem is that the much worse problem of having a symbolic representation without semantics is so bad we simply ignore it!)

The agents of applied logic are information-processing agents. What is information? Roughly speaking, the more possibilities the agent is able to rule out, the greater the amount of information the agent has about the system. An agent with complete information about the system can exclude all possibilities except for the actual state. An agent with severely limited information may be left with many candidates for being the actual state.

The *agent's-eye view* of the system is the information the agent is able to extract from the system. We distinguish between the agent and the logician. Remember the supervisor in the basic epistemic scenario? The supervisor checks whether the control room agent was correct to pull the big red lever. We may assume that the supervisor has a *god's-eye view*, in other words has complete information about both the system and the agent, so that the supervisor is able to judge whether the agent has formed a belief about the system that is accurate (true) or inaccurate (false). The supervisor is the logician. The logician is a superagent who knows everything about everything (but is unfailingly modest about it).

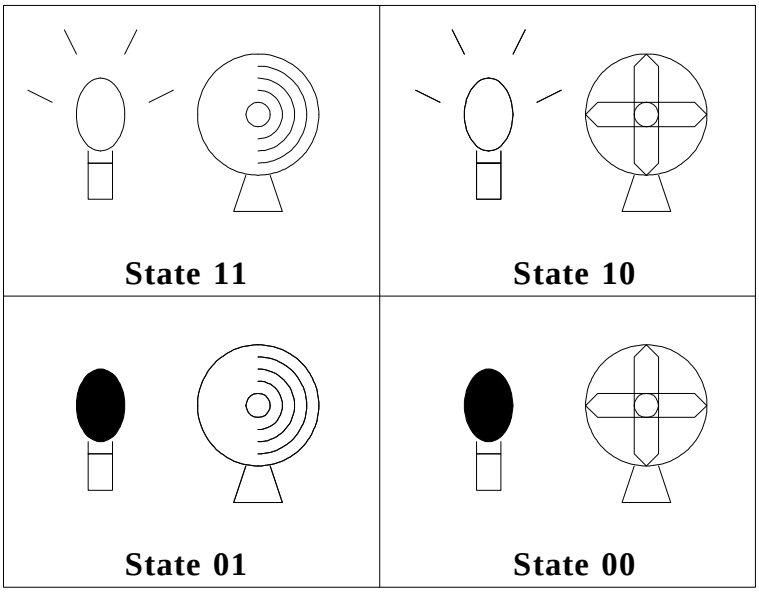
3 Semantics

Let us look at the way an agent may extract information from a system. We take the opportunity to describe a simple system that will be used repeatedly to illustrate concepts and techniques.

3.1 The Light-Fan System

The Light-Fan System has two components: a light and a fan. Each component may be either on or off. Let us assume that a state of the system is determined only by these facts, because the agent observing the system considers it irrelevant whether the components are painted pink or green, whether the light is provided by a bulb or a burning pine cone, whether the fan has three blades or two, whether the components are mounted as a joint unit on a ceiling or separately on a wall, and so on.

Given what the agent considers relevant, there are just four possible states of the system — the light and the fan may both be on, the light may be on but the fan off, the light off but the fan on, and lastly both components may be off. To start with, let us devise iconic representa-



tions for these states. We use pictures. Each pictures shows a lightbulb that is either shining or dark, and a fan that is either spinning or stationary. As convenient shorthand for each picture we use a binary string in which 1 indicates that the component is on while 0 indicates that it is off, and we agree always to mention the light first and the fan second. Thus the four possible states of the system are labelled 11, 10, 01, and 00 respectively.

If you think that this toy system is too simple to be interesting, reflect that the Light-Fan System may serve as a metaphor for such complex systems as a motor car or a nuclear power-plant, indeed for any system which has some part analogous to the light (a car's headlights, an atomic pile) and the fan (a car's engine, a cooling system) and whose other parts are not of interest at the time.

Assume that the Light-Fan System has both the light and the fan on, so that the system is in state labelled 11**.**

What information may an agent be able to extract from the system via its sensors?

If the agent is a human being near enough to see whether the light is shining and the fan is spinning, then he will be able to rule out the states labelled 10, 01, and 00, leaving 11 as the only remaining candidate for being the actual state (complete information). On the other hand, if the agent is some distance away, able to see that the light is on but unable to see (or hear) whether the fan is on, then he can rule out states 01 and 00 but is left with no way to choose between states 10 and 11 (limited information). And an agent so far away that he cannot see or hear the

system at all would be unable to exclude any of the possible states (no information).

3.2 Phases of information gathering

It is convenient to think of an agent's epistemic functioning as having three phases.

Suppose that the possible states of a system form a set S. For the Light-Fan System, $S = \{11, 10, 01, 00\}$. Initially, before the agent has acquired any information, all these states are candidates for being the actual state.

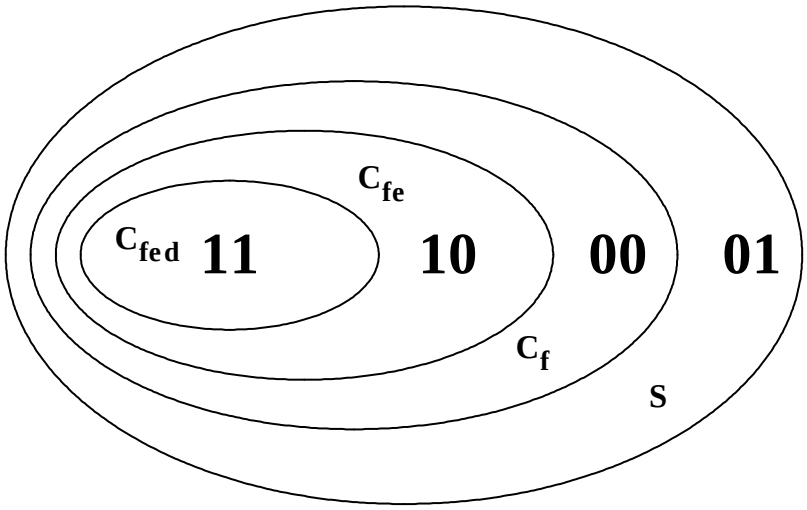
As a first step, the agent acquires his *fixed* information, for example by examining a blueprint or manual describing the system. This fixed information is reflected by the exclusion of zero, one, or more states, leaving a set $C_f \subseteq S$ of candidates for being the actual state. The amount of information is proportional to (or measured by) the set $\overline{C_f}$ of states excluded from C_f. Here we are using the notation for the *complement* of a set: given a universal set S within which to work, the complement of X relative to S is the set $\overline{X} = \{x \in S \mid x \notin X\}$.

For example, in the case of the Light-Fan System, suppose the agent studies a blueprint of the system which reveals that the light and the fan are connected in a way that makes it impossible for the fan to be on while the light is off. The agent's fixed information is represented semantically by ruling out 01 to leave the set $C_f = \{11, 10, 00\}$ of candidates.

As a second step the agent acquires *evidence*, by which we mean state-dependent information gained by observation or perhaps communicated by another agent. This is reflected by the further exclusion of states, leaving a set $C_{fe} \subseteq C_f$ of candidates. For example, suppose that the agent is able to see, from a distance, that the light is on, but is too far away to tell whether the fan is on. This evidence has the effect of ruling out 00 to leave $C_{fe} = \{11, 10\}$. We shall speak of the agent's fixed information plus his evidence as his *definite* information, and regard the definite information as being measured by the set $\overline{C_{fe}}$ of excluded states.

Finally, the agent may bring to bear a *default rule* (heuristics, statistical data, or a commonsense rule of thumb). We shall examine various kinds of default rules later on. For the moment it is sufficient to bear in mind that such default information is also reflected by the exclusion of zero, one, or more states, delivering in some way a set $C_{fed} \subseteq C_{fe}$. Keep in mind, though, that default rules admit exceptions and may lead the agent into error. We therefore think of default information as being *indefinite* rather than definite.

For example, suppose that the agent has noted, over a period of several months, that it is unusual to have the light on without having the



fan on (perhaps because there is a danger of overheating the system). This default rule, namely 'If the light is on, then normally the fan is on,' gives some tentative justification for excluding the state 10, leaving $C_{fed} = \{11\}$. Lest this successful narrowing down of possibilities create the impression that an agent's life is an easy one, note that if the system were actually in the (exceptional) state 10, then every step of the process would proceed in exactly the same way, including the final (and now incorrect) exclusion of the actual state 10 to leave the remaining candidate 11. However, mistakes such as this occur only when the system is in a state that, according to the agent's default information, is unusual or exceptional.

The successive restrictions of S may be visualised in terms of the diagram.

To summarise: the agent builds a semantic (iconic) representation of the state of the system, first on the basis of definite information (*i.e.* fixed information plus evidence) and then by further refinement based on indefinite information suggested by default rules (which may in exceptional situations lead to wrongly ruling out the actual state).

The next big question is: Can the agent's information be expressed by the sentences of some suitable language?

4 The agent's symbolic representations

Consider an agent observing the Light-Fan System. Since the state of the system is completely determined by whether the light is on and whether the fan is on, we may imagine the agent wanting to note down these basic facts in a little notebook. The agent writes down a sentence like

'The light is on' or a sentence like 'The fan is on', as the case may be. If the agent were to monitor the system for a long time, he may find it convenient to abbreviate the basic facts, for example by writing p as an abbreviation for 'The light is on' and q for 'The fan is on'. These two *atoms* p and q are the first step toward the creation of a specialised knowledge representation language.

From time to time the agent may need to express an idea that the atoms alone cannot express. Imagine that he cannot see the light or the fan and is restricted to a control panel connected to a sensor for detecting variations in temperature. This enables him to detect whether components are on, and if the sensor is sensitive enough, to detect whether it is a single component or both that are on. For example, we may imagine that when one component is on, the surrounding temperature is raised by one degree, and when both are on, by two degrees. Now there is a difficulty. Suppose the sensor detects an increase in the temperature of one degree. How can the agent record that exactly one component is on, if he cannot tell which component it is? Clearly neither atom will express this rather non-specific idea.

The solution is to equip the new knowledge representation language with *connectives*. These may be used to build, from the atoms p and q, sentences that reveal more about the agent's grasp of the situation. We shall equip the agent's language with a fairly standard set of connectives, represented by the symbols $\neg$ (called the negation symbol), $\wedge$ (the conjunction symbol), $\vee$ (disjunction), $\rightarrow$ (the conditional), and $\leftrightarrow$ (the biconditional). Grammatically well-formed sentences of the p, q-language may be constructed as follows.

(Atoms) The simplest well-formed sentences are the atoms: p and q. As it happens, we know how the agent observing the Light-Fan System wishes to interpret these atoms, but from the standpoint of grammar these intended meanings are irrelevant. It is conceivable that the language we are describing could be used by another agent observing another system, in which case the meanings of the symbols p and q would be different but the grammar the same.

(Negation) Suppose α (the Greek letter alpha) is any well-formed sentence of the p, q-language. (Initially, there are only two things α could be: the atom p or the atom q. Of course, there will soon be more, because the rules of our grammar generate an unending supply of new sentences.) For every well-formed sentence α, the negation of α, namely $(\neg\alpha)$, is a well-formed sentence too, and we read it as 'It is not the case that α'. Some examples of sentences that can be formed by this rule: $(\neg p)$, $(\neg(\neg q))$.

(Conjunction) Suppose both α and β (the Greek letter beta) are

well-formed sentences of the p, q-language. Then the conjunction of α with β, namely $(\alpha \wedge \beta)$, is well-formed, and is read as 'α and β'. We say that α and β are conjuncts of the conjunction $(\alpha \wedge \beta)$. Some examples of sentences that can be formed by the rules given so far: $(p \wedge q)$, $((\neg p) \wedge q)$, and $\neg(p \wedge (\neg q))$.

Notation 3 *We often omit parentheses for ease of reading, but must beware of ambiguity. This is made easier if we adopt the convention that* $\neg$ *applies to the shortest well-formed sentence following it, so that* $\neg p \wedge q$ *is* $((\neg p) \wedge q)$ *rather than* $(\neg(p \wedge q))$.

(Disjunction) Suppose α and β are well-formed sentences. Then so is $(\alpha \vee \beta)$, read 'α or β or both'. We say that α and β are disjuncts of the disjunction $(\alpha \vee \beta)$. Examples are $(p \vee q)$ and $q \vee (\neg p \vee q)$, where we have omitted several parentheses from the second example.

(Conditional) If α and β are well-formed, so is $(\alpha \rightarrow \beta)$, read 'if α then β'. We say that α is the antecedent and β the consequent of the conditional sentence $(\alpha \rightarrow \beta)$. Examples: $q \rightarrow p$, $\neg p \rightarrow (\neg p \vee q)$. Resist any temptation to read $\alpha \rightarrow \beta$ as 'α implies β', since the word 'implies' is ambiguous in a way that will become clear when we discuss the concept of entailment, and you should begin right now to cultivate good mental hygiene by expunging the word 'implies' from your vocabulary.

(Biconditional) If α and β are well-formed, so is the biconditional sentence $(\alpha \leftrightarrow \beta)$, read '$\alpha$ if and only if β'. Examples: $p \leftrightarrow \neg p$, $(p \rightarrow q) \leftrightarrow (\neg p \vee q)$. Resist sternly any temptation to read $\alpha \leftrightarrow \beta$ as 'α is equivalent to β', because we shall be using the word 'equivalent' to mean something else. Note that the English phrase 'if and only if' is often abbreviated 'iff'.

Given these connectives, it is now possible for an agent to construct a sentence expressing the complex notion that exactly one component of the Light-Fan System is on, even though she may not know which component it is: $(p \vee q) \wedge \neg(p \wedge q)$. In English we read this as: p or q, but not both. Do you find it easy to match the symbolic rendering with the English reading? Let us say a bit more about the connection.

For each connective we have suggested an English equivalent. For instance, $\wedge$ is like 'and'. But the English versions are more subtle and their meanings may vary according to context. Were we to say 'Today the patient ate an apple and two carrots', then the chances are that we would be describing the same situation as that reported by 'Today the patient ate two carrots and an apple'. The context suggested by these sentences is one in which the order of events is not important. But if we were to say 'The patient suffered severe indigestion and ate two carrots' we might be describing a very different situation from that reported

by 'The patient ate two carrots and suffered severe indigestion'. The context now seems to be one in which the order of events is significant. The former sentence hints that the patient had indigestion and hoped to relieve it by eating a certain fibrous vegetable, whereas the latter sentence suggests that the carrots preceded and may well have caused the indigestion. The word 'and' has here been used to mean 'and then'. Another form of context-sensitivity is the use of 'but' instead of 'and' in circumstances that involve the contrasting of ideas.

We do not want the agent's knowledge representation language to be context-sensitive the way English is. For the knowledge representation language we want to single out a specific sense of 'and' (as well as each of the other connectives) and use it in that sense always. In order to spell out the meanings of the connectives, we need to use *truth values*. The next section shows how.

Exercise 4 *1. Give English paraphrases for each of the following sentences of the p, q-language. (It helps to build insight if you first give a paraphrase that mimics fairly closely the symbols used, and then reflect on a shorter, perhaps less formal, way to express the essential idea.)*

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

2. Construct sentences of the p, q-language to paraphrase each of the following.

- *The light is on or the fan is off.*
- *The light is on or both components are off.*
- *At least one of the components is on.*
- *At most one of the components is on.*
- *Exactly one of the components is on.*
- *The fan is on if the light is on.*
- *The fan is on only if the light is on.*

5 Truth

Imagine that an agent is observing a system and using sentences of the p, q-language to represent and communicate his knowledge. Imagine that you, the all-knowing logician/supervisor, are observing both the agent and the system from a higher level. You know all about the agent's epistemic functioning and the actual state of the system. How can you indicate that one of the agent's sentences correctly describes (some aspect of) the state of the system?

The flag by which the logician marks the accuracy of a sentence is called a *truth value*. For our purposes two truth values will suffice, and we shall use the numbers 1 and 0 for the purpose. The logician is able, relative to each state of the system, to assign one truth value to every sentence. Intuitively, when the logician associates 1 with a sentence α, then the logician is indicating that α is true relative to the state of the system, *i.e.* that α is accurately describing (some aspect of) the actual state of the system. When the logician associates 0 with α, then this indicates that α is not true but instead false relative to the current state. **Note that a sentence α only has a truth value relative to a state.** It is a solecism to assert of a sentence merely that it is true or that it is false, unless the context makes it very clear what the relevant state of the system is.

The four possible states of the Light-Fan System were iconically represented by the pictures labelled 11, 10, 01, and 00. The choice of labels was no coincidence. The knowledge representation language used by the agent observing the Light-Fan System has two atoms, p and q, and it is understood that these express the elementary facts that the light is on and that the fan is on respectively. It makes sense to use the string 11 as shorthand for the state in which the light is on and the fan is on, because the sentence p is true in this state and so is the sentence q. Similarly, the state called 10 is that in which the light is on (so p is true) and the fan is off (so q is false); 01 is the state in which the light is off (p is false) and the fan is on (q is true); 00 is the state in which the light is off (p is false) and the fan is off (q is false).

We see that the logician assigns truth values to atoms according to a slightly mysterious process that takes into account their meanings and the state of the system. Later we shall look more closely at this mysterious process. For now, assume that every atom is given a truth value.

The agent's knowledge representation language contains many sentences, not just the atoms p and q. How should the logician assign truth values to the remaining sentences? As with atoms, it depends on the state of the system. After all, every sentence is built up from the atoms,

and if the atoms vary their truth values according to the state we must expect a similar variation on the part of the compound sentences. What is new is that the connectives also exert an influence, as illustrated in the truth tables below. Given this specification of how the connectives change truth values, the logician (which is to say, you) is able to take any sentence of the language and systematically calculate its truth value relative to any state of the system.

Consider negation:

α	$\neg\alpha$
1	0
0	1

Here we see a handy device — a table that conveniently displays the particular way in which we want the connective to influence the allocation of truth values.

The truth table summarises our understanding that a sentence and its negation are 'opposites'. In accordance with this intuition, the rows of the table tell us that if α were true relative to some state (i.e. had the truth value 1) then $\neg\alpha$ would be false relative to that state (i.e. would have the truth value 0), and that if α had truth value 0 relative to some state, $\neg\alpha$ would have truth value 1 relative to that state.

Let's apply our understanding of negation. Suppose the system is in state 10 and consider the sentence $\neg p$. Is $\neg p$ true relative to state 10? Since we know that the truth value of p relative to state 10 is 1, we know by virtue of the table (with p playing the role of α) that the truth value of $\neg p$ is 0 relative to state 10, so $\neg p$ is false in state 10. Continuing in this vein, we can claim that relative to the same state 10 the truth value of $\neg q$ is 1 (where q plays the role of α in the table, and the label 10 tells us that q has truth value 0). Similarly the truth value of $\neg\neg p$ is 1 (where $\neg p$ plays the role of α), and the truth value of $\neg\neg q$ is 0 (where $\neg q$ plays the role of α). For more interesting sentences than these, we need to apply our understanding of the other connectives.

Consider conjunction:

α	β	$\alpha \wedge \beta$
1	1	1
1	0	0
0	1	0
0	0	0

The table displays our understanding that the only way to make $\alpha\wedge\beta$ true is to make both of the original sentences true. In all other cases, $\alpha \wedge \beta$ has the truth value 0. This is interesting, because a moment's

thought now reveals that $\alpha \wedge \beta$ always has exactly the same truth value as $\beta \wedge \alpha$ does, and so $\wedge$ is not sensitive to the order of the conjuncts.

Let us calculate the truth value of the sentence $\neg(p \wedge \neg q)$ relative to state 00, say. To begin with, we know that p and q are both false relative to state 00. So p has truth value 0 and $\neg q$ has truth value 1, so that (by the table for conjunction, with p playing the role of α and $\neg q$ playing the role of β) sentence $p \wedge \neg q$ has the truth value 0. So $\neg(p \wedge \neg q)$ has the truth value 1, i.e. is true relative to state 00.

Consider disjunction:

α	β	$\alpha \vee \beta$
1	1	1
1	0	1
0	1	1
0	0	0

The truth table for $\vee$ indicates that $\alpha \vee \beta$ is false if and only if both of the original sentences α and β are false. In particular, $\alpha \vee \beta$ is true when the disjuncts α and β are simultaneously true. This tells us that $\vee$ corresponds to the inclusive sense of 'or' as in the phrase '... or ... or both' and not to the exclusive sense of 'or' that occurs in the phrase 'either ... or ... but not both'.

Let us try to calculate the truth value of $p \vee \neg p$ relative to state 01, say. Well, p is false in this state, so $\neg p$ is true. Letting p play the role of α and $\neg p$ the role of β, the second-last row of the table tells us that $p \vee \neg p$ gets the truth value 1.

The conditional:

α	β	$\alpha \rightarrow \beta$
1	1	1
1	0	0
0	1	1
0	0	1

The truth table for $\rightarrow$ indicates that $\alpha \rightarrow \beta$ is false if and only if α is true but at the same time β is false. The idea is that $\rightarrow$ behaves the way 'if ... then ...' does in a promise, for example 'If I pass my swimming test then I'll take you out to dinner'. The only way in which this promise could be broken is for me to pass the swimming test but then not to take you to out dinner. If I don't pass the test, then I can take you to dinner or not, as I choose, and neither option will break my promise. As an application we calculate the truth value of $q \rightarrow q$ relative to the state 10. First we note that q is false relative to this state. Now let q play the role of α and of β in the table, then the bottom row tells us that $q \rightarrow q$ is true.

Finally, the biconditional:

α	β	$\alpha \leftrightarrow \beta$
1	1	1
1	0	0
0	1	0
0	0	1

We see that $\alpha \leftrightarrow \beta$ is true whenever α and β have exactly the same truth value and is false whenever the truth values of α and β differ. Perhaps the easiest is to think of a mutually binding contract. The idea is that the contract 'I'll teach you to cook onion soup if and only if you'll teach me to swim' imposes an obligation on each of us. If we both fulfil our obligations (I teach you to cook onion soup and you teach me to swim), then everyone is happy and the contract is satisfied. On the other hand, if I decide not to teach you to cook onion soup, and you simultaneously decide not to teach me to swim, then we have in effect agreed to dissolve the contract and neither of us has any grounds for complaint. However, if one of us does our duty and the other does not, then the contract has been broken.

As an application, let's calculate the truth value of $p \leftrightarrow (\neg q)$ relative to state 11. We know that p and q are both true in this state, so that $\neg q$ is false. Letting p play the role of α and $\neg q$ the role of β, we see that $p \leftrightarrow \neg q$ gets the truth value 0.

To summarise: the binary strings we have used as labels for states of the Light-Fan System make it obvious which atoms are true in which state. Given this start, and having seen how each connective modifies truth values, we are in a position to calculate the truth value of every sentence relative to each state.

Exercise 5 *1. Find, for each sentence α below, the states in $S = \{11, 10, 01, 00\}$ at which α is true:*

- $p \wedge \neg p$
- $p \wedge q$
- $p \wedge \neg q$
- $\neg p \wedge q$
- $\neg p \wedge \neg q$
- p
- q
- $p \leftrightarrow q$

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

2. *For each of the four states, write down one sentence that is false in that state.*

6 Glossary

We introduced the following terms and symbols.

- **agent** — a not-necessarily-human entity capable of extracting information from some system.
- **atom** — a short simple sentence that can be used as a building block to construct longer compound sentences.
- **candidate** — a state left after an agent has excluded those she has reason to think are not the actual state.
- **connective** — a symbol like $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$ which can be used to build new sentences from old.
- **default rule** — heuristic information such as the rule of thumb "Cars normally stop at red lights". We have not yet said how to represent a default rule.
- **evidence** — the word we sometimes use for the state-dependent information that an agent may acquire by observation or by communicating with another agent. Evidence is assumed to be accurate as long as the relevant state of the system endures, but when the system changes state the evidence needs to be updated.
- **fixed information** — information obtained from a system description or blueprint and regarded as being always true of the system.

- **iconic representation** — picture or other representation that resembles or is quite directly linked to the thing it represents; contrasted with symbolic representation.

- **information** — information is represented semantically by dividing the set of states into two subsets, one of which consists of the excluded states and the other of states that are candidates for being the actual state of the system. The insight that information is determined by the exclusion of possibilities goes back to the philosopher Karl Popper — see Popper KR: *The Logic of Scientific Discovery* (3rd ed) Hutchinson 1972. The idea was formalised in Bar-Hillel Y & Carnap R: 'Semantic information' *British Journal for the Philosophy of Science* **4**:147-157 1953. This idea should not be confused with the statistical measure associated with Shannon's work in communication theory, and a very good explanation of the difference is given by Bar-Hillel Y: 'An examination of information theory' *Philosophy of Science* **22**:86-105 1955.

- **semantics** — giving meaning to the sentences of a language, basically by associating some iconic representation such as one of the states labelled by $11, 10, 01$, or 00 with sentences of the language. Truth is about the match-up that results.

- **state** — informally, when a system persists for a reasonable length of time exhibiting the same collection of properties, we say that the system is in a 'state' characterised by those properties. We shall have much more to say about the representation of states in the next chapter.

- **symbolic representation** — description in sentences, or some other abstract representation in which the connection with the thing represented is conventional.

- **system** — a part of the universe that is of interest to some agent.

- **truth value** — a flag or value that indicates how well a symbolic representation matches an iconic representation, in other words a label attached to a sentence to indicate whether the sentence accurately describes some aspect of a given state of the system.