COMP1009 COGNITIVE SYSTEMS AND INTELLIGENT TECHNOLOGIES
Symbolic and Subsymbolic Approaches to Artificial Intelligence

Artificial intelligence (AI) has been defined to be the science of getting machines to do things that were they to be performed by a human would be considered to require intelligence.

However there are different definitions and forms of natural intelligence, and different forms of artificial intelligence are appropriate when trying to develop systems that can be effective in these areas.

Symbolic AI

This stores knowledge in the form of verbal rules. These rules have usually been extracted from human experts who are able to solve the task in question, though there are examples of systems with a degree of 'learning' that can generate more complex or appropriate rules from verbal rule fragments. However obtained, though, the end result is the same: a set of rules of the form

IF (A is true) **THEN** (B is true)

which can then be used within the machine’s model of reasoning to give answers to users' queries (for example 'what disease might this person have, given their symptoms, and how should it be treated?').

In such an 'IF-THEN' statement, the IF part is called the premise and the THEN part is called the conclusion.

Backward chaining (deduction) starts from a likely conclusion (ie it assumes that B is indeed true) and seeks evidence that A is true. If it can find such evidence then the initial assumption, that B was true, is correct. Deduction can be regarded as top-down reasoning.

Expert systems, to be looked at in more detail later, are the clearest example of the use of backward chaining in AI.

Forward chaining (induction) in contrast considers the evidence presented to the system (ie, it assumes now the truth of the premise A) and tries to find the evidence-dependent rule (the conclusion B) that best fits these facts, sometimes 'firing' a *set* of possibly-appropriate rules and leading to a corresponding set of conclusions, weighted by their probabilities. Induction can be regarded as bottom-up reasoning.

Expert systems and other rule-based AI technologies can use both forward and backward chaining, but are mainly based on deduction.

Playing chess has long been regarded as a paradigm of intelligent behaviour and is an area in which rule-based or symbolic AI systems have had notable success.

Many people were astonished when in 1997 IBM's Deep Blue chess computer beat international grandmaster Gary Kasparov in a six-game match. Some regarded the AI system's victory over Kasparov as a proof that 'thinking machines' were on the horizon, if not here already, and that human intelligence would soon be outmatched; others, however, were more sceptical.

In the previous year it had been Kasparov who had beaten the computer, prompting well-known AI researcher Drew McDermott to tell his class that it would be "many years" before computers could could challenge the most skillful human players. However when Deep Blue won the following year, McDermott pointed out that the system would be unable even to recognise a chess piece, still less carry on a conversation about the game it had just won (McDermott, 1997). Deep Blue's 'intelligence,' he claimed, was extremely narrow in scope.

Moreover it might be said that chess -- and games-playing in general -- is a rather contrived arena for natural and artifically intelligent systems to compete in. Life, unlike chess, does not always yield to a logical and systematic consideration of available actions and their downstream consequences.

Although logical, articulate reasoning is certainly very important, other modes of intelligence, relying on unconscious pattern recognition -- 'intuition' -- are equally vital to our survival.

These other forms of intelligence, which we probably share with many other animals, are better captured by low-level subsymbolic approaches such as neural networks.

Subsymbolic AI

In this approach the 'knowledge' in the system is encoded not as a set of verbal rules, but as a set of numerical patterns that affect the way that an output is computed given a certain input.

In contrast to deductive rule-based systems subsymbolic AI can be said to be primarily using induction, as conclusions are reached via a bottom-up process in which a data-driven learned rule is applied to low-level 'sensory' (eg image data) input.

The best known subsymbolic systems are ***neural networks**,* in which the encoded patterns are in the form of the strengths and signs of connections between simple neuron-like objects. These connection strengths, or weights, are acquired by a learning process that only requires the person using the system knows what output should be produced for which input(s), not the rule that might underpin this.

Neural networks (and other subsymbolic systems such as ***genetic algorithms***, based on ideas of Darwinian evolution, and ***particle swarm optimisation***, based on observations of bird flocking and other social behaviours) do, after their learning phase, contain rules in the general sense of a lawlike association between inputs and outputs, but these are not verbally expressable.

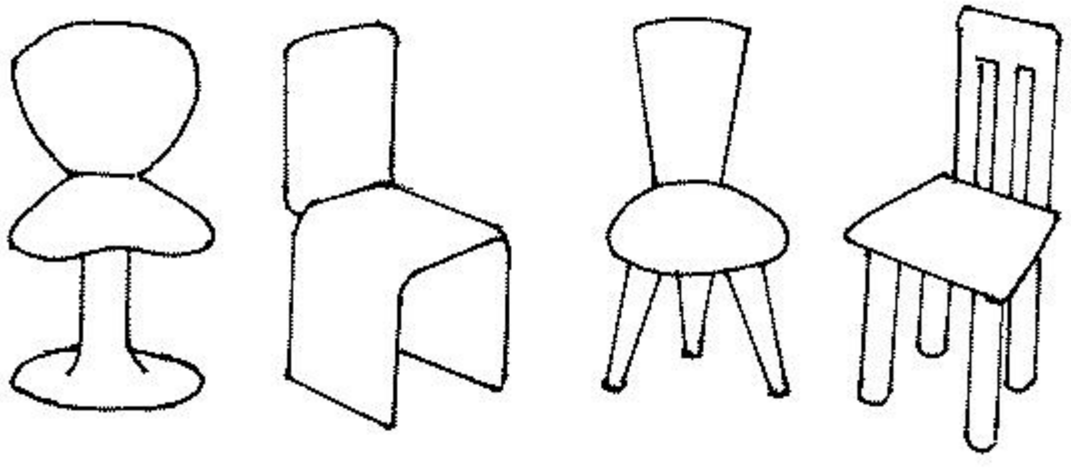
The fact that the 'rules' the system learns are only visible as internally-generated patterns of neural firing signals may be a disadvantage in some situations, but the strength of these systems is that they can learn what cannot be verbally expressed.

Applying symbolic and subsymbolic AI systems: practical problems and limitations

While we might prefer to have AI systems that operate with explicit rules and can give explanations, for many problems this just isn't practical.

Consider the problem of constructing a set of rules that could identify a chair.

Over-specific requirements such as ‘a chair has four legs’ will not do:



It isn’t relevant either whether a chair has arms, but it *does* seem relevant whether it has a back -- a chair without a back isn’t just 'a backless chair', in the English language at least it is a differently categorised object, a stool.

(Notice how, for humans at least, concept definitions are closely entangled with *language*.)

How about the following chair definition?

A chair

- is a portable object
- has a horizontal surface at a suitable height for sitting
- has a vertical surface suitably positioned for leaning against

Something like this might be a reasonable starting point -- although ideas like 'at a suitable height for sitting' would need to be firmed up. And do those words 'horizontal' and 'vertical' need to be *strictly* true...?

This is the kind of thing AI researchers trying to develop systems that encode 'common sense' or 'everyday' knowledge have had to think about. However some everyday concepts are harder to define than others.

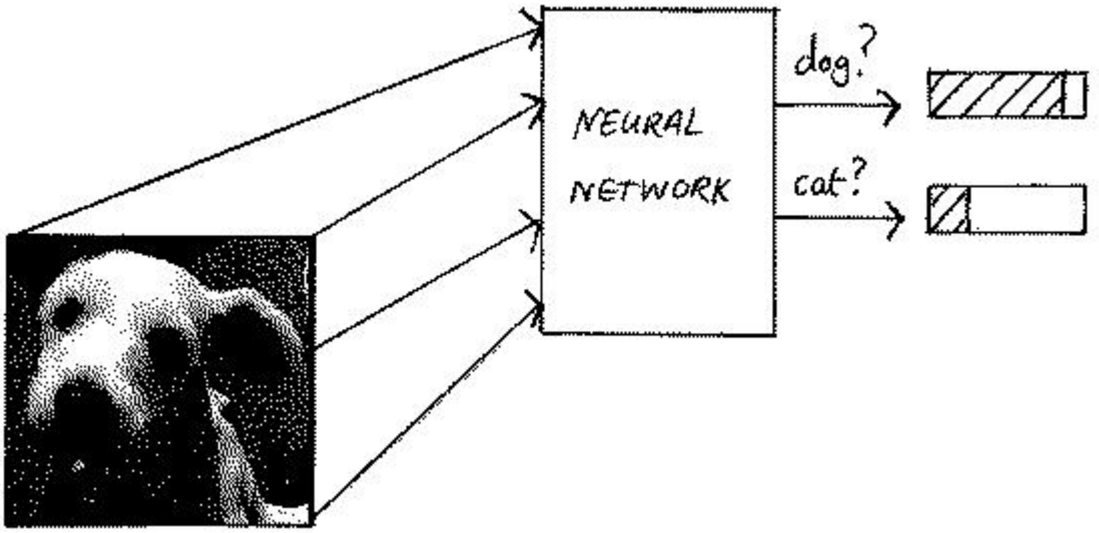
For example, what exactly is it that distinguishes cats and dogs?

(Obviously if one allowed facts like 'dogs bark' and 'cats miaow' to be included, it would be easier, but let's imagine we have to distinguish them on the basis of *image alone*).

We can tell cats and dogs apart quite easily. But is it possible to formulate a verbal definition, of the kind used for 'chair,' that could separate the two animal types?

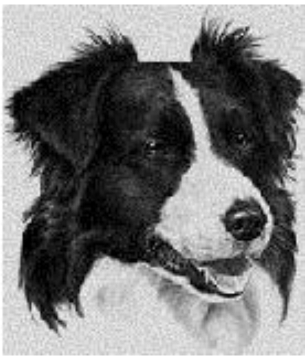
Certainly everything I can think of that cats possess -- four legs, fur, a tail, forward-facing eyes... -- dogs also, to some degree, possess... there’s *something* that distinguishes them, probably some combination of their various features, but *what*...?

The difficulty one has in articulating an explicit verbal definition which could be used to decide the cat-or-dog problem makes it alternatively seem a good candidate for a subsymbolic system such as a neural network:



The net is trained using a set of images clearly labelled as 'cat' or 'dog'; its aim is to develop, at a subsymbolic, numerical level, an internal representation that links the essential features of a cat or dog to an output code designating the type of animal. It’s not sufficient that the net store a link between each example and its target separately; in order for true category learning to have taken place -- which could be expected to generalise to new examples of cat or dog pictures -- the net must indeed have extracted some 'typical' features on which to base its judgement.

The trained neural network is then presented with a test set:



?

?

?

Would all of these have been categorised correctly as dogs?

Suppose all the dogs in the training set of pictures had been wearing collars. It’s then possible the neural network might have learned the rule 'dogs have collars, cats do not,' so that the first testing example above would have been classed as a cat!

Note that because this is a subsymbolic system the 'rule' referred to would not be directly visible and accessible to the person who had developed the classifier network, only implicitly present as a complex pattern of connection weights. The incorrect rule would only be discovered if the trainer then put in some plausible examples of dogs that were not categorised correctly and suspected because of this that something had gone wrong.

A case like this happened during the early years of neural network applications in the 1980s. A system was developed that was intended to detect from photographs (input to the net as bitwise images, one grey-scale value per pixel) whether or not there were tanks hiding amongst trees. The net found this problem extremely easy to learn to solve -- in retrospect this alone should have tipped off the developers that something might be wrong -- but, very embarassingly when it was being demonstrated to the project's sponsors on a new testing set of data the system proved to be performing no better than chance.

What had gone wrong?

It turned out all the pictures in which there were tanks amongst the trees had been taken on a sunny day, all those in which there weren't on an overcast day. All the net had actually learned was to separate patterns based on overall illumination level, the average grey-scale value per pixel, which would be an easy problem even for a single model neuron let alone a net of them, so no wonder it took no time to train!

Nowadays something like this would be much less likely to happen. It’s now appreciated that the selection of appropriate (sufficiently wide-ranging and representative) training data is as much a part of the training process of a neural network as the actual business of modifying the internal weight parameters.

But one can never be *absolutely* sure that there isn’t some subtle bias in the training data that would mean it would not perform correctly on new data. It’s this, combined that the fact that a neural net is always to some extent a 'black box' because it doesn’t make its numerically-represented rules easily accessible to a human user, that make neural nets less suitable for safety critical applications like medical diagnosis or ones for which for legal reasons a verbal justification of a decision might be required.

(The latter is true for example in the case of systems in the US which decide whether to give loans. A person refused a loan has there the right to request an explicit reason, so if the decision has been made by a neural net -- and many in fact now are -- a human loan adviser has to be on hand to give some sort of alternative justification.)

In summary, until we know more about the way natural computation is performed in the brain it will usually be necessary to choose either a subsymbolic (eg neural network) or symbolic (eg expert system) approach to AI, depending on the problem area and requirements.

Choose a subsymbolic system if

- we can solve the problem ourselves, but can't explain precisely how (eg face recognition)

or

- we can't solve the problem very well at all (eg predicting the movement of financial markets)

and

- it's not essential that the system can produce a verbal justification of its decisions.

Choose a symbolic (rule-based) system if

- we (or at least some of us) can solve the problem, and can also explain how

and

- it's important that a decision can be explicitly backtracked when a judgement is in question (eg medical diagnosis or legal opinion).