COMP310 Tutorial Week 5

Dr. T. Carroll

2020-02-{27,28}

Exercise 1 - Video set 2 (2014-15, Q1)

An agent has been described as *"... a computer system that is situated in some environment, and that is capable of autonomous action in that environment in order to meet its desired objectives..."* As such, it can be modelled abstractly.

(a) Explain what is meant by a *predicate task specification*, and how such a specification relates to utility function over runs. (**5 marks**)

(b) Explain what is meant by an *achievement* task. (**5 marks**)

(c) Explain what is meant by a *maintenance* goal. (**5 marks**)

(d) According to McCarthy, the *intentional stance* can be used to explain and predict the behaviour of machines. However, it is not always useful to do so. Give a brief explanation of what is meant by the intentional stance, and explain when and why it is useful to use the intentional stance of a machine. (**10 marks**)

SOLUTION

(a) A predicate task specification is a specification that assigns boolean utilities to agent runs, depending on whether the agent succeeded in its task for a run (i.e. the utility is 1) or if it failed (i.e. the utility is 0). A predicate task specification Ψ is given as

$$\Psi : \mathcal{R} \rightarrow \{0, 1\}$$

Thus, we can say that a run r that succeeds for a predicate task specification is given as $\Psi(r) = 1$.

(**5 marks**)

(b) An achievement task is specified by a set G of "good" or "goal" states: $G \subseteq E$. The agent succeeds if it is guaranteed to bring about at least one of these states (we don't care which, as all are considered good). The agent succeeds if it can force the environment into one of the goal states $g \in G$.

(**5 marks**)

(c) A maintenance goal is specified by a set B of "bad" states: $B \subseteq E$. The agent succeeds in a particular environment if it manages to avoid all states in B; i.e. if it never performs actions which result in any state in B occurring. In terms of games, the agent succeeds in a maintenance task if it ensures that it is never forced into one of the fail states $b \in B$.

(**5 marks**)

(d) The intentional stance is a view of the world whereby one ascribes notions of belief, free will, intentions, consciousness, abilities, or wants to a machine in a similar way to that one would use when describing a person. It is useful when the ascription helps us understand the structure of the machine, its past or future behaviour, or how to repair or improve it. It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of the machine in a particular situation may require mental qualities or qualities isomorphic to them.

The advantage of taking such a view is that it allows us to abstract what is happening at the low level within a machine, and refer to it using the same abstractions one uses for a human. So instead of describing knowledge a machine has as data stored within a data-structure, or concerning ourselves with the format, we can simply think of that knowledge as a set of beliefs. Whilst this is not always useful - for example it would not necessarily make sense to describe the actions of a light switch with the intentional stance, because humans are typically familiar with its inner workings, the intentional stance is a powerful abstraction tool. It is a convenient way of talking about complex systems, which allows us to predict and explain their behaviour without having to understand how the mechanism actually works.

(**10 marks**)