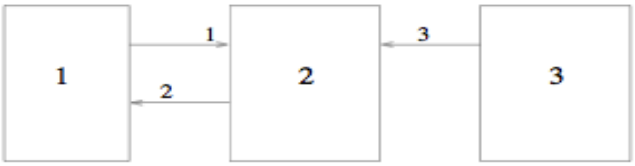
Logicist AI in Mcarthy and Hayes (1969)

By Matt Paisner

The project of logicist AI is to create a framework, or a language, within which a computer program can manipulate facts about the world. In their 1969 paper, McCarthy and Hayes present such a framework. This work is built on longstanding philosophical ideas about logic and knowledge, but attempt to adapt them to the demands of AI. Specifically, the authors attempt to formalize many ideas and concepts that are vague or incompletely defined in natural language, such as an agent's possession of knowledge or the ability to perform a certain action. Much of this summary is devoted to outlining the structure of these formalizations.

For an AI to be an agent, it has to be able to generate a set of responses to a particular situation and pick one to follow. However, a computer program is by definition deterministic (somewhat paradoxically, even random number generation is deterministic). Therefore, what it means for an agent to be able to choose an action is not immediately obvious. McCarthy and Hayes introduce an interpretation of an AI as an interconnected set of discrete finite automata that allows them to define this and several similar notions in useful ways. Consider the following diagram from their paper:



In this framework, each automaton is represented by a numbered block. The state of an automaton is a function of its previous state and the inputs to it, represented by arrows. The outputs from an automaton are a function of its previous state. An AI would be represented by a much more complex assortment of automata, which would in turn be connected by arrows (input and output) to the world.

Now, it is possible to define a notion of “can” that is both precise and stronger than simply saying that a deterministic agent can do only what it actually does or will do. If block 1 represents the agent under consideration, then the states of the world that it can bring about can be said to be the ones that are possible under any sequence of values for all the outputs from 1. In other words, if in agent 1 can generate output 0 or 1, where, in state s, output 0 puts agent 2 in state 0 and output 1 puts agent 2 in state 1, then in state s agent 1 can bring about state 0 or state 1 in agent 2. Notice that this is true even if one of these states will never actually happen, for example if agent 1 is programmed to always produce output 1 in state s.

McCarthy and Hayes go into more depth on this subject, including the question of whether an agent needs to be able to bring about a state under all possible combinations of external input/output, only some, or only the one that will actually occur. They also discuss the fact that what an agent “can” do depends on where the output of the system is defined to be – in a robot, is the output the movement of its limbs or the signals coming from its brain? The answer will determine which automaton we examine the output of, and therefore what we decide the agent can do. McCarthy and Hayes argue that neither of these interpretations is necessarily better or worse, and that we should and do use several different

versions of the idea of “can.”

The automaton model is also used to gesture at a concept of knowledge, but it is not well fleshed out and is used more as a lead-in to the next section of the paper. Here, the authors lay out the heart of their work, and the core of logicist AI: a well-defined syntactical system that allows facts about the world to be defined and used in constructing more facts. In essence, this is similar to a basic predicate logic, but it includes several changes and refinements that are intended to make it more suitable for use in inference about the world.

One building block of this system is the idea of a situation. A situation is essentially a snapshot of the universe – it is intended to represent everything that exists or pertains at a given moment in time. Of course, no situation can ever be completely defined, but the concept is still a useful tool for reasoning. Once situations are defined, it is necessary to have a way to talk about A) What facts are true in a given situation and B) How to transition between different situations. McCarthy and Hayes introduce the idea of a “fluent” to address both of these issues. A propositional fluent is a function whose domain is the set of all situations and whose range is {true, false}. For example Raining(x, s) corresponds to the statement that in situation s it is raining at location x, which is either true or false (the fluent would actually be “Raining-at-x(s)” → {true, false}, but these ideas are combined into one function with two arguments for simplicity).

A situational fluent also operates on a situation s, but it returns another situation instead of a boolean value. The situational fluent next(p, s) (where p is a propositional fluent) returns the next situation, starting from s, in which p is true. For example if s is the current situation, next(Raining, s) returns the next state of the world in which it is raining. Then, Wet(Fred, next(Raining, s)) would be true if Fred were caught outside without an umbrella the next time it rained. Fluents can also be used to represent causality. The fluent F(p, s) is defined to mean that the situation s will be followed (after some period of time) by a situation in which p is true. So Being-rained-on(Fred, s) → F(Wet(Fred), s).

Another use of fluents is in the description of actions. Specifically, the situational fluent Result(a, c, s) returns the situation that comes about when agent a chooses the action c in situation s. Consequences of actions can then be defined directly – Hand-hurt(Jeff, Result(Jeff, Punch-wall, s)) = true – or they can be used to describe conditional capabilities: Strong(Jeff, s) → Hole-in-wall(Result(Jeff, Punch-wall, s)). Actions can then by combined to form strategies in various ways, and the strategies can be evaluated to see if they would produce a desired result.

All of these formalizations are useful, and they allow a form of computable thought to take place. However, there are some problems. McCarthy and Hayes discuss a number of issues and present ideas for resolving them, and the remainder of this report will be devoted to briefly outlining the most interesting.

First, the concept of a situation brings with it several issues. It is impossible to actually describe in anything close to adequate detail the complete state of the world, or universe, at any given time. Therefore, in reasoning about these states, there are necessarily things that are going to be left out. For example, Strong(Jeff, s) → Hole-in-wall(Result(Jeff, Punch-wall, s)) might be true most of the time, but not if the wall is made of concrete. That objection could be met with a slightly modified statement - Strong(Jeff, s) ^ Not-concrete(wall) → Hole-in-wall(Result(Jeff, Punch-wall, s)) – but we could easily come up with any number of additional cases in which this statement would not be true, and it does not seem feasible to account for all of the variables that could affect a given action if we want to maintain any ability to generalize.

One solution to this issue is to come up with an idea of likelihood. One approach to this would be to attach probabilities to beliefs and outcomes, but McCarthy and Hayes dismiss this as infeasible. Instead, they want to formalize the notions of probably and normally. They do so in essence by saying

that if a fluent f is normally true in a given situation, and if the agent knows nothing to contradict f, then the agent can say f is probably true. It can then reason about different things that are probably true using formal logic, and arrive at reasonable conclusions without, hopefully, developing the characteristic AI brittleness that comes with a worldview that is too inflexible.

McCarthy and Hayes also discuss several other logics that could be relevant, including modal logic, tense logic (connecting events temporally), and action logic. However, the only only one they discuss in detail is modal logic, and since we have gone over that in class I will not do more than mention it here.

The views presented in this paper are an early attempt to formalize in a symbolic logic the way in which intelligent beings (i.e. people) think. There has been considerable debate over whether this is possible, and the argument is by no means concluded, although certainly the formalisms described in by McCarthy and Hayes are insufficient by themselves. Due to space constraints, this summary will not attempt to delve into that argument, but please see the bibliography for further reading on the subject.

References:

1) McCarthy, J., and Hayes., P. (1969), *Some Philosophical Problems From the Standpoint of Artifical Intelligence*, http://www-formal.stanford.edu/jmc/mcchay69.html

2) Bringsjord., S (1991) *Is the Connectionist-Logicist Clash one of AI's Wonderful Red Herrings?*, Journal of Experimental and Theoretical Artifical Intelligence, V.3 N.4, 319-349

3) Bringsjord., S and Zenzen., M. (1997) *Cognition is not Computation: the Argument from Irreversibility*, Synthese V.113 N.2, 285-320.

4) Kirsh, D. (1991) *Foundations of AI: The Big Issues*, Artificial Intelligence, V. 47 N.1-3, 3-30.

5) Sun, R (2009) *Motivational Representations Withing a Computational Cognitive Architecture*, Cognitive Computation, V.1 N.1, 91-103.