A modal characterisation of observable machine-behaviour.

R. Milner
Edinburgh University

1. Introduction

This paper is concerned with the interaction of an experimenter with a machine presented as a black box. We are concerned largely with properties of such machines (here called agents) which may be ascertained by a finite amount of experiment. By considering a sufficiently rich class of properties, we seek an equivalence relation over agents (that of having exactly the same properties) which is adequate for practical purposes, so that the meaning of an agent may be taken to be its equivalence class.

We have previously studied a relation, observation equivalence, of agents [HM, Mil]; it was defined intrinsically, in terms of the actions of an agent under experiment, rather than via extrinsic properties. In [HM] this was shown to agree with an extrinsic equivalence relation, but an important feature of agents was ignored—namely, the possibility of divergence. This feature has a strong bearing on what properties of agents may be ascertained effectively, i.e. by finite experiment.

We shall first define equivalence of agents in two ways—first intrinsically, and then extrinsically via properties expressed in a simple modal language. In Section 5 we show that these agree; we then exhibit conditions under which a given property A of an agent p may be effectively ascertained, and (briefly) conditions under which this is not the case.

2. Agents and experiments

Assume given:

(1) A set P, the agents,

(2) A set E, the experiments,

with the following structure:

(3) A family \{ e \rightarrow p | e \in E \} of binary relations over P. p \rightarrow p' may be read "p can undergo experiment e to become p'". If there is such a relation instance for p, we say that p can accept e.

(4) A unary predicate \uparrow over P. \uparrow(p) is written p\uparrow and may be read "p can diverge" or "p can proceed infinitely without accepting experiment". We write p\downarrow for the negation of p\uparrow.

For the present we assume nothing further about these relations. They may be infinite; even the image of p under some e \rightarrow may be infinite.
An agent may be thought of as a black box, equipped with a button for each experiment. It also has a green light, which is lit iff the agent is proceeding without responding to experiment. To attempt an experiment $e$ on agent $p$ we apply continuous pressure to the $e$-button; if the button goes down (after some time) then $p$ has accepted the experiment, and if the green light goes off without the button moving then $p$ has rejected the experiment. While neither occurs (and if $p\uparrow$ then it is possible that neither will occur) we can conclude nothing.

3. Experimental equivalence of agents

We aim to characterise the behaviour of agents by experiment only. In one method, we define a pre-order, $\preceq$, over $P$. Intuitively, $p \preceq q$ means that $p$ and $q$ have the same experimental behaviour except that $p$ may diverge where $q$ accepts an experiment. We define $\preceq$ in terms of a sequence $\preceq_k$ ($k=0,1,\ldots$) of preorders, each concerned with depth $k$ of experiment. For example, to experiment on $p$ to depth 2, we may attempt $e$ on $p$; if it succeeds - i.e. $p \xrightarrow{e} p'$ for some $p'$ - we may attempt $e'$ on $p'$.

**Definition**

$p \preceq_0 q$ always holds.

$p \preceq_{k+1} q$ iff, for all $e \in E$,

(i) $p \xrightarrow{e} p'$ implies $(q \xrightarrow{e} q'$ and $p' \preceq_k q'$, for some $q'$).

(ii) if $p\uparrow$ then

(a) $q\uparrow$

(b) $q \xrightarrow{e} q'$ implies $(p \xrightarrow{e} p'$ and $p' \preceq_k q'$, for some $p'$).

$p \preceq q$ iff $p \preceq_k q$ for all $k \geq 0$.

**Proposition 1.** $\preceq_0, \preceq_1, \ldots$ is a decreasing sequence of pre-orders, and $\preceq$ is a pre-order.

**Remark** In the absence of the predicate $\uparrow$, $\preceq$ reduces to the equivalence relation $\simeq$ studied in [HM] and [Mil], for a certain set $E$. In [HP] a similar, but not identical, pre-order is defined using divergence.

In general we cannot expect to determine $p \preceq q$, or its negation, effectively - i.e. by a finite amount of experiment. However, one purpose of this paper is to show that this relation is, under certain assumptions, at least as effective as the inclusion of recursively enumerable sets. In fact, we show that for each agent we can, by experiment, enumerate all its properties, and that $p \preceq q$ iff $q$ has all the properties of $p$ (and possibly more).

For equivalence of agents, we merely define

**Definition**

$p \preceq_k q$ iff $p \preceq_k q \preceq_k p$

$p \preceq q$ iff $p \preceq q \preceq p$