NO TURING MACHINE CAN REPLICATE AN ARBITRARILY CHOSEN TURING MACHINE BY OBSERVING ITS RESPONSES

Kerry M. Soileau

November 9, 2005

ABSTRACT

Is a brain no more than a complex Turing machine? If so, artificial intelligence research amounts to Turing machines attempting to deduce their own internal structure or that of other Turing machines. We show that no Turing machine can "analyze" (i.e., deduce the internal structure or its equivalent of) an arbitrarily chosen Turing machine if only interviewing is used. We define equivalence of Turing machines and irreducibility of sets of Turing machines. We make preliminary observations about sets of Turing machines that are analyzable.

Let us denote by $\boldsymbol{T}$ the set of all Turing machines whose halting sets are nonempty. Before proceeding, we give two definitions.

Definition: Two Turing machines T_1 and T_2 in $\boldsymbol{T}$ are equivalent if

(a) T_1 and T_2 have the same halting set, and

(b) $T_1(s)$ and $T_2(s)$ for every string s for which they halt.

Definition: A Turing machine A is said to analyze a subset $\boldsymbol{S}$ of $\boldsymbol{T}$ if by "interviewing" any Turing machine B in $\boldsymbol{S}$, A will in finite time produce a description of a Turing machine C equivalent to B. A is then called an analyst for $\boldsymbol{S}$. In an interview, the analyst machine presents an input string to the subject machine and observes its response. This cycle is repeated until the analyst makes a decision, and gives as its output a string uniquely describing the internal structure of the subject machine or an equivalent machine.

Does there exist a Turing machine that can analyze any given Turing machine? The answer is no. For, suppose such a Turing machine exists. Let us refer to it as a **universal analyst**, and denote it A. By assumption, we may choose any Turing machine B, and A will, after finite time, produce the description of an equivalent Turing machine C. For the purposes of this proof, we choose B such that the set of strings upon which B halts is infinite. Since A produced the description for C in finite time, it follows that A based its computations on a finite number of tests of the Turing machine B. In other words, A computed C based on a knowledge of B's output for only a finite number of (say n) input test strings. Let $S = \{s_1, s_2, s_3, \cdots, s_{n-1}, s_n\}$ be these strings. By our choice of B, there exists some string x, not in this set, for which B (and therefore C) halts. Since B and C are equivalent, we have $B(x) = C(x)$. Define a new Turing machine D such that D halts on x, B and D agree over the strings in S, and $B(x)$ is not equal to $D(x)$. If D is now presented to A, A will again produce C as a result, since the interview will proceed exactly as it did before, and thus must come to the same conclusion. Hence D is equivalent to C, and since D halts on x, we have $D(x) = C(x)$. Since $B(x) = C(x)$, it follows that $D(x) = B(x)$. But we earlier had that $D(x) \neq B(x)$. This contradiction proves that no such universal analyst can exist.

The proof of the previous Proposition, while somewhat dismaying, suggests that suitably restricting the subset to be analyzed might render that subset analyzable. To pursue this idea, we need another

Definition: A nonempty subset $\boldsymbol{S}$ of $\boldsymbol{T}$ is reducible if there exist distinct equivalent Turing machines in $\boldsymbol{S}$. A nonempty subset $\boldsymbol{S}$ is irreducible if no two members of $\boldsymbol{S}$ are equivalent. In particular, any singleton subset of $\boldsymbol{T}$ is irreducible.

Note: Any nonempty subset of an irreducible subset is irreducible.

Note: Only an irreducible set may have an analyst.

Proposition: If $\boldsymbol{S}$ is a finite subset of $\boldsymbol{T}$ and $\boldsymbol{S}$ is irreducible, then there exists a Turing machine A such that the union of $\boldsymbol{S}$ and $\{A\}$ is irreducible.

Proof: Label the members of $\boldsymbol{S}$: $\{S_1, S_2, S_3, \cdots, S_n\}$. Choose strings $\{x_1, x_2, x_3, \cdots, x_n\}$ such that S_i halts on x_i for $i = 1, 2, 3, \cdots, n$. Create a Turing machine A such that A halts on each x_i and $A(x_i) \neq S_i(x_i)$ for any i. Now suppose the union of $\boldsymbol{S}$ and $\{A\}$ is reducible. Then there exist distinct Turing machines B and C in this union such that B is equivalent to C. Since $\boldsymbol{S}$ is irreducible, and B and C are distinct, they cannot both be members of $\boldsymbol{S}$, and cannot both be equal to A. Without loss of generality we may assume that $B = A$ and C is a member of $\boldsymbol{S}$. Thus A is equivalent to C. Since C is a member of $\boldsymbol{S}$, C is equal to some S_j. Thus A is equivalent to S_j, so in particular we must have $A(x_j) = S_j(x_j)$. But by the construction of A we assured that $A(x_j) \neq S_j(x_j)$, providing the desired contradiction.

It is not true that all finite irreducible subsets are analyzable. To demonstrate this, we need only consider a finite set of Turing machines, no two of which halt on the same string. This set is irreducible, yet not analyzable, because the attempt of any analyst to identify one of its members would frequently be frustrated by the member's failure to halt. We define a property stronger than irreducibility in the following

Definition: A subset $\boldsymbol{S}$ of $\boldsymbol{T}$ is called separable if the halting sets of its members has a nonempty intersection, and for any distinct members T_1 and T_2, there exists some string x in this intersection for which they return different strings. Note that separability implies irreducibility, but the converse is not true. For reasons of convenience we declare singleton subsets to be separable.

Proposition: If $\boldsymbol{S}$ is finite and separable, then $\boldsymbol{S}$ has an analyst.

Proof: If $\boldsymbol{S}$ is a singleton set, we have already seen that it has an analyst. In the following suppose $\boldsymbol{S}$ has more than one member. Let n be the number of members of $\boldsymbol{S}$, and suppose $\boldsymbol{S}$ is separable. Label the members of $\boldsymbol{S}$: $\{S_1, S_2, S_3, \cdots, S_n\}$. Let H denote the intersection of the halting sets of the members of $\boldsymbol{S}$. For each $i = 1, 2, 3, \cdots, n-1$ and $j = i+1, i+2, i+3, \cdots, n$, there exists a string $x_{ij} \in H$ such that $S_i(x_{ij}) \neq S_j(x_{ij})$. Any member of $\boldsymbol{S}$ may now uniquely be identified by means of its responses to the strings $\{x_{ij}\}$, and these responses can be collected in finite time. A Turing machine implementing this test is an analyst for $\boldsymbol{S}$.

Definition: An irreducible subset is called maximal irreducible if none of its supersets is irreducible. Note that $\boldsymbol{T}$ is reducible, hence any maximal irreducible subset would be a proper subset of $\boldsymbol{T}$.

Proposition: There exist maximal irreducible subsets.

Proof: Note that equivalence of Turing machines, as defined above, is an equivalence relation. Hence equivalence induces a partition of the set $\boldsymbol{T}$ into a countable collection of disjoint subsets $\{\boldsymbol{S_1}, \boldsymbol{S_2}, \boldsymbol{S_3}, \cdots\}$. It is then clear that if we create a subset by choosing exactly one member of each subset $\boldsymbol{S_i}$, the resulting subset is maximal

irreducible. Conversely, any maximal irreducible subset contains exactly one member of each subset $\boldsymbol{S}_i$. We have thus characterized the general form of maximal irreducible subsets of the set of all Turing machines whose halting sets are nonempty. We note that every irreducible subset is a subset of some maximal irreducible subset. We also note that if one maximal irreducible subset has an analyst, that analyst analyzes all maximal irreducible subsets.

So far we have shown that every finite separable subset has an analyst, and that the maximal irreducible subsets are the largest that could have an analyst, although whether they do is yet to be seen. In addition, we have so far exhibited the existence of analysts only for finite separable subsets. Does some infinite subset have an analyst? The answer is yes, as we show in the following

Proposition: There exist analyzable subsets of infinite cardinality.

Proof: Choose a Turing machine B such that B halts on all strings over some finite alphabet A. Let & represent a symbol which is not a member of A. Define a sequence $\boldsymbol{S} = \{S_1, S_2, S_3, \cdots\}$ of Turing machines over the extended alphabet $A \cup \{\&\}$, according to

$$S_i(x_j) = \begin{cases} B(x_j) & \text{if } i \neq j, \\ '\&\&\&\cdots\&' & \text{where } \& \text{ is repeated } i \text{ times, if } i = j, \end{cases}$$

where $\{x_1, x_2, x_3, \cdots\}$ is an ordered enumeration of all possible input strings over the alphabet A. It is clear that $\boldsymbol{S}$ is both infinite and separable. Now we need only observe the existence of the following analyst D: Let D conduct an interview of a subject machine C belonging to $\boldsymbol{S}$ by presenting the strings x_i one after the other, beginning

with x_1. Since $C = S_j$ for some unique j, in finite time x_j will be presented to C and C will return a sequence of the form $'\&\&\&\cdots\&'$. The length of this string uniquely identifies C as identical to S_j. Using the known specification of B, the specifications of each of the Turing machines S_i are computable. Hence D is an analyst for the infinite subset $\boldsymbol{S}$.

For Further Investigation: Are all infinite separable subsets analyzable? Do there exist analyzable subsets which are not separable?