Working Paper Series
ISSN 1170-487X

Survival of the Species vs Survival of the Individual

by R H Barbour & K Hopper

Working Paper 94/14
August, 1994

© 1994 by R H Barbour & K Hopper
Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Survival of the Species vs Survival of the Individual

by

RH Barbour & K Hopper

Dept of Computer Science, University of Waikato, Private Bag, Hamilton, New Zealand

Abstract

This paper examines the relationships between human and computing entities. It develops the biological ethical imperative towards survival into a study of the forms inherent in human beings and implied in computer systems. The theory of paradoxes is used to show that a computer system cannot in general make a self-referential decision. Based upon this philosophical analysis it is argued that human and machine forms of survival are fundamentally different. Further research into the consequences of this fundamental difference is needed to ensure the diversity necessary for human survival.

Key-words : Survival, will, understanding, machine, culture, natural selection, halting problem, computability, paradox.

Introduction

While it is frequently stated in developed societies that computers are an invaluable influence on the society which uses them, there are other societies in which the computer has had no or very little influence. Similar sentiments have been expressed over millenia about various technological developments from writing to the latest advances in medicine.

Computers are undoubtedly an important aspect of modern technology. They are often attributed with similar characteristics to the humans who create, use and interact with them. They are even treated as sentient artefacts. This ascription of a potentially sentient nature suggests that, like other sentient beings, they could well be recognised as developing a "sense of survival" when the technology to do this has been developed.

These potential developments could have an extraordinary influence on the survival of humanity. Using arguments founded on well-known philosophical principles, it is shown that the nature of computer system behaviour is fundamentally different from that of humans. This difference leads to a different form of inherent survival mechnism which could lead to both major problems for people as well as to unforeseen benefits.

IT and the Moral Imperative

> *The arithmetical machine produces effects which approach nearer to thought than all the actions of animals. But it does nothing which would enable us to attribute will to it, as to the animals.*
>
> **B Pascal**, Pensées No 340

This thought of Pascal provides the motive for much of the technical and logical developments of modern Information Technology. It introduces, however, the notion of human will; it is this aspect of the thought and its implications in the light of modern technology which motivate the theoretical investigations described here. The natural will attributed to all animate species is considered to be the moral imperative to survival. This will to survive has been recognised biologically in Darwin's principle of natural selection as requiring diversity to produce 'better' survivors. This paper extends this principle to show its application in humans who produce cultural diversity as a survival strategy.

In searching for new markets, the software industry seeks to increase international use of applications; the need to understand the implications of the limitations and differences can no longer be ignored. It has become an increasingly complex and multi-faceted activity with strong tendencies towards monopoly.

These tendencies are inconsistent with our commitment to promote the diversity of human cultures. The need to understand the implications of diversity arises because software solutions have become available for and have been applied to an ever wider range of human problems experienced in so-called western society. Much human interaction is now supported by software; the software industry has as a result influenced and altered human relations. Much of this influence has arisen

from the research and development activities of western societies whose common cultural perspectives tend to mask important human differences in how, and over what, human interactions take place.

Philosophical Foundations

The intersection of these concerns highlights the pressing problems within computer science considered in this paper. Is there a sound theoretical foundation on which to build a framework for considering the relationships between people, irrespective of culture, and computing entities?

Two possible frameworks suggest themselves – computing theory and post-modern philosophy per se. Since computing theory may not be applicable to human behaviour and can be accommodated within post-modern philosophy, the latter is the chosen foundation for this discussion.

A well-established tactic would be to identify a basis for determining a priori the truth of statements which could be applied about computer systems and people. Such a priori truths are held to be truth by virtue of their meaning and thus (analytically) true in all possible worlds by virtue of their meaning. Zalta[12] shows that there are logical and analytical truths that are not necessary and that there are falsehoods which are possible. The combination of logic and a priori truth of Zalta's findings does not therefore appear to support 'truth' in this form as a suitable candidate framework.

Brouwer, however, argues[4] that mathematical understanding can be generated from the concepts of the perception of difference and any choice of representation of differences by sequences of sounds or marks. The particular set of sounds and marks that are used to represent mathematical understandings will vary from one cultural group to another. He also showed that there is no one special set of sounds and marks and that any agreed set will suffice.

Subsequently Brouwer showed that while both syllogism and contradiction are acceptable as a basis for a mathematical system, the Law of the Excluded Middle is not tenable. He argues that this principle reduces to the question of whether unsolvable mathematical problems exist – stating also that "there is not a shred of proof that there exist no unsolvable mathematical problems". Perhaps the most interesting notion in Brouwer's work is his contention that "In mathematics it is uncertain whether the whole of logic is admissible and it is uncertain whether the problem of its admissibility is decidable"[4].

Brouwer's notion is prescient of Turing's later statement[9] of the Halting Problem – given a Turing machine in an arbitrary configuration with arbitrary input, will the machine eventually halt or will it go into an infinite loop. While there are solutions to particular cases of this problem, there is no general solution – it is one example of a classically unsolvable mathematical problem as defined by Brouwer.

The relation between these two kinds of problems is that Brouwer was concerned with decidability, while Turing was concerned with failure to terminate. A necessary preliminary argument to the thesis of this paper is a proof that the ability to decide and the ability to terminate are equivalent in nature. The nature of Paradoxes of Self-Reference (as, for example, discussed by Priest[7]), which are intrinsically undecidable, is the starting point for this demonstration.

Russell's well-known Paradox may be defined axiomatically as – "Find the set consisting of all elements not contained in any set". This may be expressed operationally as a search for a set according to given rules – but without a specific goal (since the searcher does not know what is the set to be found).

Consider the Halting Problem (as stated above) as a potential paradox of self-reference. If the Turing machine may be in an arbitrary configuration and given an arbitrary input then the problem could be expressed as "Given a copy of itself as input can a Turing machine always decide whether it will halt?" In practice this particular form of the problem is the one used to show that in general the machine will fail to halt[5]. This statement for the purposes of proof (together with the proof) indicates that it is indeed an operational definition of a paradox of self-reference.

The Halting Problem, however, is of the nature of a dual of Russell's Paradox. The axiomatic form of paradox has a set of rules but no goal, whereas this operational dual has a goal but no adequate rules which would allow the goal to be reached. This may be expressed in the form of the following table

Paradox Form	Goal	Rules
Operational	*Yes*	*Inconsistent*
Axiomatic	*No*	*Yes*

in which the inconsistency of the operational rules may result from ambiguity or over-definition. In both cases sufficiency is necessary. We believe that this insight may be an important contribution to the understanding of paradox because the table identifies ways in which paradoxes may be created.

If both forms of expressing a paradox of self-reference are considered in operational terms, then both express the idea of a search for something which fails to terminate. The ability to decide and the ability to terminate are therefore, for paradoxes of self-reference, equivalent in nature. This is the first cornerstone of this paper – that decidability and ability to terminate are equivalent in respect of interaction between two entities.

The second cornerstone of the argument is concerned with the nature of interaction between active entities. Wittgenstein[11] argues that languages are tools for creating and relating to experience. This view contrasts with earlier views in which language 'represented' or 'fitted the world'. On the Wittgensteinian view Davidson[1] argues that words get their meaning from association with other words in particular contexts. Even within one culture people have diverse histories and diverse responses to those histories so that their appreciation of the meanings which they individually associate with words is likely to differ.

The existence of diversity between humans must not preclude the existence of human interaction. In order for such interaction to take place it is necessary that the entities (people) interacting have a means of communicating with each other. While the exact mechanism of such communication is not important, it can only proceed by transmitting sounds/marks/gestures, etc (denoted generically in the following as an utterance) from one to the other in a way which can be attended to by the recipient of such an utterance.

The essential pre-conditions for useful interaction are :–

a. The utterance must be perceivable by the recipient. This means that there must be a common medium for communication to take place.

b. Any structure imposed on the sequence forming the utterance may be described by a set of rules for such a structure, an instance of which would be recognisable by utterer and recipient.

c. The meaning ascribed by the utterer to the utterance must relate to a concept which is meaningful both to the utterer and the recipient and hence have a 'shared' meaning for both participants.

Unfortunately, in order for these three requirements to be met there is need for prior agreement in each case. Such prior agreement can, of course, only be achieved by prior communication on the basis of shared experience between the participants involved. While this may seem to presage an infinite recursion in definition terms, the notion that there has to be some shared experience, however little, for communication to take place at all is the key to the solution. Note that this infinite recursion is another application of the operational view of paradoxes of self-reference — that they fail to terminate!

Interaction between Entities

The development of higher human concepts, other than basic physical shared experience, rests upon the ability to develop more complicated ways of communicating based upon those primitive shared experiences. Behaviour which can be viewed as a mechanism which permits the definition of concepts so that such a definition can be passed from an utterer to recipients is postulated here.

The availability of computer systems as active agents, which appear to be able to communicate with each other in a way similar to humans, has led to the hoped for possibility that at some future time they could be used as intelligent participants in human thinking activity[10], interacting with their human counterparts just as people interact.

In order to explore the possibilities of interactions between humans, between computers and between a human and a computer, it will be assumed initially that a computer system is an entity which can be given such human-like behaviour.

To be able to interact in these kinds of way, both humans and computer systems must behave in a manner which reflects the general notion of a shared concept. In order to share such concepts, the entities must generate a means of describing an arbitrary concept. Since the concept to be described is arbitrary it is useful to employ the concept of such description – a meta-definition in fact – as the subject of an interaction. Using such a meta-definition in this discussion ensures that further communication required for learning will always be possible.

In discussing such an interaction it is essential to ignore the particular (human or computer) nature of the communicating entities and, equally, to ignore the nature of any medium which may be needed to effect practical communication between entities. The argument which follows focusses, therefore, on behaviour which has observable outcomes.

Recalling the initial premise that human and computer interactions may be considered equivalent, it is possible to write

$$H \leftarrow x \rightarrow H' = C \leftarrow x \rightarrow C'$$

where H represents a communicating human,
C represents a communicating computer system
and $\leftarrow x \rightarrow$ represents the communication of x by the entity on the left to the entity on the right, although interaction may be in both directions.

If computer systems are indeed to be considered equivalent to humans in this form of interaction, then the possible behaviour of the computer system in relation to x during the interaction must be equivalent to the behaviour of the two humans during the interaction. In other words

$$B_C(x) \equiv B_H(x)$$

where B denotes the behaviour of the subscripted entity (eg C) in relation to the parenthesised concept (eg x).

Since the interaction on the right of the communication equivalence given is solely between computer systems, it is necessary to conduct an experiment designed to show if the interaction is indeed equivalent and, if not, gain some insight into the nature of any difference revealed.

Given that the experimenter needs to make decisions about the behaviour of a computer system, it is reasonable to test the hypothesis in an equivalent modified form, as

$$H \leftarrow x \rightarrow C = H \leftarrow x \rightarrow H'$$

Since the equivalence between H and C is hypothesised, it does not matter in this context whether a human or computer is involved in any particular interaction and the above is equivalent to the original statement.

Thesis Given

(1) Two humans (H and H') and a computer system (C),

(2) The concept of a meta-definition unknown to the computer system (C) and one (H') of the two humans,

(3) The concept of a meta-definition known to the other human (H),

(4) The concept to be communicated is the meta-definition concept.

then the behaviour of the computer system after having been given the meta-definition by interaction with the human who knows it is *not* equivalent to the behaviour of either human after an equivalent human-human interaction.

This may be formally expressed as

$$H_{pre} \rightarrow B_H(I_{md}) \rightarrow C_{pre} \not\equiv H_{pre} \rightarrow B_H(I_{md}) \rightarrow H'_{pre}$$

for which

$$B_H(I_{md}) \in B(H_{pre})$$

and

$$B_H(I_{md}) \notin B(C_{pre})$$

where $B(E)$ is the set of all possible behaviours of E,

and I_{md} is the meta-definition concept information.

there is an interaction

$$H_{pre} \leftarrow B_H(I_{md}) \rightarrow C_{pre}$$

which is terminated by H for which

$$B(H_{post}) = B(H_{pre})$$

and

$$B(C_{post}) = B(C_{pre}) \cup \{B_H(I_{md})\}$$

Justification

Since the recipient of the meta-definition in the experiment suggested by this theorem is a computer system, the only way in which a proof may be obtained is by conducting a subsequent interaction between the same computer system and the same human participant which will enable the human participant to conclude that the behaviour of the computer system in respect of the meta-definition passed to it is the same as would be the human's own behaviour.

This second interaction may be characterised as

$$C \leftarrow ? \rightarrow H$$

where the burden of proof is now to consider what the nature of the concept to be passed must be in order to be conclusive evidence to the human participant in the interaction. Note that no other human participant is possible in this second interaction as the question to be decided has now become, "Is the behaviour of the computer system now identical to my behaviour in respect of the concept of a meta-definition?" which is only decidable by the original human participant since it necessarily refers to behaviour which is generated internally.

Human behaviour in interacting with others in respect of some behaviour learnt may take one of three forms :-

a. Reporting the behaviour learnt.

b. Applying the behaviour learnt.

c. Generating a description of (ie teaching) how the behaviour was learnt.

Each of these three is a potential candidate interaction for the purposes of justifying the thesis. Consider them individually :

a. $C \leftarrow B_H(I_{md}) \rightarrow H$ is not an acceptable behaviour by the computer system as justification, because it merely requires the ability to copy the original interaction, acting as a sort of mirror. The termination point of this interaction is pre-defined by the original message.

b. $C \leftarrow X \rightarrow H$ where $X = Apply(B_H(I_{md})).Y$ for some Y is not acceptable as justification either, since it is a purely mechanistic process implied in the original description received. The termination point of this interaction is also implied by the termination of the application of the defined behaviour.

c. $C \leftarrow Z \rightarrow H$ where $Z = B_C(Generate).B_H(I_{md})$ which produces a description of how to generate the description $B_H(I_{md})$ is, however, an acceptable interaction for the purposes of justification since it requires the ability in the computer system to constructively use the acquired meta-definition *and* also the ability itself to decide when to terminate the interaction Z.

The key component of interaction in c above is the need for the computer system to decide when to terminate the interaction. Such a decision requires the ability to refer to its own behaviour – which is a paradox of self-reference! This form of paradox has been shown to be equivalent in nature to the Halting Problem. The inability of the computer to terminate this interaction therefore means that its behaviour cannot be identical to that of its human counterpart.

The original premise that a computer system could be made to behave like a human being with respect to this problem is therefore shewn to be false. This finding, incidentally, also raises serious doubts about the veracity of the Turing Test.

The Nature of the Difference

Russell's Paradox manifests itself in a conversation context in the guise of the inability for two participants to know that the concept which has been the topic of some conversation interaction is known *identically* by both participants. While the participants may have indicated that they are prepared to state concurrence, this may be more of the nature of a stipulation in respect of future conversation which may need to use such a stipulation as the basis for further discussion.

This stipulation is a mechanism by which human participants in a conversation can unilaterally indicate that they comprehend something by moving outside the context given. Hence, so far as they are concerned agreement has been reached. The reasons for such stipulation are of no concern at present, since the other participant has no way of confirming that the concept which is internalised in the other person's head is identical to that which (s)he has internalised. It is exactly this uncertainty which perturbs the exactness of concepts passed from one person to another and provides random variation for the generation of new ideas, assisting in providing the diversity between individuals necessary for the survival of the human species.

Since computers are incapable of determining the conditions for their own termination, termination of any interaction must be imposed from outside (ie by some human programmer). The necessary imposition of agreement upon a computer system contrasts markedly with the view of human agreement. Imposition in this way implies that a computer system is necessarily rooted in analytical world 'data'; the computer necessarily embodies an analytical model of interaction. The need for this model of reality for a computer system suggests that a computer system is 'designed' not for diversity and self-survival in the sense attributable to humans, but to design diversity out and impose on those with which it interacts, actions/restrictions to ensure a single perception of survival, irrespective of the survival of others (whether other computers or other humans). Since a machine encapsulates a single analytical truth, it is designed for the survival of the individual entity only.

By its very nature, therefore, a computer system will impose a statement rather than come to agreement with its human counterpart. The inability to negotiate agreement implies either human acceptance of the statement or ignoring the computer system entirely. Accepting the statement is accepting an imposed idea. This will tend to reduce the human to conformance and weaken the diversity essential to the survival of the species.

Conclusion

There is a fundamental limitation on the interactive behaviour of a computer system in relation to a human being. This limitation arises because of the inability of a computer system (a Turing machine) to decide when to terminate an arbitrary interaction. The insight gained by examining this limitation of computer behaviour suggests that the nature and directions of research into application portability, human-machine interaction and artificial intelligence subject to this limit need to be rethought.

References

1 Davidson D, *A Nice Derangement of Epithets,* in le Poré E (ed) *Truth and Interpretation: Perspectives on the philosophy of Donald Davidson*, Blackwell, (Oxford, 1986).

2 Dennett Daniel C, *Consciousness Explained*, Penguin, (London, 1991).

3 Gerard RW, *Units and Concepts in Biology,* in Buckley W (ed), *Modern Systems Research for the Behavioural Scientist*, Aldine, (Chicago 1968).

4 Heyting A (ed), *Brouwer LEJ Collected Works: 1 Philosophy and the Foundations of Mathematics*, North-Holland/American Elsevier, (New York 1975).

5 Hopcroft JE & Ullman JD, *Formal Languages and their Relation to Automata*, Addison-Wesley, (Reading, Mass 1969).

6 von Neumann J, *Theory of Self-reproducing Automata*, University of Illinois Press, (Urbana, 1966).

7 Priest G, *The Structure of the Paradoxes of Self-Reference*, Mind **103**(409) 1994, (pp25-34).

8 Rorty R, *Contingency, Irony and Solidarity*, CUP, (Cambridge, 1989).

9 Turing AM, *On computable numbers with an application to the Entscheidungsproblem*, Proc London Math Soc **42**(2) 1936, (pp230-265).

10 Turing AM, *Computing Machinery and intelligence*, Mind, **95**, 1950, (pp433-460).

11 Wittgenstein L, *Tractatus Logico-Pilosophicus*, Routledge & Kegan Paul, (London, 1961).

12 Zalta EN, *Logical and Analytical Truths that are not necessary*, J Phil **85**(2), Feb 1988, (pp57-74).