AI is Dead: Long Live AI

Paul Qualtrough
Department of Computer Science
University of Auckland
email: paulq@cs.auckland.ac.nz

March 1999

Abstract

The time seems ripe to reflect on the state of AI and re-evaluate its goals and techniques. It is argued that a fundamental oversight has been made in adopting a focus on knowledge and reasoning without simultaneously emphasising understanding and meaning. Despite not being able to define the latter term precisely, we can exploit our limited knowledge of meaning to argue that most AI efforts will simply never achieve artificial intelligence of the kind that many pioneers in the field envisaged, or that most members of the public imagine when they hear this term. The only route to this goal is argued to be via learning, and objections to this approach are argued to be weak or flawed. A number of implications of learning with the specific aim of acquiring meaning and understanding capabilities are discussed.

1 Introduction

A spate of recent anniversaries has served to remind us that it is barely fifty years since the first computers were switched on and the computer revolution began. Today it is easy to be impressed by the impact of computer technology on diverse fields from cell biology to cosmology, or from power generation to publishing. But in one much-publicised field where computer technology seemed to promise so much, viz. *artificial intelligence* (AI), so little seems to have been delivered. Despite confident predictions that artificial intelligence was just around the corner (e.g. see [2]), when it comes to the archetypical pinnacle of AI research — a mobile robot — the present truth is that they have considerable difficulty in recognising a corner, let alone moving around it.

In fact, many AI systems, including mobile robots, suffer from problems of *brittleness* — they may function well under controlled conditions, but outside their unintentionally narrow and ill-defined domains, they can fail catastrophically. For example, a robot which operates well in a diffusely lit laboratory may be confused by its own shadow in direct sunlight. Such failures may be easy to understand when they occur, but are very difficult to predict in advance because of the ill-definition of the system's domain.

So where, if anywhere, has AI gone wrong? Is there a better approach than those we have adopted? Have we overlooked something? There seems to be steady progress, albeit slow, but is it in the right direction? Is there even a clear goal for AI? I shall consider only some of these questions, the clearest answer to any of them proving to be a "yes" to the third one just posed. This conclusion will have some clear implications to answering some of the other questions, while remaining ones are only posed to encourage readers to undertake their own assessment of the past, present and future of AI.

2 Intelligence

What *is* intelligence? Despite many attempts to answer this question in the history of AI, no satisfactory definition of intelligence has ever been produced. I shall not attempt to redress this situation but would suggest that one of the earliest landmark papers on artificial intelligence remains one of its most important.

Alan Turing's 1950 paper [23] questioned whether computers might be capable of demonstrating intelligence, how one might test for this intelligence, and how one might go about producing it. Having assumed more than justified a "yes" answer to the first of these questions, and gone into some detail on the second, Turing briefly outlined two proposals to address the last of them. The second of these was to produce artificial intelligence by programming a computer to learn from interactive experiences with our human world. The idea appeared to fall out of favour very quickly, in part due to the difficulties involved in implementing it. Recently it has resurfaced, been discussed in a positive light, and used to inspire new directions in intelligent systems research (e.g. [24, 4]).

In contrast, while it has had many critics, Turing's proposed test for intelligence has never lost its relevance. Due to our inability to define intelligence in precise terms, we remain highly reliant on observations to distinguish between behaviours we consider intelligent, and those which fall short of this mark. The way Turing proposes to test the behaviour of a potentially intelligent machine is interesting to note.

Turing pictures an anonymous interview between a human judge and two unknown parties, only one of which is a woman. The woman's task is to help the judge identify her as the woman. The other party is either a man or machine which seeks to be identified (incorrectly) as the woman. All communication between the judge and these parties occurs via a medium which removes any physical differences between the woman and the man or machine. The machine is deemed to be as intelligent as a man when it can cause the judge to incorrectly identify it as a woman at least as often as a man can.

The judge issues questions to probe the knowledge of the parties, to test their reasoning, their recall of events, and so on. He[1] is well aware that the man or machine

[1] This is definitely not intended as a sexist statement. Rather, if there is any possibility of there being

will be lying, misleading and/or exaggerating. But there is something deeper than these superficial features, which I believe to be an implicit test for *understanding*.

Without understanding, I believe Turing thought the judge would be able to trick one party into making some mistake, or producing an inconsistency from which it could not satisfactorily recover. This failure would reveal that party to be artificial, and by implication *less* intelligent than a human being. With understanding, the implication is that there may be fallibility, but there would also be mechanisms to cope with any such events — producing explanations for them, refusing to be drawn in to them, or doing something else which is otherwise characteristic of (human) intelligence.

Whether or not this interpretation of Turing's paper is what he intended, understanding seems to be an important component of our intelligence. So rather than focus on the indefinable conglomerate which is intelligence, let us turn our attention to the (hopefully) more tangible phenomenon of understanding.

3 Understanding

What *is* understanding? I have only recently become aware of the large amount of philosophical and psychological literature on this and related questions, and have not yet had time to survey it to the degree that it warrants. Nevertheless, the relatively small amount of material I have read, being an assortment of introductory chapters in AI texts, overviews such as [1] and [22], and more in-depth papers such as [25] and [20], seems consistent with the conclusions I will draw below from my own observations.

Before I begin discussing understanding, I should point out that because of my limited exposure to the philosophical and psychological literature proper, it may be that some of the words I use will not be employed in the sense which is normal in those fields. My apologies if this causes any confusion. To avoid one potential source of confusion, I shall henceforth restrict myself to describing and using understanding as a verb. As a noun, it may convey either "knowledge" or "meaning", things which are better described using those words.

It seems to me that understanding is primarily an act of making a meaningful mental connection between something previously unknown and something known already. More generally, it is attributing meaning to some event or observation, such as a gesture or statement; it is coming to know what something means.[2] Certainly, meaning is essential for understanding, and appears to play two roles in the main process we use to understand things. That process dominates to the extent that it alone is usually identified with the

things which only a woman can know, then the judge must be male. Since we cannot prove that there are no such things, it is safest to use a masculine judge.

[2]Regrettably, meaning is as difficult to define as intelligence. So if you feel the need for these statements to be grounded, you will have to find your own way to do this, and try to bear with me. The only definitions I could provide would be circular, such as "meaning is what is intended to be understood by some statement". This is similar to what my Webster dictionary does explicitly, and my Concise Oxford dictionary implies.

term "understanding". However, there is an important secondary mechanism which is so closely related to this kind of understanding that I will include it in my definition of understanding. I will discuss that mechanism a little later.

The first role that meaning plays in most of our acts of understanding is simply to exist in or otherwise be associated with an item of prior knowledge — the "something known". This knowledge has already been understood; it has previously been made meaningful in some way. The second role of meaning here is that it must exist in the way that such knowledge is connected to the entity which is unknown. This connecting process may be viewed as a generalised form of reasoning. It may be purely rational, but it can also simply be associative, or even quite illogical — one can indeed be "right for the wrong reasons". As long as the process is meaningful in some way to the agent making the connection to prior meaningful knowledge, an initially unknown, meaningless entity can itself become meaningful, and be understood. In being understood this item too becomes *knowledge.*

In my view, to say "meaningful knowledge" is redundant — knowledge must have meaning, or it is not knowledge at all. However, I will occasionally use both words together in this section to stress the relationship between them since there are some subtleties involved. As a contrasting example, making a meaningless connection, such as *automatically* equating an unfamiliar statement with truth, does not constitute understanding. In this case it is acceptance, and the process turns the statement into a *belief*, not knowledge per se. Since forcing computers to accept and act on statements blindly is the basis of much conventional AI research, I would suggest strongly that despite extensive use of the word "knowledge", this work is better characterised as having been concerned with "belief-engineering" and having produced "belief-bases".

Now while one can believe absolutely anything one likes — even in the face of evidence to the contrary — to know and understand something is a different matter. Accordingly, the problem of brittleness faced by many AI systems seems very likely to be case of *misunderstanding* — they have been forced to accept things and reproduce behaviour in a rote fashion, but they do not *know* anything per se. In the case of a robot confused by its shadow, a lay person would plausibly classify its behaviour as being due to misunderstanding. Can we argue convincingly that they would be wrong in this assessment? It seems far more plausible that they would in fact be right.

The problem of brittleness is widely argued (if not accepted) to be caused by a lack of "common sense". The conjecture is that systems which suffer from it generally don't have enough of the sort of "facts" we take for granted. As a result they are often unable to cope with unforeseen situations or even certain variations on known themes. As I have written it, I agree with this conjecture, but that agreement only comes by interpreting it in a somewhat unconventional way. The "facts" conventionally deemed necessary are factual ones; my opinion is that their validity in any absolute sense is a red herring. Rather, such information as an intelligent system has, simply must be *meaningful* to it to allow robust operation; it must be genuinely understood.

Put another way, the conventional assumption is that the quality of this common sense is fine but the *quantity* is insufficient. In contrast, my argument is that the *quality* is the primary cause of brittleness, while quantity issues are secondary. While Douglas Lenat has championed the quantitative cause and emphasised bulk common sense [12, 11], I would argue that qualitative aspects have been shortchanged and that we should emphasise *meaningful* knowledge. I have no particular objection to the term "common sense" save to warn that it is not as common as we might like to think it is. All things considered, I agree that calling this problem the *common sense knowledge problem* is a good characterisation of the root cause of brittleness.[3]

Now, returning to the distinction made between beliefs and knowledge, let me emphasise that the common sense knowledge required to avoid brittleness must include meaningful ways of reasoning. In some ways this is more important than meaningful facts, since today's knowledge tends to become tomorrow's beliefs. Nevertheless, we observe that understanding by our forebears, or during our own childhoods, was quite possible despite some of the things once considered factual or rational now being completely laughable to us. What we understand at any time is always a result of the meaning in them — *as far as we are aware*. When new discoveries expand our awareness, that meaning does not disappear or diminish, but it comes to account for less of our enlarged reality. The way we deal with such new phenomena, or shine new light on older observations, is at the centre of the other way we understand things.

This second process, which I also consider to be an act of understanding (but also accept that others may not, and don't see this as a problem), is at least conceptually straightforward: we *abstract* meaning from observations; we induce new concepts or new ways of connecting them. In doing so, we create and associate item(s) of knowledge with event(s) from our experience. For example, we see a number of blue objects, we hear people saying the word "blue", and we abstract — and understand — *blue* as a concept. Or perhaps we acquire and understand these concepts in advance of associating them with any word or linguistic construct. Either way, if we try to explain such a concept, we are usually left without any recourse except to provide examples of it — or to give circular definitions. So while we understand concepts like this, we do so in a different way to our more utilised form of understanding which involves making connections.

To draw this section to a close, let me summarise my perspective on understanding. Understanding is a either generalised reasoning process which connects observations to prior knowledge, or abstraction from direct experience to create completely new knowledge. The first process dominates our thinking about understanding, and its two components — reasoning and knowledge — have dominated AI research for most of its history. So a focus on understanding, even if one excludes abstraction from it, conceivably loses no generality as far as most AI researchers might be concerned. But abstraction is such an important supplement to this process that it belongs in all discussions about understanding, if not under the banner of understanding itself. Together, both processes are critically dependent on meaning. My preferred definition, adopted from this point forwards, is:

[3] One article I have come across [13] suggests that the common sense knowledge problem is an umbrella term for several related problems. Brittleness is not mentioned in this context at all, but the component problems are stated as perceived causes rather than as symptoms to be addressed. I infer these symptoms to be brittleness. If I have made an incorrect identification here between brittleness and what is referred to as the common sense knowledge problem, I apologise.

understanding *is the act of making something meaningful.*

I see no reason to limit what can be made meaningful. So long as something is made meaningful, then I shall believe that whatever it was that was involved has been understood. But if we are going to rely on meaningfulness as the foundation of understanding, then we would no doubt like to understand what meaning means.

4 Meaning

What *is* meaning? This question has apparently also occupied many philosophers and others for a very long time, and intensively so in the earlier part of this century [20]. Unfortunately, the answer is still not clear. While at times (including here on occasions) it is discussed as if it exists in a tangible sense, it is more often seen as an abstract property of "mental objects"; a kind of beauty in the eye (or mind's eye) of its beholder. Whatever meaning is, it is widely accepted that it could only exist in one's mind (whatever that may be) rather than in physical symbols or acts such as words or gestures.

For example, when one person talks to another, the meaning is *represented* rather than conveyed; it remains with the speaker, and some facsimile or functionally equivalent version of it is generated by the listener. Walter Freeman, who reinforces this view, goes further and argues that meaning has a concrete existence in the stable patterns of activity amongst brain cells [8]. Most other researchers fall short of suggesting specific physical relationships between meanings, minds and brains.

However, as with understanding, it is not my intention here to review the various opinions, favour any of them, or shed any further light on this question. I simply don't think I can. Rather, let us consider the implications for an intelligent system which is required to understand in the sense just defined. It would need to abstract and/or exploit meaning to do this. I will argue that in accepting four things that I have already touched on, namely:

1. meaning is a purely internal, "personal" thing, distinct from external representations exchanged by autonomous agents;
2. meaning can not exist independently of an intentional object — it is always expressible as "meaning of ... ";
3. the definition of understanding above is correct, and therefore (from 2) the existance of either understanding or meaning necessarily implies the simultaneous existance of the other;
4. there are no other ways an agent may understand other than by abstraction or by making meaningful connections to prior meaningful knowledge;

we have enough information about meaning to place some important limits on how it could be utilised, despite not knowing precisely what it is. Unfortunately, until we specify exactly what meaning is, we will not be able to definitively argue whether or not a particular agent has understood.

Nevertheless, let us consider how synthetic creatures might obtain the meaning they would need to exhibit understanding. Broadly speaking there are only three possibilities: we take responsibility for giving them meaning; they take responsibility for acquiring their own meaning; or we argue that meaning is somehow inherent in them and they already possess forms of understanding. I believe we are in a position to rule out the first of these possibilities, and dismiss the last of them as being of no particular help to us.

We may do this by considering two unlike agents, A and B, and some abstract statement of a "common sense" fact. Suppose this statement is "atomic" to A in that it is "just the way things are"; it may be directly experienced, and meaning may be abstracted from it directly. Suppose, however, that B is unable to make any corresponding direct observation of the phenomenon. For example, a statement such as "when one closes one's eyes, one can't see" may be an "atom" of common sense knowledge to a human but not to a computer.[4]

How could B understand such a statement? Since B has no basis for abstracting the required meaning directly, its only recourse is to relate the statement to its prior knowledge about the functions of eyes and eyelids. Any such knowledge must have meaning in it to be knowledge, and that meaning is B's alone. But then B's prior knowledge could only have become meaningful either by being related to other meaningful knowledge or by being abstracted from B's direct experiences. At some point, B is forced to abstract its own meaning independently, and make its own meaningful connections to this abstracted meaning in order to ground the whole process. While B may be able to understand such a statement in due course, it could not do so until it had abstracted an equivalent set of concepts to those held by A, and perhaps an equivalent set of reasoning or connecting techniques.

If meaning could be inherent in an agent then it would most likely be in the sense that the actions an agent takes were meaningful to it simply because that is what it did; meaning could be argued to arise through being. This particular line of thinking does not strike me as very plausible, chiefly since it may imply that things such as fire demonstrated understanding by my definition. In any case, this possibility will not be of much help to an agent unless it also has ways of acquiring other knowledge, and, even if it were lucky enough to have that inherently as well, ways of seeking out new meaning when faced with new situations. So, whichever way one looks at it, an agent which understands, and can also build on that understanding, must have some way of abstracting new meaning.

This important secondary mechanism for understanding has been frequently overlooked in the past — giving it a different name such as abstraction provides a basis for excluding it from discussions on understanding. By including it under the banner of understanding, we are forced to acknowledge the vital role it plays; we are encouraged to forgo our analytic mindset for a more synthetic one. With this we conclude that, in order

[4]This particular example is used to allow understanding here. However, I suspect that this sort of common sense fact is far in advance of what any agent, artificial or not, might first use as an atomic basis for its knowledge.

to understand, an agent must be able to *learn* from the outset. Here learning is used as a general, inclusive term which encompasses all the plausible mechanisms by which an agent might discover, develop or otherwise acquire and exploit meaning.

So, having initially sought to produce intelligence, our focus has shifted from there to understanding and meaning, and now to learning. This establishes a connection between Turing's intelligence test, and his suggestion that artificial intelligence might be produced in an agent capable of learning from sensory interactions with the world. While not stating that learning *will* lead to understanding, I do argue that if understanding is possible in an artificial agent, then it will only occur through learning. I would also argue that understanding is such a vital element in what we consider to be intelligence, that artificial intelligence itself can only be achieved through learning.

Befre we consider the implications of these conclusions, there is a final point to make about meaning. While directly transferring meaning between unlike agents has been ruled out, there is some possibility that meaning could be transferred through some copying or cloning process between identical or sufficiently similar types of agents. Such transfers would be desirable in mass-producing the likes of mobile robots intended for domestic use, where no two environments are alike, all environments change to some degree over time, and an ability to understand the environment and events within it would be highly advantageous to such robots.

However, we cannot rule in the possibility of directly transferring meaning either — if Walter Freeman has it right, then any disruptions to stable dynamic patterns of behaviour may well destroy the meaning in them. Clearly this is an issue to be researched. Perhaps it will motivate another flurry of activity on the question of "what is meaning?".

5 Learning

What is Learning? Little has changed since Minsky observed that there were "too many notions associated with learning" [14]. But is this necessarily a bad thing? If we accept that understanding is required for intelligence, and that this necessarily means an artificial agent must acquire its own meaning, then these agents are going to need all the help they can get — we would like them to understand every aspect of their operations, which means that they should learn everything at some point. Any mechanism which adds to or accelerates the acquisition of meaning and understanding capabilities should rightly be called learning.

It is beyond the scope of this paper to investigate specific forms of learning suited to the task (as has been done in [18]) but there are a number of observations we can make and constraints we can place on any learning system which is employed for this purpose.

Firstly, the notion of *symbol grounding* [9] becomes crucial to efforts along these lines. It is clear that a great deal of our mental activity is of a symbolic nature, but equally clear that these symbols do not float as castles in the air, detached from the reality of the non-symbolic individual nerve impulses which give rise to them. Symbol grounding is intimately connected to understanding: it asks how meaning can be made intrinsic in these symbols rather than parasitic on the meaning in our own minds which we associate with symbols.

Unfortunately the word "grounding" implies that this occurs in a top-down fashion. Like Harnad [9], I see this problem as able to be tackled only from the bottom up: we must ask how we can abstract meaning from experience and, in effect, *encapsulate* that in a symbol, not how we can take a symbol and relate it to experience. To obtain a valid symbol to relate to experience implies that we have already grounded it in part. Symbol grounding is the major open problem to address in attempting to emulate the sort of understanding which we use in our everyday lives.

A second important consideration in learning is the role of the learning agent's environment. Significantly, there should be one if we want learning to be an ongoing, "lifelong" process. Otherwise agents will simply run out of things to learn. The way the agent interacts with and experiences this environment will have implications to the meaning it can abstract from it. For example, autonomy in the choice of actions taken in the environment will provide more scope for learning than prescribed (automatic) actions. Prescribed actions may even preclude understanding since there is no meaning intrinsic in the particular choice of actions.

The wide variety of learning activity evident in the natural world has many lessons which can be transferred to our artificially intelligent systems. However, the wholesale application of observations and insights, as commonly occurs in much AI research, simply cannot be expected to work if understanding is our goal. So we must question what we can legitimately give to our learning agents, and how they might both take that material, and gather other material from their environments. A number of principles along these lines, suggested for use with autonomous mobile robots but having considerably more general application, have been developed in [18].

6 On the Other Hand...

The conclusion that autonomous agents must learn from the lowest level upwards, in order to understand and form the foundation for intelligent systems, is somewhat daunting. Although the need for learning in robots, for example, is widely acknowledged (e.g. [7]), in general, learning tends to be seen as best deferred until most other issues are dealt with [3, 10, 16]. It is difficult to find anywhere a comprehensive learning-based approach to either robotics or general AI aims gets more than a passing or indirect mention (e.g. in [26]). This seems to be due to arguments against it such as the following from [6]:

> *One idea that has fascinated the Western mind is that there is a general-purpose learning mechanism that accounts for almost all of the state of an adult human being. ... AI students often rediscover [this idea], and propose to dispense with the study of reasoning and problem solving, and instead build a baby and let it learn these things. We believe this idea is dead, killed off by research in AI (and linguistics and other branches of "cognitive science"). What this research has revealed is that for an organism to learn anything, it must already know a lot. Learning begins with organised knowledge, which grows and becomes better organised. Without strong clues to what is to be learned, nothing will get learned.*

If this argument is not simply based on the principle known as the *hermeneutic circle* — interpretation requires understanding; in turn, understanding requires interpretation [25] — then it is certainly very closely related to it. And in turn, both are closely related to what I have been saying. However, both deal only with our dominant way of understanding, viz. developing knowledge (understanding as a noun) from existing knowledge through generalised reasoning (cf. interpretation). While the last sentence quoted above raises an issue I shall have to return to later, neither that argument nor the hermeneutic circle appear to take any account of how knowledge or understanding might arise or occur in the first place.

For example, consider the fact that each of us is an adult human being teeming with knowledge, and yet each of us began life as a fusion of two simple cells. Neither of these cells has any knowledge as far as anyone can tell, at least not in the explicit declarative sense implied by those arguing against learning from this standpoint. There are two points to make from this. First, the development of humans from embryo to adult occurs on a continuum. This suggests that any observations consistent with the objection above would be better explained as a tightly coiled but ever-advancing "hermeneutic helix" than as a circle which suddenly comes into existence at a certain stage of development.[5]

Second, this argument can be seen to focus solely on "knowing that" (declarative knowledge) while ignoring "knowing how" (procedural knowledge). We might then ask how it could be that an embryo devoid of "knowing that" acquires it, and the meaning implicit in it, if it is not inherent in the embryo from the beginning?[6] I shall shorten this term to "know-that" and agree with Ryle's argument that the path to acquiring it lies in utilising the right kinds of the corresponding "know-how" [19]. For example, when a baby is born it does not *know that* crying will bring it attention, but it does *know how* to cry — or does it?

[5] I have recently wondered whether basket weaving may provide a better metaphor for the understanding processes involved (unfortunate associations with "basket cases" aside...). By the time human subjects are old enough to provide data in which clear relationships can be seen, the basket is taking shape, and it is easy to ignore the loose ends. But a basket's strands all start as individual threads, and only a few are added to the basket at a time in its early stages. Despite its scruffy, ad hoc nature, the way a bird constructs a nest may be a more accurate model still — new strands can be added at any time, can be of a wide variety of forms, but soon lose their individual importance as the nest takes shape. Importantly, the nest is made using imprecise skills out of what is available; in contrast, an apparently simple, repetitive basket weaving pattern can mask a good deal of complex preparation and precise skills.

[6] The radical suggestion that declarative knowledge is indeed present in an embryo has been made — and (thankfully) rebutted. For example, see [25].

I am more comfortable with the suggestion that the baby simply cries, and does not really *know* much (if anything) at all. To know implies meaning, which must be abstracted somehow or be inherent in the first place. My conjecture would be that a baby possesses completely meaningless actions which soon become know-how by being carried out in *contexts* which emerge as the baby learns. The meaning is self-abstracting if you like — it is context that "gives meaning" to actions. With no reference to a context (e.g. if executed randomly) or with no effective context (e.g. if always executed, such as in blindly accepting facts) an action is meaningless.

Whether or not this conjecture is plausible, there appears to be no corresponding simple trick when it comes to acquiring know-that. To note that some action in some context yielded some result does not imply that the agent *knows* this, although it seems likely that not storing such information would equally prevent any know-that about it. Regrettably, know-that is the kind of knowledge we usually mean when we talk of understanding in its noun sense. Just what kind of act would be required to produce an item of know-that I have no particular idea. The research I have read has not been particularly helpful in identifying specific acts which might be emulated. So there will be no hint in this paper of any mechanisms for developing identifiable know-that.

Not all is lost, however. No artificial agent may be able to understand in this way (yet), but *we* can. For example, we can usefully relate the above suggestions to the argument that "strong clues" are necessary for learning, and make some additional observations. That is, providing pertinent contexts for actions which the agent can accept or reject as it chooses, may well be a sufficiently strong form of clue which is consistent with having an agent acquire meaning. The more *useful* an action is in a context, as assessed by an *objective* observer — such as a critic mechanism within a learning agent — the more meaningful one could argue that action becomes. The action does not have to change to increase that usefulness or meaningfulness — just consider the act of eating in the contexts of being hungry versus it being a certain time of the day or it being windy.

However, an agent is not compelled to be objective, and can virtually choose any *subjective* context it likes to give meaning to its actions — eating at sunrise and sunset can be just as effective despite not addressing the underlying problem directly. The same argument applies to other aspects related to understanding: an artificial agent which is effective due to making subjective choices, and increases that effectiveness through learning, may be easier to produce and consequently preferable to one in which optimality and/or objectivity is explicitly required. Clues of this and other forms are certainly necessary to provide an *inductive bias* [15] so that an agent is able to learn at all, but strong clues, for all their good intention, may equally bias an agent away from perceiving and acting in its world in useful ways.

To summarise, the strongest argument against a learning-based approach is founded on a limited view of what is involved in understanding and knowledge. It identifies understanding with explicit, declarative *knowing that*, which is popular, but it does not tell the whole story. Even so, it seems likely that acquiring "know-that" involves know-how, and this know-how may arise when actions are executed in useful contexts. Noting these actions, and their associated contexts and results are likely to be vital in producing know-that. If this is indeed the way know-that is to arise, I perceive a counter-argument to taking a learning approach.

7 Two Major Problems

One might ask what is wrong with programming an intelligent system such as a mobile robot "normally". In doing so, one prescribes actions which are taken in appropriate contexts, and therefore could be argued to have meaning in them which could conceivably be exploited. But who is it that puts the actions in context, and thereby gains the meaning if there is any to gain? Clearly it is the programmer, not the system in question, and consequently there can be no understanding in such a system via this route alone.

The alternative route to understanding via conventional programming would be to argue that meaning is somehow inherent in the system and only needs to be exploited. Once again, I don't see any particular merit in this possibility nor do I see precisely how it could be, so I cannot mount a direct argument against it. However, it is clear that any such meaning may provide the foundation for the system's understanding. One could argue about how thick this foundation should be, and from what starting point a system should learn. The thinner this foundation, the less biased the agent will be towards learning particular things; it will be able to keep "an open mind" so to speak. Whether this is good or bad is perhaps a matter of personal opinion.

My opinion is that the weaker this bias, the more likely the learner is to "see things through its own eyes" and find a form of understanding which is more efficient and natural to it. I believe this to be the sort of understanding we desire in the longer term, although it would come at a price — the learner would have very little to guide it through an enormous space of possibilities, and would take far longer to learn as a result.

Nevertheless, it seems useful to consider the implications of starting with a minimal bias — no know-how per se, just a *meaningless* set of actions. This is also consistent with the possibility that meaning cannot be inherent in an agent when created. Also, if the conjecture regarding context made earlier is correct, then none of the actions should be associated with any contexts; any such *instincts* which do exist will have to be understood in other ways if they are to be understood at all. The assumption that an agent should start with no meaning whatsoever seems a useful one to make, and will apply from this point onwards.

There are some immediate implications of making this assumption. Such a set of actions would be of no use unless the agent can learn to describe the conditions under which each action should be deployed. In other words, the agent requires some sense of state to provide the necessary context for each action. This state may be explicitly represented as such, or merely implicit in the conditions learned for invoking each action.

We are therefore presenting two credit assignment problems to our agent: the first, structural — finding the features which best describe its state with respect to its environment at any time; the second, temporal — determining which actions in sequence achieve whatever result is desired from each state. The conjecture that contexts give meaning to actions could therefore be interpreted as saying that solving the structural credit assignment problem is an act of understanding.

Unfortunately there seems to be a circular dependence between these two problems and their solution is not straightforward. In the temporal credit assignment problem, a learning agent seeks the best action for its state; in the structural credit assignment problem, it seeks the best state for each action. Without finding useful (meaningful?) features to provide context, it cannot effectively prescribe suitable actions; without actions taken in context, it may not gain useful (meaningful?) experience from which it could discover the salient features in its world.

The agent is rescued from this dilemma by the notions of subjectivity and objectivity. The dilemma only occurs if we require that any behaviour exhibited by a learning agent be useful in an objective sense. Do we have any good reason to demand this at all times? We would certainly like an intelligent agent to take objectively useful actions in objectively sensible contexts in due course, but knowing that it must learn these things we must be prepared to give it the time to do so. We should therefore find it acceptable if the agent forms its own subjective view of the world, and gradually makes this more objective over time.

However, to say that an agent can contextualise its actions in any way it likes is a little misleading. An agent will be limited in the things it can use to do this. Loosely speaking, the only things it has available to provide contexts for its actions in the first instance are other actions it has taken and/or the results it has perceived from taking them. The reason for this is that the agent will start with no meaning, and the only thing it can do to change this fact is act. The only ways actions can make a meaningful difference is through some perception that an action or actions have been executed and/or some equivalent perception of what happened as a result of that action or actions.

Having talked at times of "first principles" and "atomic" common sense, we now see that, at the lowest level, a learning agent's knowledge can only be expressed in terms of actions and the perceived results of those actions. So the crucial issue of knowledge representation, which has been the preoccupation of many AI researchers for most of its history, becomes highly constrained. There is both good and bad to this: it makes the problem far harder, as can be seen in the discussion and application of a particular representation of this kind in [18], but it also focuses our efforts.

An important secondary point is that action cannot effect change to the agent's subsequent activity unless there is perception as well. Furthermore, it does an agent no immediate good whatsoever to store away some perceived result of an action in some part of memory. The agent can no longer perceive such a result; it must recall it if it needs it — by taking an action to do so. It is difficult to imagine how these things might occur other than in some controlling centre of an artificial intelligence — something analogous to our human consciousness. Regardless, consciousness is an important phenomenon to consider in any discussion of intelligence, and one which might provide another objection to the learning approach to AI.

8 Consciousness

An argument which some people might raise in response to this paper is that quite simply I have "missed the boat". Rather than understanding being key to intelligence, they would argue that its most important component is consciousness. The most controversial proponent of this point of view in recent times has perhaps been Roger Penrose [17]. While there are certainly more moderate perspectives, I have not encountered any which has categorically defined consciousness and its relationship to intelligence so that we can assess its importance. As a result it is hard to mount an argument against this position. However, if we consider recent arguments a little further, we can see that consciousness may be something that will take care of itself.

A plausible perspective on consciousness is that it is *epiphenomenal*, in other words, something which emerges out of other phenomena which are not themselves conscious or designed to produce consciousness. This viewpoint suggests that consciousness is related to the brain in a way which is similar to how shadows relate to objects — consciousness is intimately related to brain activity, but is merely a reflection of it, not something which has the power to influence that activity.

In line with this view, it does not seem unreasonable that consciousness may simply be an artefact of our need to abstract ourselves away from the immediate space of our own actions and their results (or the motor and sensory nerve signals which implement them) and come to operate in a world described in terms of space, time and other concepts. Consciousness could simply be the sensation of perceiving the world the way we do.

This line of thinking has not yet been developed or explored much beyond what is stated here. Since it is also unlikely to be well received, I will not devote any more time or space here to this possibility. The only statement about consciouness which is likely to bring any agreement at present is that it remains a controversial issue, and this is likely to be the case for some time. Even if it proves to be the most important element of intelligence, until we have a much clearer idea of what it is, arguments about consciousness seem unlikely to advance artificial intelligence in a practical sense.

9 Implications

There are a few lesser objections to the learning approach to AI which conincidentially serve as useful introductions to some of the implications of the arguments I have made above. For example, conventional AI researchers might emphasise goal-directed problem solving as the most important part of intelligent activity. People holding this opinion might ask why we should demand understanding when, for example, provided a robot solves the problem of carrying out the correct actions in response to some voiced command, that is all we require.

There is little one can do to defend against this argument other than note that the primary tools of these researchers are knowledge and reasoning. I would be very surprised if a system could be produced on this basis which was demonstrably robust in an unpredictable ("real world") environment and in which there was clearly no understanding occurring in the sense I have discussed. By focusing on understanding, I would hope to be inclusive rather than dismissive of work which takes this perspective on intelligence.

But it is clear to me that the basis upon which this work is currently undertaken is untenable, insofar as it genuinely seeks to produce artificial intelligence. Autonomy tends to be very limited in these systems, — even in many of those that learn — and the less control an agent has over its affairs, the less meaning it can hope to acquire and exploit as the basis for its understanding. Declaring that AI is dead, insofar as we mean the dominant "knowledge-based", problem-solving approach to it, seems as justified as declaring that the learning approach to AI is dead, as was quoted earlier.

Of course, I have argued that the learning approach is very much alive. Equally, even though I am convinced that due recognition of the roles of meaning and understanding is long overdue in AI circles, I see no reason to abandon any of the work currently undertaken under the AI banner. After all, once we humans have come to grips with the basics of living and laid the foundations of our common sense, we advance our knowledge principally by learning about and adopting what others have done before us. Our own experience will undoubtedly prove invaluable once artificial systems start to genuinely understand our world.

So the title of this paper is a double-entendre: yes, traditional approaches to AI will not achieve intelligence at all, but, even though we should devote more resource to examining issues of meaning and understanding, there is genuine value in the AI work which has been done. There also appears to be new life in an approach to AI which had previously been declared dead and buried.

The remaining arguments that I have encountered against the learning approach are comparitively minor. Something of a "straw man" argument comes from Herbert Simon, Nobel Laureate and respected AI pioneer. While apparently committed to the learning cause, he chose to question why learning should have any place in AI systems at all [21]. His argument can be loosely summarised as "why bother programming a computer to learn to do something, when it is far easier to program it to do the thing directly". This presumes that if one has the specific knowledge available to program a machine to learn to perform a specific task, then one also has the knowledge to program the machine to do the task itself.

While this may be true, can we really expect it to scale up? Our world is simply too diverse, dynamic and complex to expect that we could ever program all the knowledge an intelligent system might need to get by in it. If we are not to program these systems with worldly knowledge, then who will?[7] The obvious candidates are the systems themselves, that is, they should learn. Of course, this does not imply that systems should learn

[7]Rodney Brooks has objected that worldly knowledge need not be programmed into autonomous mobile robots at all. In [18] I argue that this argument also suffers from what amounts to a scaling problem. Yes, there are certain types of intelligent behaviour — even forms of reasoning and planning [5] — which might plausibly arise through interactions with the environment, but it is more plausible that we will need some form of world modelling to do these things in the sophisticated way we desire.

virtually everything through experience. Surely if we are able to concisely describe certain things, it would be advantageous to program these descriptions into these systems?

My response to this suggestion is just as pragmatic as Simon's argument. If one intends to learn in the longer term, then, especially given the doubts raised here over the absolute need to have existing knowledge to acquire more knowledge, why not learn in the short term as well? Why change horses in midstream so to speak? If one has useful information, then it seems more defensible to *teach* it to the agent, rather than program it, if one seeks a mature, experienced sort of agent that one can have confidence in. And if what I have argued earlier has merit, then an artificially intelligent agent will need this ability to learn from the moment it finds itself in a position to experience new things, that is, from the outset.

Lastly, and in a similar vein, it may be that I am vastly overestimating the extent to which knowledge, as I have discussed it, dictates our behaviour. It may be that beliefs underlie a lot of what we consider to be intelligence. If this were the case, it might seem to make sense to program things in the current fashion, and hope that, in due course, artificially intelligent agents could make genuine sense of these things.

There are three things to say about this. First, while beliefs may be at the heart of much of our intelligence — for example, in the statement of Hooke's Law which gave me the courage to have a bungy cord tied to my ankles and launch myself off a bridge — we can gain meaning and understanding from beliefs by placing them in hypothetical contexts ("if Hooke's Law were true, then I should survive this…"), or by reasoning about them and relationships between them (…since when an object attached to bungy cord falls to its lowest point where the cord is most likely to break, that object must be essentially stationary and a fall from there would be safe). Provided this understanding is borne out by experience, we have no need to challenge these beliefs. While they may be cornerstones of our existance, they do not necessarily spread like a cancer through our knowledge, allowing only other beliefs to arise from them.

Second, the only way an agent will in due course make sense of any of the beliefs we try to give it, will be in the way I have described above, namely by learning in a way which leads to meaning and understanding. Last, the further we go down the path of encoding more and more beliefs using so-called knowledge representations, the bigger the potential waste of effort if it proves that these representations cannot be adopted by agents with the capacity to understand. There are already enough eggs in this basket. I suggest it is time to transfer a good deal of our efforts from this potentially fruitless task to one which, while difficult, at least addresses the fundamental issue of intelligence.

10 Conclusion

AI systems are not what they used to be. In fact, they never were — fifty years of advances in computing technology has done little to turn the wishful thoughts of yesteryear into practical sophisticated systems for today. Despite a seemingly reasonable focus during much of this period on knowledge and reasoning, these things are impotent unless placed in the wider context of understanding, and allowed to take on real meaning.

A failure to pinpoint precisely what meaning is does not prevent us from placing limits on how it may be exploited as a basis for understanding, provided we accept the widely held viewpoint that meaning is purely an internal, "personal" property of intentional objects, that meaning and understanding are inseparable, and that there are only two mechanisms by which understanding may occur. Together these imply that meaningful knowledge cannot be transferred from an agent of one kind (e.g. human) to one of a different kind (e.g. a robot). All knowledge must be regenerated in a receiving agent *by that agent* in order to be genuinely understood by it. Understanding is an act of attributing meaning to something, and this meaning can only be arise through learning.

The learning approach to AI, while having considerable intuitive appeal, has been discredited in the past. However, the strongest of the arguments against it fails to account for how knowledge first comes into existence. It ignores issues of symbol grounding and, in effect, identifies meaning with representation. Other arguments can be mounted against the learning approach, but none are particularly strong. Perhaps the greatest concern about this approach is making it practical.

In this regard, prior to, if not concurrent with, addressing the issue of symbol grounding, learning agents will encounter two other important problems: the structural credit assignment problem and the temporal credit assignment problem. The only things an agent can use to represent the solutions to these problems in a meaningful way are the actions it takes in its environment, and the results it perceives from doing so.

These considerations will restrict the kinds of learning techniques which can be used to achieve artificial intelligence and understanding. Nevertheless, there is merit in persisting with otherwise unsuitable methods, since they will likely bring at least insights which will benefit agents capable of understanding them.

But the central message is clear: AI, as it is practiced by most researchers, is dead. It cannot hope to achieve its aim because there is no meaning in the systems it creates; there is only meaning in the minds of its creators. So let us embark on a quest for meaning: what it is, how it arises, and how we can use it to create the kind of systems that people had in mind when they coined the term "artificial intelligence". Long live *that* kind of AI!

References

[1] John R. Anderson. *Cognitive Psychology and its Implications.* W. H. Freeman, New York, 1995. (4th Edition).

[2] Paul Armer. Attitudes towards intelligent machines. *Symposium on Bionics*, 1960. Reprinted in Edward A. Feigenbaum and Julian Feldman (eds), *Computers and Thought*, McGraw-Hill, 1963, pp. 389–405.

[3] Rodney A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2(1):14–23, March 1986.

[4] Rodney A. Brooks. Intelligence without reason. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, pages 569–595, 1991.

[5] David Chapman and Philip E. Agre. Abstract reason as emergent from concrete activity. In M. P. Georgeff and A. L. Lansky, editors, *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop at Timberline, Oregon*, pages 411–424, Los Altos, California, 1987. Morgan Kaufmann.

[6] Eugene Charniak and Drew McDermott. *Introduction to Artificial Intelligence.* Addison-Wesley, 1985.

[7] Marco Dorigo. Introduction to the special issue on learning autonomous robots. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 26(3):361–364, 1996.

[8] Walter Freeman. A neurobiological interpretation of semiotics: Meaning vs. representation. In *Proc. IEEE Systems, Man and Cybernetics Conference*, pages 1481–1486, Orlando, Florida, October 1997.

[9] Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[10] John E. Laird, Allen Newell, and Paul S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

[11] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, November 1995.

[12] Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. CYC: Toward programs with common sense. *Communications of the ACM*, 33(8):30–49, August 1990.

[13] John McCarthy. Book review: “What computers still can’t do” by Hubert Dreyfus. *Artificial Intelligence*, 80(1):143–150, 1996.

[14] Marvin Minsky. Steps towards artificial intelligence. *Proc. Institute of Radio Engineers*, 49:8–30, January 1961. Reprinted in Edward A. Feigenbaum and Julian Feldman (eds), *Computers and Thought*, McGraw-Hill, 1963, pp 406–450.

[15] Tom M. Mitchell. The need for biases in learning generalization. Tech. Rep. CBM-TR-117, Dept. of Computer Science, Rutgers University, New Brunswick, N.J., 1980.

[16] A. Newell and G. W. Ernst. *GPS: A Case Study in Generality and Problem Solving.* Academic Press, 1969.

[17] Roger Penrose. *The Emperor’s New Mind.* Oxford University Press, 1989.

[18] Paul Qualtrough. *Learn First, Ask Questions Later: a Meaningful Approach to Robotics.* PhD thesis, Department of Computer Science, University of Auckland, New Zealand, 1998.

[19] Gilbert Ryle. Knowing how and knowing that. In *Collected Papers*, volume 2, pages 212–225. Hutchinson, London, 1971.

[20] Gilbert Ryle. The theory of meaning. In *Collected Papers*, volume 2, pages 350–372. Hutchinson, London, 1971.

[21] Herbert A. Simon. Why should machines learn? In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 25–37. Tioga, 1983.

[22] Mel Thompson. *Teach Yourself Philosophy*. Hodder and Stoughton, London, 1995.

[23] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, October 1950. Reprinted in Edward A. Feigenbaum and Julian Feldman (eds), *Computers and Thought*, McGraw-Hill, 1963, pp. 11–35.

[24] Stewart W. Wilson. The animat path to AI. In *Proceedings of the 1990 Conference on Simulating Adaptive Behavior*, pages 15–21, 1990.

[25] Terry Winograd. What does it mean to understand language? *Cognitive Science*, 4(3):209–242, July/September 1980. Also in D. Norman (ed.), *Perspectives on Cognitive Science*, Ablex and Erlbaum Associates, 1981, pp. 231–264.

[26] Patrick Henry Winston and Sarah Alexandra Shellard. *Artificial Intelligence at MIT: Expanding Frontiers*. MIT Press, 1990.