On our best behaviour

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu

Abstract

The science of AI is concerned with the study of intelligent forms of behaviour in computational terms. But what does it tell us when a good semblance of a behaviour can be achieved using cheap tricks that seem to have little to do with what we intuitively imagine intelligence to be? Are these intuitions wrong, and is intelligence really just a bag of tricks? Or are the philosophers right, and is a behavioural understanding of intelligence simply too weak? I think both of these are wrong. I suggest in the context of question-answering that what matters when it comes to the science of AI is not a good semblance of intelligent behaviour at all, but the behaviour itself, what it depends on, and how it can be achieved. I go on to discuss two major hurdles that I believe will need to be cleared.

1 Intelligent behaviour

This paper[1] is about the *science* of AI. Unfortunately, it is the *technology* of AI that gets all the attention. The general public could be forgiven for thinking that AI is just about all those whiz-bang applications, smart *this* and autonomous *that*. Those of us in the field know that for many applications, the term "intelligent" is no more than a buzzword (like the term "delicious" in "red delicious apples"). And along with the many possibly beneficial AI applications under consideration, we often have serious misgivings about the potential misuse of AI technology (in areas like weaponry).

But AI is more than just technology. Many of us are motivated not by any of the AI applications currently being considered, but by the scientific enterprise, the attempt to understand the world around us. Different sciences have different subject matters, and AI is the study of *intelligent behaviour* in computational terms. What could be more fascinating? The human brain is a remarkable thing, perhaps the single most complex object we know of in the universe. But even more remarkable is what a human brain is capable of *doing*. Our intelligent behaviour at its best goes well beyond what we have any right to expect to emerge out of purely physical matter. Indeed, the overarching question for the science of AI is:

> How is it possible for something physical (like people, for instance) to actually do X?

where X is one of the many instances of intelligent behaviour. This needs to be contrasted with a related question:

> Can we engineer a computer system to do something that is vaguely X-ish?

about which we will have much more to say later.

Note that the science of AI studies intelligent behaviour, not *who* or *what* is producing the behaviour. It studies natural language understanding, for instance, not natural language understanders. This is what makes AI quite different from the study of *people* (in neuroscience, psychology, cognitive science, evolutionary biology, and so on).

What sort of behaviour do we care about? Different researchers will quite naturally focus on different aspects. The behaviour may or may not depend on perceptual or motor skills. It may or may not include learning. It may or may not be grounded in emotional responses, or in social interactions. For some researchers, the main concern is intelligent behaviour seen in a variety of animals, like the ability to find a desired object in a room. For others, the focus is on behaviour seen in humans only, like the ability to play chess. (These two groups sometimes engage in methodological disputes, with the former arguing that we cannot expect to understand human behaviour until we understand its more basic forms, and the latter responding that this is not how science works at all. At this stage of the game, there is really no reason to take a doctrinaire position one way or another.)

1.1 Answering questions

In this paper, I intend to examine one basic form of intelligent behaviour: answering certain *ad-hoc* questions posed in English. Consider a question like the following:

> *Could a crocodile run a steeplechase?*

Even if you know what crocodiles and steeplechases are,[2] you have never really thought about this question before, unless you happened to have read an early paper of mine [6]. Nor

[1] This paper is a written version of the Research Excellence Lecture presented in Beijing at the IJCAI-13 conference. Thanks to Vaishak Belle and Ernie Davis for helpful comments.

[2] For those who do not know, a steeplechase is a horse race, similar to the usual ones, but where the horses must jump over a number of hedges on the racetrack. So it is like hurdles for horses.

can you simply look up the correct answer somewhere. And yet, an answer does occur to you almost immediately. Here is another question from the same paper:

> *Should baseball players be allowed to glue small wings onto their caps?*

Again, you have never thought of this before, but again an answer occurs to you. (In this case, you might even wonder if there is some sort of trick to the question that you may have missed. There is none.)

In this paper, I want to consider our ability to answer one-shot questions like these, and for four reasons:

1. This is behaviour that is clearly exhibited by people. We are indeed capable of answering questions like these without any special training or instructions.
2. This is behaviour that is difficult to crack. We have as yet no good idea about what people do to answer them. No existing computer program can duplicate our ability.
3. Our behaviour in answering questions like these appears to underly other more complex (and more ecologically significant) forms of behaviour.
4. Being clear and precise about the form of behaviour we care about even in this simple case will also help clarify what it means for the science of AI to be successful.

As we will see, however, there will be good reasons to move to answering questions of a more restricted form.

2 Behavioural tests

Given some form of intelligent behaviour, how do we know that the computational story told by AI researchers actually explains the behaviour. The answer, going all the way back to Turing, is this: a computational account is adequate if it is able to generate behaviour that cannot be distinguished over the long haul from the behaviour produced by people.

This, of course, harks back to the famous Turing Test [11]. We imagine an extended conversation over a teletype between an interrogator and two participants, a person and a computer. The conversation is natural, free-flowing, and about any topic whatsoever. The computer is said to *pass the Turing Test* if no matter how long the conversation, the interrogator cannot tell which of the two participants is the person.

Turing's point in all this, it seems to me, is this: Terms like "intelligent," "thinking," "understanding," and the like are much too vague and emotionally charged to be worth arguing about. If we insist on using them in a scientific context at all, we should be willing to say that a program that can pass a suitable behavioural test has the property in question as much as the person. Adapting the dictum of the movie character Forest Gump who said "*Stupid is as stupid does*," we can imagine Turing saying "*Intelligent is as intelligent does*." This is a very sensible position, it seems to me, and I have defended it elsewhere [7].

2.1 The trouble with the Turing Test

However, I do feel that the Turing Test has a serious problem: it relies too much on *deception*. A computer program passes the test iff it can *fool* an interrogator into thinking she is dealing with a person not a computer. Consider the interrogator asking questions like these:

> *How tall are you?*

or

> *Tell me about your parents.*

To pass the test, a program will either have to be evasive (and duck the question) or manufacture some sort of false identity (and be prepared to lie convincingly). In fact, evasiveness is seen quite clearly in the annual *Loebner Competition*, a restricted version of the Turing Test.[3] The "chatterbots" (as the computer entrants in the competition are called) rely heavily on wordplay, jokes, quotations, asides, emotional outbursts, points of order, and so on. Everything, it would appear, except clear and direct answers to questions!

The ability to fool people is interesting, no doubt, but not really what is at issue here.[4] We might well ask: is there a better behaviour test than having a free-form conversation?

There are some quite reasonable non-English options to consider, such as "captchas" [12] and the program at `www.areyouhuman.com`. But English is an excellent medium since it allows us to range over topics broadly and flexibly (and guard for biases: age, education, culture, *etc.*).

But here is another option: what if instead of a conversation, the interrogator only asks a number of *multiple-choice questions*? This has some distinct advantages:

- Verbal dodges are no longer possible. A program can no longer game the test using evasive maneuvers.
- It does not require the ability to generate "credible" English. The program will not need to worry about choosing words or syntax to accurately mimic actual speakers.
- The tests can be automated (administered and graded by machine). Success on the test does not depend on the judged similarity to people, but on the correctness of the answers.

2.2 Cheap tricks

We want multiple-choice questions that people can answer easily. But we also want to avoid as much as possible questions that can be answered using cheap tricks (*aka* heuristics).

Consider for example, the question posed earlier:

> *Could a crocodile run a steeplechase?*
> - *yes*
> - *no*

The intent here is clear. The question can be answered by thinking it through: a crocodile has short legs; the hedges in a steeplechase would be too tall for the crocodile to jump over; so no, a crocodile cannot run a steeplechase.

The trouble is that there is another way to answer the question that does not require this level of understanding. The idea is to use the *closed world assumption* [10; 3]. This assumption says (among other things) the following:

[3] See the book by Brian Christian [2] for an interesting account of what it was like to play the human in a Loebner contest.

[4] The ELIZA program [13] is a good place to start on that issue.

> If you can find no evidence for the existence of something, assume that it does not exist.

For the question above, since I have never heard of a crocodile being able to run a steeplechase, I conclude that it cannot. End of story. Note that this is a *cheap* trick: it gets the answer right, but for dubious reasons. It would produce the wrong answer for a question about *gazelles*, for example. Nonetheless, if all we care about is answering the crocodile question correctly, then this cheap trick does the trick.

Can we find questions where cheap tricks like this will not be sufficient to produce the desired behaviour? This unfortunately has no easy answer. The best we can do, perhaps, is to come up with a suite of multiple-choice questions *carefully* and then study the sorts of computer programs that might be able to answer them. Here are some obvious guidelines:

- Make the questions Google-proof. Access to a large corpus of English text data should not *by itself* be sufficient.
- Avoid questions with common patterns. An example is "*Is x older than y?*" Perhaps no single Google-accessible web page has the answer, but once we map the word "older" to "birth date," the rest comes quickly.[5]
- Watch for unintended bias. The word order, vocabulary, grammar and so on all need to be selected very carefully not to betray the desired answer.

One existing promising approach in this direction is the *recognizing textual entailment* challenge [4; 1]. But it has problems of its own, and so here we propose a different one.

3 Winograd schema questions

Our approach is best illustrated with an example question:[6]

> *Joan made sure to thank Susan for all the help she had given. Who had given the help?*
> - *Joan*
> - *Susan*

A *Winograd schema question* is a binary-choice question with these properties:

- Two parties are mentioned in the question (both are males, females, objects, or groups).
- A pronoun is used to refer to one of them ("he," "she," "it," or "they," according to the parties).
- The question is always the same: what is the referent of the pronoun?
- Behind the scenes, there are two *special words* for the schema. There is a slot in the schema that can be filled by either word. The correct answer depends on which special word is chosen.

In the above, the special word used is "given," and the other word is "received." So each Winograd schema actually generates two very similar questions:

> *Joan made sure to thank Susan for all the help she had **given**. Who had **given** the help?*
> - *Joan*
> - *Susan* ✓

and

> *Joan made sure to thank Susan for all the help she had **received**. Who had **received** the help?*
> - *Joan* ✓
> - *Susan*

It is this one-word difference between the two questions that helps guard against using the cheapest of tricks on them.

Here are some additional examples. The first is one that is suitable even for young children:

> *The trophy would not fit in the brown suitcase because it was so small. What was so small?*
> - *the trophy*
> - *the brown suitcase*

In this case, the special word used is "small" and the other word is "big." Here is the original example due to Terry Winograd [14] for whom the schema is named:

> *The town councillors refused to give the angry demonstrators a permit because they feared violence. Who feared violence?*
> - *the town councillors*
> - *the angry demonstrators*

Here the special word is "feared" and the alternative word is "advocated."

With a bit of care, it is possible to come up with Winograd schema questions that exercise different kinds of expertise. Here is an example concerning certain materials:

> *The large ball crashed right through the table because it was made of styrofoam. What was made of styrofoam?*
> - *the large ball*
> - *the table*

The special word is "styrofoam" and the alternative is "steel." This one tests for problem-solving skill:

> *The sack of potatoes had been placed below the bag of flour, so it had to be moved first. What had to be moved first?*
> - *the sack of potatoes*
> - *the bag of flour*

The special word is "below" and the alternative is "above." This example tests for an ability to visualize:

> *Sam tried to paint a picture of shepherds with sheep, but they ended up looking more like golfers. What looked like golfers?*
> - *the shepherds*
> - *the sheep*

The special word used is "golfers" and the other is "dogs."

[5] The program at `www.trueknowledge.com` appears to work this way.

[6] This section is drawn mainly from [8]. I thank Ernie Davis and Leora Morgenstern for their contribution.

Of course not just any question in this form will do the job here. It is possible to construct questions that are too "easy," like this one:

> *The racecar easily passed the school bus because it was going so fast. What was going so fast?*
> - *the racecar*
> - *the school bus* (Special=*fast;* other=*slow)*

The problem is that this question can be answered using the following trick: ignore the given sentence, and check which two words co-occur more frequently (according to Google, say): "racecar" with "fast" or "school bus" with "fast." Questions can also be too "hard," like this one:

> *Frank was jealous when Bill said that he was the winner of the competition. Who was the winner?*
> - *Frank*
> - *Bill* (Special=*jealous;* other=*happy)*

The problem is that this question is ambiguous when the "happy" variant is used. Frank could plausibly be happy because he is the winner or because Bill is. Further discussion on these and other issues can be found in [8].

3.1 A new test

It is now possible to formulate an alternative to the Turing Test. A collection of pre-tested Winograd schemas can be hidden in a library.[7] A Winograd Schema Test involves asking a number of these questions with a strong penalty for wrong answers (to preclude guessing). A test can be administered and graded in a fully automated way:

1. select N (*e.g.*, $N = 25$) questions that are suitable (with respect to vocabulary, expertise, *etc.*);
2. randomly use one of the special words in the question;
3. present the test to the subject, and obtain the N binary replies;

The final grade for the test is

$$\frac{\max(0,\ N - k \cdot Wrong)}{N}$$

where k codes the penalty for guessing (*e.g.*, $k = 5$). The main claim here is that normally-abled English-speaking adults will pass the test easily. So, if we want to produce behaviour that is indistinguishable from that of people, we will need to come up with a program that can also pass the test.

To summarize: With respect to the Turing Test, we agree with Turing that the substantive question is whether or not a certain *intelligent behaviour* can be achieved by a computer program. But a free-form *conversation* as advocated by Turing may not be the best vehicle for a formal test, as it allows a cagey subject to hide behind a smokescreen of playfulness, verbal tricks, and canned responses. Our position is that an alternative test based on Winograd schema questions is less subject to abuse, though clearly much less demanding intellectually than engaging in a cooperative conversation (about sonnets, for example, as imagined by Turing).

[7] See, for example, the collection at `http://www.cs.nyu.edu/faculty/davise/papers/WS.html`.

4 Passing the test

What would it take for a computer program to pass a Winograd Schema Test. My feeling is that we can go quite some distance with the following:

1. Take a Winograd schema question such as

   > *The trophy would not fit in the brown suitcase because it was so small. What was so small?*
   > - *the trophy*
   > - *the brown suitcase*

   and parse it into the following form:

   > Two parties are in relation R.
   > One of them has property P. Which?

   For the question above, this gives the following:

   > R = does not fit in; P = is so small.

2. Then use *big data*: search all the English text on the web to determine which is the more common pattern:
   - x does not fit in y **+** x is so small *vs.*
   - x does not fit in y **+** y is so small

This "big data" approach is an excellent trick, but unfortunately, it is still too cheap. Among other things, it ignores the *connective* between R and P. Consider this:

> *The trophy would not fit in the brown suitcase* ***despite the fact that*** *it was so small. What was so small?*
> - *the trophy*
> - *the brown suitcase*

Note that the R and P here would be the same as before, even though the answer must be different this time.

Now consider the following example:

> *Fred is the only man alive who still remembers my father as an infant. When Fred first saw my father, he was twelve years old. Who was twelve years old?*
> - *Fred*
> - *my father* (Special=*years;* other=*months)*

Here the relationship between any R and P is clearly much more complex.

So what do we conclude from this? Do we simply need a bigger bag of tricks?

4.1 The lure of statistics

There is a tendency in AI to focus on behaviour in a purely statistical sense. We ask:

> Can we engineer a system to produce a desired behaviour with no more errors than people would produce (with confidence level z)?

Looking at behaviour this way can allow some of the more challenging examples that arise (like the question concerning Fred above) to simply be *ignored* when they are not statistically significant.

Unfortunately, this can lead us to systems with very impressive performance that are nonetheless *idiot-savants*. We

might produce prodigies at chess, face-recognition, *Jeopardy*, and so on, that are completely hopeless outside their area of expertise.[8]

But there is another way of looking at all this. Think of the behaviour of people on Winograd schema questions as a *natural phenomenon* to be explained, not unlike photosynthesis or gravity. In this case, even a *single example* can tell us something important about how people are able to behave, however insignificant statistically.

4.2 A thought experiment

Reconsider, for instance, the styrofoam / steel question from above. We might consider using other special words in the question: for "balsa wood," the answer would be "the table," for "granite," it would be "the large ball," and so on. But suppose we use an unknown word in the question:

> *The large ball crashed right through the table because it was made of XYZZY. What was made of XYZZY?*
> - *the large ball*
> - *the table*

Here there is no "correct" answer: subjects should not really favor one answer much over the other.

But suppose we had told the subjects some facts about the XYZZY material:[9]

1. It is a trademarked product of Dow Chemical.
2. It is usually white, but there are green and blue varieties.
3. It is ninety-eight percent air, making it lightweight and buoyant.
4. It was first discovered by a Swedish inventor, Carl Georg Munters.

We can ask, on learning any of these facts, at what point do the subjects stop guessing? It should be clear that only one of these facts really matters, the third one. But more generally, people get the right answer for *styrofoam* precisely because they already know something like the third fact above about the makeup of styrofoam. This background knowledge is critical; without it, the behaviour is quite different.

4.3 The lesson

So what do we learn from this experiment about the answering of Winograd schema questions? From a pure *technology* point of view, a reasonable question to ask here is this:

> Can we produce a good semblance of the target behaviour without having to deal with background knowledge like this?

But from a *science* point of view, we must take a different stance. We want to understand what it takes to produce the intelligent behaviour that people exhibit. So the question really needs to be more like this:

> What kind of system would have the necessary background knowledge to be able to behave the way people do?

4.4 A radical approach

So to account for what people are actually able to do, we need to consider what it would take to have a system that *knows* a lot about its world and can apply that knowledge as needed, the way people can.

One possibility is this:

- some part of what needs to be known is represented symbolically (call it the knowledge base);
- procedures operate on this knowledge base, deriving new symbolic representations (call it reasoning);
- some of the derived conclusions concern what actions should be taken next (including answering questions).

This is a very radical idea, first proposed by John McCarthy in a quite extraordinary and unprecedented paper [9]. It suggests that we should put aside any idea of tricks and shortcuts, and focus instead on what needs to be *known*, how to represent it symbolically, and how to use the representations.

5 Two scientific hurdles

I do not want to suggest that with McCarthy's radical idea on board, it is all smooth sailing from here. A good question to ask is why, after 55 years, we have so little to show for it regarding the science of intelligent behaviour. The answer, I believe, is that it leaves some major issues unresolved.

My Computers and Thought Lecture at IJCAI-85 [5] was in part a reaction to the "*Knowledge is Power*" slogan which was quite in vogue at the time. It all seemed too facile to me, even back then. My sense was that knowledge was *not* power if it could not be acquired in a suitable symbolic form, or if it could not be applied in a tractable way. These point to two significant hurdles faced by the McCarthy approach:

1. Much of what we come to know about world and the people around us is not from personal experience, but is due to our use of *language*.

   > People talk to us, we listen to weather reports and to the dialogue in movies, and we read: text messages, sport scores, mystery novels, *etc*.

   And yet, it appears that we need to use extensive knowledge to make good sense of all this language.

2. Even the most basic child-level knowledge seems to call upon a wide range of logical constructs.

   > Cause and effect and non-effect, counterfactuals, generalized quantifiers, uncertainty, other agents' beliefs, desires and intentions, *etc*.

   And yet, symbolic reasoning over these constructs seems to be much too demanding computationally.

I believe that these two hurdles are as serious and as challenging to the science of AI as an accelerating universe is to astrophysics. After 55 years, we might well wonder if an AI researcher will *ever* be able to overcome them.

[8] Indeed, it would be good fun to try *Watson* on Winograd schema questions: the category is "Pronoun referents," the clue is "Joan made sure to thank Susan for all the help she had given," and the desired answer in the form of a question is "Who is Susan?"

[9] These facts were lifted from the Wikipedia page for styrofoam.

Life being short (and "time to market" even shorter), it is perhaps not surprising that many AI researchers have returned to less radical methods (*e.g.*, more biologically-based, more like statistical mechanics) to focus on behaviours that are seemingly less knowledge-intensive (*e.g.*, recognizing handwritten digits, following faces in a crowd, walking over rough terrain). And the results have been terrific!

But these terrific results should not put us into denial. Our best behaviour *does* include knowledge-intensive activities such as participating in natural conversations, or responding to Winograd schema questions. It is my hope that enough of us stay focused on this sort of intelligent behaviour to allow progress to continue here as well.

This will require hard work! I think it is unreasonable to expect solutions to emerge spontaneously out of a few general principles, obviating any real effort on our parts. For example, I do not think we will ever be able to build a small computer program, give it a camera and a microphone or put it on the web, and expect it to acquire what it needs all by itself.

So the work will be hard. But to my way of thinking, it will be more like scaling a mountain than shoveling a driveway. Hard work, yes, but an exhilarating adventure!

5.1 Some suggestions

What about those hurdles? Obviously, I have no solutions. However, I do have some suggestions for my colleagues in the Knowledge Representation area:

1. We need to return to our roots in Knowledge Representation and Reasoning *for* language and *from* language.

   We should *not* treat English text as a monolithic source of information. Instead, we should carefully study how simple knowledge bases might be used to make sense of the simple language needed to build slightly more complex knowledge bases, and so on.

2. It is not enough to build knowledge bases without paying closer attention to the demands arising from their use.

   We should explore more thoroughly the space of computations between fact retrieval and full automated logical reasoning. We should study in detail the effectiveness of *linear* modes of reasoning (like unit propagation, say) over constructs that logically seem to demand more.

As to the rest of the AI community, I do have a final recommendation:

> We should avoid being overly swayed by what appears to be the most promising approach of the day.

As a field, I believe that we tend to suffer from what might be called *serial silver bulletism*, defined as follows:

> *the tendency to believe in a silver bullet for AI, coupled with the belief that previous beliefs about silver bullets were hopelessly naïve.*

We see this in the fads and fashions of AI research over the years: first, automated theorem proving is going to solve it all; then, the methods appear too weak, and we favour expert systems; then the programs are not situated enough, and we move to behaviour-based robotics; then we come to believe that learning from big data is the answer; and on it goes.

I think there is a lot to be gained by recognizing more fully what our own research does *not* address, and being willing to admit that other AI approaches may be needed for dealing with it. I believe this will help minimize the hype, put us in better standing with our colleagues, and allow progress in AI to proceed in a steadier fashion.

5.2 The prospects

Finally, let me conclude with a question about the future:

> Will a computer ever pass the Turing Test (as first envisaged by Turing) or even a broad Winograd Schema Test (without cheap tricks)?

The answer to this question, I believe, lies in a quote from Alan Kay: "*The best way to predict the future is to invent it.*" I take this to mean that the question is not really for the pundits to debate. The question, in the end, is really about *us*, how much perseverance and inventiveness we will bring to the task. And I, for one, have the greatest confidence in what we can do when we set our minds to it.

References

[1] D. G. Bobrow, C. Condoravdi, R. Crouch, V. de Paiva, L. Karttunen, T. H. King, R. Mairn, L. Price, A. Zaenen, Precision-focussed textual inference, *Proc. of the ACL Workshop*, Prague, 2007.

[2] B. Christian, *The Most Human Human*, Doubleday, 2011.

[3] A. Collins, E. Warnock, N. Aiello, M. Miller, Reasoning from incomplete knowledge, in *Representation and understanding*, Academic Press, 1975.

[4] I. Dagan, O. Glickman, B. Magnini, The PASCAL recognising textual entailment challenge, in *Machine Learning Challenges*, Springer Verlag, 2006.

[5] H. J. Levesque, Making believers out of computers, *Artificial Intelligence*, **30**, 1986.

[6] H. J. Levesque, Logic and the complexity of reasoning, *The Journal of Philosophical Logic*, **17**, 1988, 355–389.

[7] H. J. Levesque, Is it enough to get the behaviour right?, *Proc. of IJCAI-09*, Pasadena, CA, 2009.

[8] H. J. Levesque, E. Davis, L. Morgenstern, The Winograd Schema challenge, *Proc. of KR-2012*, Rome, 2012.

[9] J. McCarthy, The advice taker, in *Semantic Information Processing*, MIT Press, 1968.

[10] R. Reiter, On closed world databases, in *Logic and Databases*, Plenum Press, 1978.

[11] A. Turing, Computing machinery and intelligence, *Mind* **59**, 433–460, 1950.

[12] L. von Ahn, M. Blum, N. Hopper, J. Langford, CAPTCHA: Using Hard AI Problems for Security, in *Advances in Cryptology, Eurocrypt 2003*, 294–311.

[13] J. Weizenbaum, ELIZA, *CACM* **9**, 36–45, 1966.

[14] T. Winograd, *Understanding Natural Language*. Academic Press, New York, 1972.