On Our Best Behaviour

Hector J. Levesque
Dept. of Computer Science
University of Toronto

Thanks to my coauthors!

Fahiem Bacchus
Vaishak Belle
Ron Brachman
Phil Cohen
Ernie Davis
Jim Delgrande
Giuseppe de Giacomo
Jim des Rivières
Richard Fikes
Hojjat Ghaderi
Victoria Gilbert
Joe Halpern
Koen Hindricks
Toby Hu
Marcus Huber
Michael Jenkin
Sanjeev Kumar
Gerhard Lakemeyer
Yves Lespérance
Fangzhen Lin
Yongmei Liu
Jeff Lloyd
Daniel Marcu
Gord McCalla
David McGee
David Mitchell
John Mylopoulos
Leora Morgenstern
José Nunes
Sharon Oviatt
Maurice Pagnucco
Ron Petrick
Fiora Pirri
Ray Reiter
Shane Ruman
Sebastian Sardina
Richard Scherl
Bart Selman
Steven Shapiro
Ira Smith
Kenneth Tan
Stavros Vassos

AI science and technology

AI technology is what gets all the attention. But it has a downside too:

- systems that are intelligent in name only
- systems in dubious areas of application

But AI is more than just technology!

Instead: the study of intelligent forms of behaviour

How is it possible that people are able to do X?

vs.

Can we engineer a system to do something X-ish?

Not the study of *who* or *what* is producing the behaviour

(so $\neq$ neuroscience, psychology, cognitive science, ...)

Our best behaviour

What sort of intelligent behaviour do we care about?

Different researchers will focus on different aspects.

May or may not involve
- learning (via training or via language)
- perception or motor skills
- emotional responses or social interactions

For some: behaviour that is uniquely human.

For others: behaviour also seen in other animals.

Today, one seemingly simple form of intelligent behaviour:

responding to certain questions

"In science one can learn the most by studying the least."

— Marvin Minsky

Getting the behaviour right

When will we have accounted for some intelligent behaviour?

The answer from Turing: when the behaviour of an AI program is indistinguishable over the long haul from that produced by people.

> **The Turing Test**: 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.
>
> **Passing the Turing Test**: no matter how long the conversation, the interrogator cannot tell which of the two participants is the person.

Turing's point: if we insist on using vague terms like "intelligent," "thinking," or "understanding" at all, we should be willing to say that a program that can pass the behavioural test has the property as much as the person.

> *cf.* Forest Gump: "Stupid is as stupid does."

What is wrong with the Turing Test?

The problem with the Turing Test is that it is based on *deception*.

> A computer program is successful iff it is able to *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?
- Tell me about your parents.

To pass the Turing Test, the program will either have to be *evasive* or manufacture some sort of *false identity*.

Evasiveness is seen very clearly in the annual *Loebner Competition*, a restricted version of the Turing Test [Christian 11].

> The "chatterbots" often use wordplay, jokes, quotations, asides, emotional outbursts, points of order, etc.

Beyond a conversation

The ability to fool people is interesting, but not really what is at issue here.

cf. the ELIZA system [Weizenbaum 66]

Is there a better behaviour test than having a free-form conversation?

There are some very reasonable non-English options to consider.

e.g. "captchas" [von Ahn *et al* 03], also see `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.).

What if the interrogator only asks a number of *multiple-choice questions*?

- verbal dodges are no longer possible (so harder to game)
- does not require the ability to generate "credible" English
- tests can be automated (administered and graded by machine)

Answering questions

We want questions that people can answer easily using what they know.

But we also want to avoid as much as possible questions that can be answered using cheap tricks (*aka* heuristics).

Could a crocodile run a steeplechase? [Levesque 88]

- Yes
- No

The intended thinking: short legs, tall hedges $\Rightarrow$ No!

The cheap trick: the closed-world assumption [Reiter 78, Collins *et al* 75]

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

(Note: the heuristic gives the wrong answer for *gazelles* perhaps.)

Can cheap tricks be circumvented?

Maybe not. The best we can do is to come up with our questions *carefully,* and then study the sorts of programs that might pass the test.

Make the questions Google-proof: access to a very large corpus of English text data should not *by itself* be sufficient.

Our motto: *"It's not the size of the corpus; it's how you use it!"*

Avoid questions with common patterns: 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.

(This is largely how the program at `www.trueknowledge.com` works.)

Watch for unintended bias: word order, vocabulary, grammar etc.

One existing promising direction is the *recognizing textual entailment* challenge, but it has problems of its own. [Dagan *et al* 06, Bobrow *et al* 07]

A new proposal

[Levesque,Davis,Morgenstern 12]

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

- Joan
- Susan

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

- Two parties are mentioned (males, females, objects, groups).
- A pronoun is used to refer to one of them (he, she, it, they).
- The question is: 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 is received.

Two more examples

The original example due to Terry Winograd (1972):

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

The special word used is feared and the alternate is advocated.

An example involving visual resemblance:

> 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.

etc.

A Winograd Schema Test

A collection of pre-tested Winograd schemas can be hidden in a library.

(E.g. http://www.cs.nyu.edu/faculty/davise/papers/WS.html)

A Winograd Schema Test involves asking a number of these questions with a strong penalty for wrong answers (to preclude guessing).

The test can then be administered and graded in a fully automated way:

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

The final grade: $\dfrac{\max(0,\ N - k \cdot Wrong)}{N}$, (e.g. $k = 5$).

Claim: normally-abled English-speaking adults will pass the test easily!

Regarding the Turing Test ...

The question as to whether computers can or ever will *really* think (understand, be intelligent) remains as controversial as ever.

The Turing Test suggests that we should focus the question on whether or not a certain *intelligent behaviour* can be achieved by a computer program.

> Aside: The attempt to refute this by using an instruction book to produce the behaviour *sans* the understanding does not stand up. [Levesque 09]

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.

However: An alternate test based on Winograd schema questions is less subject to abuse, though clearly much less demanding intellectually.

What does it take to pass the test?

It is possible to go quite some distance with the following:

1. Parse the Winograd schema question into the following form:

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

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

   This gives R = does not fit in; P = is so small

2. 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

And the result is ...

This *big data* approach is an excellent trick, but it is still too cheap!

Among other things, it ignores the *connective* between R and P.

> 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

This gives R = does not fit in; P = is so small, just like before!

And now consider this one:

> 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; alternate=months)

Do we need a bigger bag of tricks?

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

> Can we engineer a system to produce a desired behaviour with no more errors than people would produce (with confidence level z)?
>
> This can allow some of the more challenging examples to be *ignored*!

But taken strictly, this can also lead to systems with very impressive performance that are nonetheless *idiot-savants*.

> prodigies at chess, face-recognition, *Jeopardy*, etc.

But there is another way of looking at this ...

Think of people's behaviour as a *natural phenomenon* to be explained.

Even a *single example* can tell us something important about what people are able to do (however statistically insignificant).

A thought experiment

Consider a question about materials:

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

Now suppose that you learn some facts about XYZZY.

1. It is a trademarked product of the Dow Chemical Company.
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.

Ask: At what point does the answer stop being just a guess?

The lesson

From a *technology* point of view, a reasonable question here is

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

But from a *science* point of view, we want to understand what it takes to produce the behaviour that people are able to exhibit.

So the question really needs to be more like

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

No tricks!

What would it take to build a system that *knows* a lot about its world and can apply that knowledge as needed, the way people can?

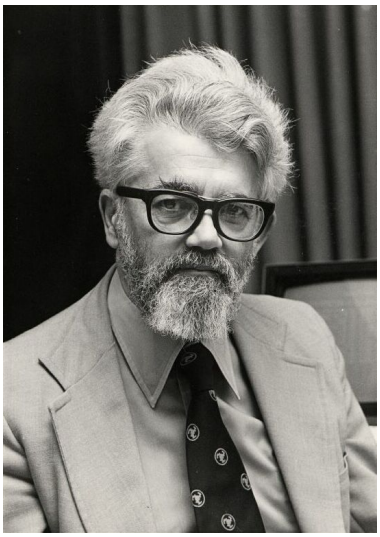
One possible answer:

1. some part of what needs to be known is represented symbolically;
   (call it the KB)
2. procedures operate on this KB, deriving new symbolic representations;
   (call it reasoning)
3. some of the derived conclusions concern what actions should be taken next
   (including answers to pressing questions)

A radical proposal: instead of tricks or shortcuts, focus attention on
- what knowledge is needed;
- how to represent it symbolically;
- how to use the representations.

Where did this radical idea come from?

This idea was first proposed in 1958 by John McCarthy in an extraordinary and unprecedented paper entitled “Programs with Common Sense.”



John McCarthy (1927 – 2011)

Two hurdles for knowledge-based AI

1. Much of our knowledge of the world comes not from personal experience, but from the 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 have 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, performing symbolic reasoning over these logical constructs appears to be much too demanding computationally.

AI longa, vita brevis

These hurdles are as serious and as challenging to the science of AI as an accelerating universe is to astrophysics.

Given that it's been 55 years since the McCarthy paper, we might well wonder if AI researchers will *ever* be able to overcome them.

Many of us have been compelled to return to more traditional methods

e.g., more biologically-based,
more like statistical mechanics

to focus on behaviours that are seemingly less knowledge-intensive

e.g., recognizing hand-written digits,
following faces in a crowd,
walking over rough terrain

. . . and with terrific results!

Our best behaviour (again)

But our best behaviour *does* include knowledge-intensive activities

such as { participating in a natural conversation
responding to Winograd schema questions

My hope: enough of us stay focused on this sort of intelligent behaviour to allow progress to continue here as well.

This will require hard work!

> We cannot expect solutions to emerge spontaneously out of a few general principles, without any real effort on our parts.

Hard work, yes, but an exhilarating adventure !

And what about those two hurdles? ...

Advice to the KR scientist

1. We 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.
   >
   > 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, *etc.*

2. It is not enough to build knowledge bases without paying very close attention to the demands arising from their full 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 over constructs that logically demand more.

Advice to the AI scientist

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

serial silver bulletism:

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

It will all be solved by ~~automated theorem-proving!~~ ~~expert systems!~~ ~~behaviour-based-robotics!~~ ~~learning from big data!~~ ⟨*the next big thing*⟩

We recognize more fully what our own research does *not* address, and 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 / allow progress to proceed in a steadier fashion

What are the prospects?

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

A: *"The best way to predict the future is to invent it."*

— Alan Kay

THE END

These slides and a written version of this talk can be found at

`http://www.cs.toronto.edu/~hector/Papers/`

References

- 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 ACL Workshop*, Prague, 2007.
- B. Christian, *The Most Human Human*, Doubleday, 2011.
- A. Collins, E. Warnock, N. Aiello, M. Miller, Reasoning from incomplete knowledge, in *Representation and understanding*, Academic Press, 1975.
- I. Dagan, O. Glickman, B. Magnini, The PASCAL recognising textual entailment challenge, in *Machine Learning Challenges*, Springer Verlag, LNAI 3944, 2006.
- H. J. Levesque, Logic and the complexity of reasoning, *The Journal of Philosophical Logic*, **17**, 1988.
- —, Is it enough to get the behaviour right?, *Proc. of IJCAI-09*, Pasadena, CA, 2009.
- —, E. Davis, L. Morgenstern, The Winograd Schema challenge, *Proc. of KR-2012*, Rome, 2012.
- J. McCarthy, The advice taker, in *Semantic Information Processing*, MIT Press, 1968.
- R. Reiter, On closed world databases, in *Logic and Databases*, Plenum Press, 1978.
- A. Turing, Computing machinery and intelligence, *Mind* **59**, 433–460, 1950.
- L. von Ahn, M. Blum, N. Hopper, J. Langford, CAPTCHA: Using Hard AI Problems for Security, in *Advances in Cryptology, Eurocrypt 2003*, 294–311.
- J. Weizenbaum, ELIZA, *CACM* **9**, 36–45, 1966.
- T. Winograd, *Understanding Natural Language*. Academic Press, New York, 1972.

He who clings to his work will create nothing that endures.
If you want to accord with the Tao, just do your job, then let go.

— Lao-Tzu