WORK IN PROGRESS

Notes expanding
Comments on Royal Society Survey on Machine Learning

Aaron Sloman
http://www.cs.bham.ac.uk/~axs

Response to Royal Society Call for evidence on Machine Learning
https://royalsociety.org/topics-policy/projects/machine-learning/call-for-evidence/

From: Aaron Sloman
School of Computer Science
The University of Birmingham
Edgbaston, B15 2TT, UK
http://www.cs.bham.ac.uk/~axs

Background

I was asked at late stage to submit comments. I sent in some messy, hastily written comments on 7th Jan 2016. A short time later I posted my comments here, and began to improve, correct, clarify and expand them. This version is
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/royalsoc-deep-learning.html
Also (pdf).

NOTE: The original version of this document was written in great haste and submitted on 7th Jan 2016, with an apology for lateness. Since then I have made a number of changes, attempting to clarify and document some of the claims, and adding some references. There is still a great deal more that could be done, to clarify, extend, organise, and provide evidence for the detailed claims about what still needs to be explained, and limitations of currently popular machine learning approaches to AI.

Last updated: 11 Dec 2018

CONTENTS

- **Note2 added 7 Jan 2016**
- **Kinds of stuff**
- **Playing with blocks and drawing them**
-

Clarifying requirements for visual perception and learning

- **A question about processing power needed**

Need for balance

The call for evidence is related to this web site:
https://royalsociety.org/topics-policy/projects/machine-learning/ which unfortunately echoes much of the "hype" surrounding recent developments in AI, some of it justified, but not without careful analysis of its scope, and what may be missing.

I would have expected the Royal Society to produce a more balanced presentation, e.g. including examples of limitations of what has been achieved so far, and discussion of hard problems remaining to be solved. Perhaps those were addressed in documents I have not found. For science, the main business of the Royal Society, identifying what we don't yet understand is at least as important as reporting the progress that has been made.

There have certainly been extremely useful advances in machine learning techniques, though it is important to be aware that demonstrations do not always present analyses of limitations and failures of the systems being demonstrated. Moreover, as Geoffrey Hinton pointed out in his online video presentation to the Royal Society and also in the Forum programme on BBC world service recently http://www.bbc.co.uk/programmes/p02kmqt1, the recent advances in machine learning have come mainly from old ideas that can take advantage of recent of advances in speed and memory capacity, along with dramatic reductions in cost, size and weight of computers available for research and applications over the last 30 years.

I am not opposed to supporting and promoting research in machine learning. But there are many aspects of intelligence in humans and other animals that the recent advances do not address, and they are at least as important (for science) as the learning techniques. Unfortunately, they seem to go unnoticed.

A little history and some background comments on "large" projects

It may be worth dwelling a little on some history. Provoked by the Japanese "5th Generation" project, The UK Government's Alvey Project, begun around 1983, attempted to address the problems of applying results of AI research. Here's a small sample of publications about the project.
https://en.wikipedia.org/wiki/Fifth_generation_computer
https://en.wikipedia.org/wiki/Alvey
http://www.chilton-computing.org.uk/inf/alvey/p001.htm
http://www.amazon.co.uk/gp/search?index=books&linkCode=qs&keywords=9780115152818
Evaluation of the Alvey Programme for Advanced Information Technology:
A Report by Science Policy Research Unit (1991)

The Japanese project assumed that all the main problems could be solved by using logic programming implemented on highly parallel hardware. That assumption was challenged by commentators in the UK and USA, and the Alvey project was broadened to allow more varied hardware and languages to be used, and to address important omissions -- e.g. the original proposal did not include work on visual perception. Most of the aims were not met, though a lot of learning happened.

In my view the single most important result of the Alvey project was that it introduced to UK computing/software companies the idea of collaboration between different companies and between academe and industry, in an area in which such collaboration had not previously happened. The main hoped for results were not achieved (as several people, including Edward Feigenbaum, in the USA, predicted in advance), though a lot was learnt in the process. Moreover, the Alvey project gave birth to a special Joint Council Initiative (JCI) in Cognitive Science and Human-Computer Interaction, more narrowly focused on AI/HCI and Cognitive science:
http://www0.cs.ucl.ac.uk/staff/R.Young/jci-evaluation/intro.html

The JCI initiative, like many such initiatives, provoked research of very mixed quality, but was just beginning to lead to new worthwhile collaborations when it was deemed (mistakenly in my opinion) to be no longer in need of any special support, and the initiative fizzled out.

A decade later, the EC launched a new *Cognitive systems* initiative, announced in 2003, funded from 2004, combining robotics & AI, cognitive science, linguistics, and related fields. It was unusual among EC Framework initiatives insofar as it emphasised both multi-disciplinary collaboration and also *research* rather than *applications*, in recognition of the huge gaps in our scientific knowledge and understanding. The tail end of this is still running in the form of the "European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics"
http://www.eucognition.org/

More recently the EU Human Brain project was launched:
https://en.wikipedia.org/wiki/Human_Brain_Project

I was invited to join, but declined because I thought its assumptions were mistaken and, like most research, ignored some of the unexplained competences of humans and other intelligent animals, thereby addressing only a subset of the problems, and especially a subset of brain functions, without acknowledging the fact.

I do not claim that the above list is complete. I merely wish to warn against enthusiastic national or international initiatives that ignore the problems, failures, partial successes, and unfinished business of previous initiatives addressing the same or closely related problems.

I think all of these projects failed to ensure that adequate attention was given to the problem of identifying what needs to be explained (or replicated) in the long term, so that there is an “external” benchmark with which achievements can be compared. Of course, any list of long term objectives/requirements should be subject to revision. If scientists cannot agree, after analysis and discussion, on what should be in the list, the union of objectives could be provisionally adopted, subject to revision later in the light of new findings.

I have noticed two facts that repeatedly interfere with the success of such major new initiatives.

(a) As indicated above, the projects usually do not include sufficiently broad and deep *"requirements analyses".* They make over-simple assumptions about what needs to be explained, for example, and end up focusing on a small subset of problems, ignoring other equally important, or more important, problems. In many cases, the problems that are ignored need to be solved in order to address the problems that are attended to.

My personal impression is that the most important advances do not come from grand, ambitious national or international government-led projects but from individuals or groups beavering away at very hard problems, eventually producing results that could not have been anticipated or planned. The main practical benefits of some of the deepest new discoveries are not realised until decades later.

(There are obvious counter examples to the emphasis on individuals or small groups: namely projects that depend essentially on large expensive new instruments, machines, expeditions, etc., such as high energy physics, astronomy, space research, oceanographic surveys, etc. But these are usually cases where there is already far more shared factual and theoretical knowledge at the time proposals are formulated, than in typical multi-site AI/Robotics projects.)

(b) In addition, many projects are limited by the difficulty of recruiting researchers (doctoral and post-doctoral) whose education is broad enough and deep enough for the task. This is particularly true for projects in AI or natural cognition. Even researchers with degrees in AI or cognitive science often turn out to have learnt only the recently most fashionable techniques and theories, sometimes having been brainwashed into thinking that there was nothing of value in earlier work. (Sometimes this is also true of project leaders.)

One consequence of (b) is that researchers are selected not because they are really suitable for the jobs, but because funding contracts require projects to start by a certain date, so project leaders are forced to accept the *best* applicants available by that date, instead of continuing to search for the *right* researchers. That is also a problem with large multi-team projects that need to be synchronised.

Moreover, the general background education of many researchers is now grossly inadequate, e.g. knowing little or nothing about Euclidean geometry and topology, many have mathematical backgrounds restricted to numerical mathematics rather than mathematics of structures. And usually they know nothing of relevant areas of linguistics or philosophy, including philosophy of mind, philosophy of language, philosophy of science, and philosophy of mathematics. So they invent or read up and use bad philosophy, unaware of any alternative and in some cases adopt linguistic goals that do not match what is already known about human language. Another common example is work aimed at modelling affective states, such as emotions, based on definitions or pronouncements by some authority or research group, taking no account of published criticisms or major extensions of those views -- leading to research that ignores published criticisms of its assumptions. (I am not saying all research in the area makes all these mistakes, but most of the examples I have looked at do.)

Similarly, researchers' knowledge of biology and achievements of evolution, and their knowledge of new forms of computation being explored outside computer science (e.g. chemical computation) is often inadequate. Many, especially those with qualifications from psychology, often have a narrow naive-Popperian view of the nature of science and the criteria for evaluation of scientific theories – with an over-emphasis on statistical validity that cannot accommodate deep individual variations in

development or competences.

They are usually unaware that Popper revised his opinions on requirements for science, partly as a result of coming to appreciate the great achievement of Darwin and Wallace even though he had previously pointed out that the theory of natural selection was not falsifiable. (I have extended Popper's criteria to include research on *what is possible* and on *explanations of* what is possible. Claims about what is possible are not empirically falsifiable, but have often been at the heart of major scientific advances.)

Deep research into understanding, modelling, and replicating aspects of intelligence of humans and other animals requires a new cohort of graduates with a much deeper and broader education in science, mathematics and philosophy than our schools and universities are now able to provide. Moreover, there is far too much pressure on new young lecturers to get grants and citations, so that they don't have an appropriate period, e.g. 5 to 10 years, in which they extend their education by reading broadly, attending seminars in other disciplines, etc., while doing all the teaching required of them and contributing to departmental administration. These are general problems that are especially important for researchers in highly multidisciplinary fields, e.g. research in cognition, artificial intelligence, robotics, neuroscience, animal behaviour, and related fields. I turn now to the Royal Society initiative.

The Royal Society Call for evidence

The web site asks:

How important do you think machine learning will be for you and/or society in the next 10 years?

Machine learning is one of my main research interests, but not for practical engineering reasons: rather because as a philosopher and a scientist I am trying to understand biological forms of learning and I think any explanatory theory will need to be tested by being implemented in working machines. So work on machine learning is an end in itself for some people and just one part of a large collection of problems, for others.

My own work, straddling philosophy, AI, cognitive science and biology, is mainly on assembling *requirements* for deep explanatory theories – i.e. identifying phenomena that are currently hard to explain or replicate.

These phenomena (e.g. aspects of mathematical discovery by humans, especially in topology and geometry, that are related to everyday competences of humans and other animals, and closely related aspects of visual perception) generate requirements that need to be met by machine learning systems if they are to match or model the forms of learning that occur in humans and other intelligent animals.

Some of the links between mathematical discoveries and perception of affordances are discussed in a paper under construction here:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/impossible.html

In particular, in 1781 Immanuel Kant pointed out some features of mathematical discovery that need to be explained, which he summarised in his claim that mathematical knowledge is not empirical, not analytic (derivable from definitions and logic), and not contingent (i.e. the discoveries are concerned with what is necessarily the case or necessarily not the case – conclusions that *cannot* be based on statistical evidence).

Some details of his claims need to be modified in the light of things learnt since his time, but I think he was basically right. However the methods of machine learning developed so far are incapable of producing the kinds of knowledge he discussed, including knowledge of geometry and topology apparently acquired and used by pre-verbal toddlers and other animals, which appear to be precursors to the achievements of Euclid.

Specifying requirements is often one of the hardest parts of an engineering project and many IT projects (both national and commercial) have been disastrous because of grossly inadequate requirements analysis.

Likewise, a failure to specify detailed and accurate requirements for human-like or animal-like learning in machines can lead to failed, or seriously inadequate projects, or more insidiously, projects that appear to be successful because the public are unwittingly persuaded to accept shallow and inadequate criteria for success – e.g. performance in narrowly constrained domains.

Unfortunately, despite the tremendous theoretical advances and hugely varied and useful applications of machine learning, there remain extremely important aspects of human and animal learning that either have been ignored completely or have not been characterised adequately. As a result the fact that the phenomena have not been explained largely go unnoticed, and that can have harmful effects for science.

In part that is because many of the problems are very subtle and difficult to characterise.

That can be illustrated by the fact that humans have been using spoken and written languages for centuries but it was not until the last century that some of the main requirements for biological or artificial brains to use and understand language were understood and characterised mathematically, though even now there are serious gaps and inadequacies in theories of language, which is partly why the linguistic abilities of machines are still so restricted.

One source of problems is the wide-spread assumption that children, in effect, *learn* their native language(s) by doing data-mining in the examples of expert language use that they encounter. This *must* be wrong, because originally there were no language users from whom to learn.

More direct evidence comes from the deaf children in Nicaragua who could not have learnt their sign-language from data because they (mostly) created the language themselves. This short video documents some features of the episode:
https://www.youtube.com/watch?v=pjtioIFuNf8

See also
Ann Senghas, (2005), Language Emergence: Clues from a New Bedouin Sign Language,
in *Current Biology*, 15, 12, pp. R463--R465, Elsevier

So it is possible at least for humans to develop competence in a language that they have not learnt by data-mining. As far as I know, there is nothing in current AI that models the language development processes that actually occur normally in children, though some AI systems model processes observed in artificially constrained laboratory tests. Moreover, insofar as the statistics-based language learning mechanisms are constrained by the externally provided data and human language-creators are not, the current AI methods *cannot* provide a basis for replicating human intelligence in future machines. Exactly what kind of creativity is missing and how to provide it is a topic for non-trivial long term research. I have been collecting examples for

over four decades, many of them examples of mathematical discoveries. (The examples are assembled in a large and messy, steadily growing, publicly accessible, web site, that I shall not try to summarise here.)

The point extends far beyond language learning. Consider all the products of human intelligence that most people encounter by learning from external sources, e.g. perceiving, reading, being told. These include paintings, stories, musical compositions, architectural designs, and, since the distant past, a stream of discoveries and creations leading to new tools, new techniques, new designs, new applications, theories, games, new notations (for music, architectures, mechanical designs, dance moves, etc.) new criticisms of previously accepted products, techniques, ideas, etc. and unfortunately also new ways of doing evil.

It could be argued that once such novelties have somehow been produced, everyone who needs to learn about them uses data-mining, so we can at least aim to produce intelligent machines that are always intelligent students, not intelligent innovators. But the Nicaraguan case refutes that. Moreover many educators have criticised that theory of education, stressing the importance of learning by producing creative solutions to carefully graded challenges (sometimes referred to as "scaffolding"). E.g. that is how I was taught mathematics. As far as I can tell, the data-mining mechanisms do not (yet) incorporate the ability to learn through a steady stream of creative responses to a steady stream of challenges, most clearly evident in good teaching of mathematics, philosophy, engineering and some areas of science.

Moreover, insofar as they discover only statistical **regularities**, not mathematical (e.g. geometrical, topological) **necessities**, they cannot be extended to achieve our research goals.

I don't think anyone has good (i.e. implementable) theories about the mechanisms that support the kinds of creative learning process that enable humans and other intelligent animals to come up with novel solutions to novel or old problems, or even to adopt new types of goals that are unrelated to old goals or needs. It must have something to do with the products of natural selection: discovering exactly how natural selection starts from a lifeless planet and eventually produces those creative learning procedures is one of the aims of the Meta-Morphogenesis project.

My background and qualifications as commentator

As I am a critic of the current state of AI, I should perhaps make it clear that I am not one of those who *hope* or *predict* that AI will fail. I am trying to understand requirements for it to succeed.

After I started learning about AI (in 1969 mainly from Max Clowes) I began to get deeply involved in trying to use AI to model human mental capabilities, and to teach new ways of doing philosophy and cognitive science, first at Sussex University, where I was one of the founders of COGS, the school of Cognitive and Computing Sciences (Margaret Boden was the first Dean). I was also co-developer and for a while local manager of development of Poplog, an AI toolkit developed at Sussex University and used for teaching, research and product development (https://en.wikipedia.org/wiki/Poplog). I continued doing philosophically motivated AI research after I moved to Birmingham University in 1991.

So I am not merely a philosophical *commentator* on AI: I have (with colleagues and students) designed and built working systems to test out ideas, and was the main developer of a Poplog-based toolkit (SimAgent) that was used by students and researchers at Birmingham and elsewhere to explore alternative information-processing architectures for "complete" agents of

varying types. I think it has important features still not found in other AI architectural tools, e.g. support for flow of symbolic or other information between subsystems of different sorts while they are processing information (a process sometimes referred to as 'barge in'), not based on adjustment of numerical parameters, and support for various kinds of meta-cognition.
E.g. used in these two papers by Catriona M. Kennedy, who helped to specify some of the architectural requirements for SimAgent (building on earlier work by Luc Beaudoin and Ian Wright):

Distributed Reflective Architectures for Adjustable Autonomy,
in *Proc. IJCAI 1999 Workshop on Adjustable Autonomy,* July, 1999, Stockholm, Sweden,
http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#56

(With A.Sloman):
Autonomous recovery from hostile code insertion using distributed reflection,
Journal of Cognitive Systems Research 4, 2, 2003. pp. 89--117,
http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200301

With that background of long term goals and practical expertise, I can say with confidence that some extremely important aspects of human and animal intelligence have not been noticed by the majority of AI researchers (and many other researchers, e.g. in psychology and neuroscience). Full replication in AI systems does not seem to be close.

In most of my research, instead of focusing only on particular aspects of intelligence (vision, speech, learning, planning, manipulation, etc.) and trying to model or replicate them, I have been trying to survey a wide variety of mutually supportive aspects of intelligence that need to be characterised, explained or modelled, including forms of motivation and their effects (e.g. emotions, moods, states of grief, etc.)

Among the particularly interesting unexplained phenomena are the kinds of discoveries that led to the production of Euclid's *Elements* about 2,500 years ago especially discoveries in geometry and topology. These seem to be beyond the scope of current AI systems, not just deep learning systems. Current artificial learning systems can't even make some of the discoveries made by pre-verbal human toddlers, or squirrels defeating squirrel-proof bird-feeders.

For several years I have been collecting examples and analyses of "toddler theorems". E.g. here's a video of a pre-verbal toddler holding a pencil, picking up a sheet of card with two holes, and going through carefully controlled movements: pushing the pencil into one of the holes, pulling it out, rotating the sheet of card to bring the other side of the hole into view, pushing the pencil into the hole in the reverse direction, pulling it out, pushing it again through the hole from the original direction, pulling it out, then moving on to do something else.
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/toddler-theorems.html#pencil
She seems to have very definite intentions and very expert abilities to bring them about. But at her age (about 17.5 months) she could not say what she was doing, how she did it, why she decided to do it or how she knew it was possible.

That implies that *long* before she could express in words things like "I am going to push this pencil through that hole", "The hole can also be entered from the other side". "I am going to move the pencil through space, rotating it, until it can be pushed through the hole in the opposite direction" she clearly had intentions with contents related to my verbal descriptions here, and she was able to derive the appropriate movements of her hands and head, including controlling

eye-gaze when the pencil was being moved towards the hole.

(She did this with no prompting, no social interaction, no imitation of anyone else: she apparently just happened to see the opportunity, and I just happened to have a cheap camera available, and saw my opportunity.)

Her expert, untrained, ability presupposes one or more rich internal languages (information-bearing systems) capable of representing both perceived and future *possible structured configurations* and *possible configuration-changing processes.* The precise forms of the languages are unknown. But they must include abilities to represent complex structures and processes, of varying complexity and form.

N.B. Similar arguments apply to many complex, creative, problem-solving achievements of other animals that never develop human-like languages for communicating with one another: they must also have powerful internal languages with structural variability and compositional semantics, for representing percepts, goals, intentions, and possibly also questions and hypothetical answers to guide investigations.

I am not aware of any currently used formalisms for representing visual contents in robots that are capable of expressing the percepts, intentions, plans, etc. apparently involved in motivating and controlling her actions: and the knowledge of 3-D topology that she seems to have deployed.

What language could she have used? Where did it come from? What role does it play in the child's learning to talk, later on? How is it implemented? What kind of neuroscientific research could suggest answers?

Similar questions are triggered by observing intelligent behaviours of other animals, including squirrels, crows, weaver birds, elephants, orangutans, dolphins, octopuses, and many more.

Use of modal concepts

At a later stage, the child's information processing resources must make possible expression of modal judgements (e.g. X is *possible*, X is *impossible*, X is *necessarily* the case, etc.) It is not clear that such modal concepts can be learnt by any sort of data-mining: As Kant noticed, something genetically determined needs to be in the architecture that supports their discovery and use. I do not believe that currently popular theories of the semantics of such modal operators are correct (e.g. "possible world semantics".)

Although AI researchers have designed various sorts of "modal" representation and reasoning systems, they cannot yet claim to have modelled natural modal representations. Moreover, as far as I know, no current AI learning systems even attempt to learn the kinds of "modal" representation and reasoning capabilities the toddler with the pencil seemed to be implicitly(?) using. Moreover, as far as I know, no neuroscientist has tried to explain how brains are able to represent modalities (e.g. possible, impossible, contingent, necessary).

These are not examples of probabilistic information. They cannot be derived from statistics, except that observation of a single example of a type (a minimal statistic) does demonstrate the possibility of instances of that type, and may refute an impossibility claim. However, sometimes the exact characterisation of the observed example (like my characterisation of the toddler with the pencil) is debatable.

I discuss some of these problems and related issues concerning evolution of language in this slide presentation:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk111
What are the functions of vision?
How did human language evolve?
(Languages are needed for internal information processing, including visual processing)

The Meta-Morphogenesis project

These problems have been the focus of my research for many years, most recently in the framework of the (Turing-inspired) Meta-Morphogenesis project, which aims to identify important transitions in biological information processing since the very earliest organisms and pre-biota. Some of those previously unnoticed transitions may give us clues as to what we are currently failing to identify in brain functions and (therefore) in brain mechanisms: e.g. capabilities that will be needed in more complete future AI systems.

The project was triggered by wondering what Turing might have done if he had died three or more decades after his 1952 morphogenesis paper, rather than two years later. Details available at the Meta-Morphogenesis (M-M) project web site, below.

This has given a new shape to work I've been doing for half a century, starting before I encountered AI, especially work in philosophy of mathematics.

Since presenting a critique of the logicist manifesto of McCarthy and Hayes at IJCAI 1971, my aim has not been to prove that AI must fail (e.g. like Dreyfus) but to identify gaps that need to be filled so that it can succeed in its long term (scientific, explanatory) aims.
http://www.cs.bham.ac.uk/research/cogaff/62-80.html#1971-02
A. Sloman, Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence, *Proc 2nd IJCAI,* London, 1971, pp. 209--226,
Also Chapter 7 of The Computer Revolution in Philosophy, 1978:
http://www.cs.bham.ac.uk/research/projects/cogaff/crp/

There's still a long way to go, mainly because of gaps that go unnoticed by most AI researchers. (Like the people who once thought Newton had all the answers and the rest was just a matter of filling in details).

This is not AI as *engineering* but AI as *science* (and philosophy: the two overlap when done well), i.e. attempting to construct theories that explain or model natural forms of intelligence that at present are not understood -- like the intelligence that led up to Euclid, the intelligence of squirrels defeating bird-feeders, the intelligence of weaver birds making nests using several thousand knitted/knotted leaves, the intelligence of composers who produce great music, the intelligence of listeners who respond to such music the first time they hear it, without needing to have it explained, even centuries later, and the intelligence of human toddlers exploring 3-D topology.

These scientific AI goals seem to have recently been sidelined, though understanding and modelling natural intelligence was an important goal for founders of AI, including Turing (in a letter to Ashby), McCarthy, Minsky, Simon and others.
Added 11 Nov 2018
Turing's letter is here:

https://www.bl.uk/collection-items/letter-from-alan-turing-to-w-ross-ashby
Added 11 Nov 2018
For a discussion of Turing's distinction (in 1938) between mathematical intuition and mathematical ingenuity, see
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/turing-intuition.html (also pdf)

Moreover, I think the more ambitious engineering goals will not be achieved while so many gaps in scientific understanding remain. Identifying and filling the gaps may take longer than the rest of this century -- even if applied AI continues to make spectacular progress in many constrained sub-fields.

More examples

Things that have proved hard to analyse from the designer stance include the kinds of perception, learning, and reasoning processes that might have led to the production of Euclid's *Elements* -- long before the discovery of modern logic-based, formal, mathematics. (Arguably Euclid's book is the single most important publication ever produced on this planet. Its results are still used every day by scientists and engineers all over the planet e.g.
http://www.cut-the-knot.org/pythagoras/Proof1.shtml.)

Modern AI theorem provers that start with axioms and rules expressed in a logical notation, and attempt to find proofs derived from the axioms in accordance with the rules, do not model processes of the sort in question (in part that was Frege's criticism of Hilbert's attempt to logicise Euclidean geometry, though I would make a similar criticism of Frege's great work attempting to logicise arithmetic -- as a result of which he produced some of the powerful constructs now commonplace in AI programming languages, and some others, e.g. higher order functions).

I think Immanuel Kant, in his discussion of the nature of mathematical knowledge in *The Critique of Pure Reason* (1781) started moving in the right direction, and would probably have used AI with glee if it had been available then.

Anyhow, for several decades, in addition to working on AI projects (including building tools and formalisms used by students and colleagues) I have been collecting examples of capabilities that don't seem to fit current AI techniques (e.g. trying, in the 1980s, to specify the functions of vision ignored by Marr, Gibson and others, and the architectural requirements for a wide variety of emotional and motivational phenomena, including some ignored by most researchers, like long term grief). I have recently been trying to assemble and organise these long term requirements and relate them to products of biological evolution, in the messy and growing Turing-inspired Meta-Morphogenesis (M-M) project web site, mentioned above.

Here are a few example challenges.

A key feature of biological intelligence (almost, but not quite, recognized by James Gibson, though he made some moves in the right direction) is the ability to grasp sets of possibilities that have nothing to do with probabilities, but do have absolute limitations, i.e. things that are impossible.

E.g. a child playing with similar blocks on a table top could discover ways of arranging groups of blocks, in regular arrays, as illustrated in these three examples:

```
(a) oooooooooo

(b) ooooo
    ooooo

(c) ooo
    ooo
    ooo
```

Every group can be arranged in a line, like the first example. Sometimes one group can be arranged in more than one way, e.g. as a line or a rectangular or square array, or in several different ways, e.g. 64 blocks.

But sometimes if you add or remove a block the possibilities change dramatically. E.g. 68 blocks can be arranged in several different configurations, but if you remove one block only one possible configuration remains. Why?

(Gibson, apparently did not notice "negative" affordances with mathematical explanations.)

How could a young robot playing with such blocks come to realise that some of the rearrangements are impossible? (I don't know how many humans can, unaided, but some can. I suspect more would be able to if primary schools were run differently.) Different sorts of impossibilities involving blocks are mentioned below.

I hope readers of this document will have recognized the connection between my examples and the fundamental theorem of arithmetic: every natural number (positive, non-zero, integer) has exactly one decomposition (ignoring variations in order) into a product of primes greater than 1. Should we expect work on machine learning to lead to a machine capable of discovering and proving this theorem without first having to be programmed with general knowledge and techniques of logic and set theory, which the original discoverers did not have?

A quite different example: Three or more straight lines drawn on a plane can enclose a finite region of the plane. Why can't that be done with two lines? (One of the examples discussed by Kant.)

Is there a similar limitation on plane surfaces arranged in a 3-D space to enclose a finite volume of the space? How could a machine discover, and understand the limitation?

(Does anyone have a geometric theorem prover that can find the answer? Would it have to use geometry arithmetised, following Descartes? You probably have another way of thinking about it. Can your abilities be programmed into a robot now? Does any neuroscientist know how your brain supports such abilities?)

Similar discoveries about impossible spatial structures might be useful for future robot architects -- saving a lot of time trying to build impossible buildings mistakenly thought to be possible and useful.

Strings

There are many ways flexible strings can be moved around. In particular, a string can be threaded through one or more holes in a piece of leather (as in a shoe). Suppose it goes through only two holes: how many different ways are there of removing the string from the holes? How can you be sure that you have counted them all? [Assume two removal processes are the same if the ends of

the string go through the same holes in the same direction.]

You can remove the string by pulling one end, or by pulling the other end. Why can't you remove it even faster by pulling both ends? What needs to be added to current robots to enable them to (a) discover such impossibilities, (b) understand **why** they are not possible?

If you pull both ends at the same time, there is a configuration that can be achieved faster: what configuration? The ability to answer that might be based on searching through a mass of data concerning previous pulling episodes. But that isn't required. What sort of ability would enable a robot to answer the question without resorting to experiments with strings and holes, and without searching through stored records of previous such experiments? How do you answer the question?

Shirts/Sweaters, etc.

If you want to put a tight fitting shirt on a child, or a doll, why is it a mistake to start by pulling a sleeve up one of the arms?

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/shirt.html

Impossible linking-unlinking

What would enable a young robot to have the intelligence to be amazed at a stage performer who seems to be able to make two disconnected solid rings become linked together?

Young humans (after what age?) are amazed: they don't need lectures in topology to understand that what they appear to be seeing cannot happen. It's not just unfamiliarity. I can do many totally unfamiliar things that will not be seen as impossible, for example, holding an egg in one hand, brushing it with a toothbrush held in the other, while I repeatedly recite Pythagoras' theorem. You can probably easily come up with equally unfamiliar, but possible, scenarios. How?

What sort of robot seeing apparently solid rings apparently being linked and unlinked would be as amazed as the human audience, and for the same reasons?

This sort of amazement is quite different from amazement on learning about something previously thought to be impossible, simply because it has never been encountered, or because of empirical evidence regarding limits of materials, or limits of human abilities. Piaget's last two books included examples of children of various ages answering questions about possibilities and necessary connections. He collected very interesting examples, and understood the differences between knowledge based on empirical evidence and knowledge based on logical, geometrical, or arithmetical reasoning, but had not learnt about computational models and was not able to propose designs for explanatory models. I don't know of any psychologist, neuroscientist, or AI researcher who can. Kant took some steps in a promising direction. Perhaps the work on "Representational Redescription" by Annette Karmiloff-Smith will turn out to be relevant, discussed here http://www.cs.bham.ac.uk/research/projects/cogaff/misc/beyond-modularity.html

More geometry and topology

Why must the three internal angles of a triangle sum to half a rotation?

What would have to go into a future AI system to enable it to discover or understand the proofs discussed here:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-sum.html

It is well known that there is no construction that will trisect an arbitrary angle in Euclidean geometry.

But Archimedes was aware of a fairly *simple* extension to Euclid that makes it possible to trisect any angle, discussed here:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/trisect.html

(It can also be done using origami geometry.)

What sort of AI system could discover that sort of extension to Euclid, and discover that it could be used to trisect angles? Could it be done by data-mining in a space of possible diagrams with changing parts? Or perhaps data-mining in a space of experiments with simple 2-D and 3-D manipulable objects, in a relatively unfamiliar domain, such as polyflaps?
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/polyflaps

Torus Theorems

You and I can think about closed, non-self-crossing curves (i.e. "simple curves") on the surface of a plane, a sphere or a torus (doughnut). We can also discover that there are classes of simple curves that are equivalent in that each curve in a class can be **continuously** deformed into any other in that class, i.e. without any cutting to produce a gap that did not previously exist and without any joining, to remove a gap. E.g. any simple closed curve in a plane can be continuously deformed into any other simple closed curve in the plane. How do you know? The same is true of simple closed curves on a sphere. How do you know? On a torus it is possible for two simple closed curves to exist that can be continuously deformed into each other, two curves going round the 'sidewall' of a tyre, or one of those and a curve on the 'rim' of the tyre, and many more.

If you think for a while you'll discover that there are some pairs of simple closed curves on a torus that cannot be continuously deformed into each other, (e.g. e.g. a circle on the sidewall of a tyre surrounding the hole, and a curve going round the "tube", i.e. going through the hole and coming round the outer rim to re-join itself). How do you know that neither can be continuously deformed into the other? How could a future robot know that? No matter how many attempts it has seen end in failure that does not prove it is impossible, since it will not have seen all possible pairs of curves and all possible ways of attempting to transform one to the other.

How do you know that a curve round the sidewall cannot be continuously deformed (in the surface of the torus) into a curve round the tube, going through the hole?

How can such discoveries be made for the first time?

If C1 can be continuously deformed into C2, then C2 can be continuously deformed into C1. Why? How do you know? How could a robot know, without being told?

If two curves C1 and C2 can't be continuously deformed into each other on the surface, they are in distinct equivalence classes, otherwise the same equivalence class. How many distinct classes of simple continuous, closed, non-self-crossing curves on a torus are there? How do you convince yourself?
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/torus.html

Could you have two equivalence classes of curves EC1 and EC2, such that EC1 contains EC2, but not vice versa? How will your robot know, without being told?

How do all these mathematical capabilities grow out of products of natural selection: what were the biological requirements that were being met by our ancestors ancestors... that later made mathematicians possible?

(I suspect there were several stages, some shared with other species, followed by three layers of meta-cognition apparently unique to humans: but not all available at birth -- for good reasons their epigenetic development has to be delayed: why?)

I am not claiming that these mathematical discovery mechanisms are *infallible* : the work of Imre Lakatos (in *Proofs and Refutations* (1976)) on the ups and downs of Euler's theorem about polyhedra
(E=V+F-2 where E:number of edges, V:number of vertices, F:number of faces)

demonstrates the fallibility of human mathematical abilities -- and some of the debugging and recovery processes that are possible in intelligent systems to compensate for the fallibility.

Conjecture:
Perhaps a future intelligent robot could replicate Euclid's, and Archimedes', discoveries by doing data-mining not in a database of human-supplied facts, but in a database of percepts of structures and processes generated by playing with various sorts of construction kits, e.g. wooden cubes, meccano, tinkertoy, lego, plasticine, sand, mud, foldable paper, string, scissors, etc.

What forms of representation would the percepts use? What kinds of data-mining/deep learning algorithms could operate on the recorded percepts? In what ways might visual and motor records have to be transformed in order to be usable in such learning systems? What sorts of mechanisms could discover not merely that something *never occurs* , but that it is *impossible* ? What could discover not merely that some operation *always* produces a certain result but that it *necessarily* does so. What sort of mechanism could discover that something that has never been observed, is nevertheless *possible* – e.g. a regular planar polyhedron with a billion and three sides?

Note1 added 7 Jan 2016
I suspect that the normal semantics of (alethic) modal concepts used by humans (e.g. "impossible", "possible", "necessarily", etc.) are derived from this sort of situation rather than consideration of possible complete worlds. "Possible world semantics" may provide a nice mathematical theory, but I don't think it has anything to do with ordinary human, or animal, thinking and reasoning about what is and is not possible, or necessary consequences of possible changes -- for example, some of the contents of perception of affordances.

Compare the discussion of "possibility transducers" in
A. Sloman, (1996), Actual Possibilities, in Principles of Knowledge Representation and Reasoning: Proc. 5th Int. Conf. (KR '96),
Eds. L.C. Aiello & S.C. Shapiro, Morgan Kaufmann Publishers, Boston, MA, pp. 627--638,
http://www.cs.bham.ac.uk/research/cogaff/96-99.html#15

Note2 added 7 Jan 2016
*In 2002 I was invited to a DARPA workshop in Virginia to discuss goals for the new Cognitive Systems initiative. I was the only person from outside the USA. At the end of the workshop we were invited to propose goals for the project. I suggested a "baby" robot of the sort hinted at above, developing an ontology and theories about its physical environment through playful exploration of the environment. However the DARPA director, Tony Tether, had previously expressed an interest in building "personal assistants" (e.g. inspired by Corporal Radar O'Reilly, in the M*A*S*H film and TV series). So my proposal was rejected in favour of personal assistants. I don't know what the funded research actually achieved.*

Kinds of stuff

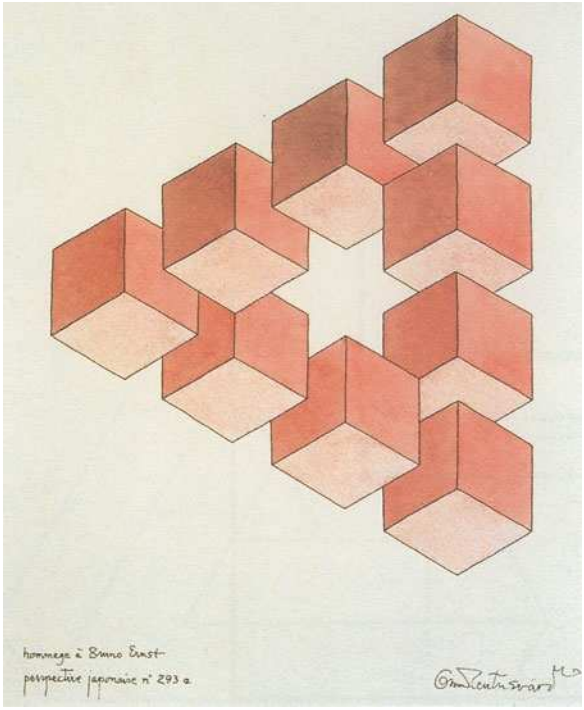
The above examples concern abstract shapes and the possibilities and impossibilities of various transformations of those shapes. Humans and many other animals also learn about different kinds of space-filling stuff, e.g some rigid, some with various kinds of non-rigidity (e.g. elastic, inelastic deformity). Many kinds of animal intelligence depend on abilities to perceive, understand and use kinds of deformity various kinds of stuff are capable of: e.g. the orangutans that use different sorts of compliance in their motions through trees.

How can robots be given similar capabilities? Will all their knowledge have to come from training, or could they have some deeper capabilities that enable them to make discoveries analogous to discoveries in Euclidean geometry but subject to various possible shape deformities of different kinds of matter.
[This point needs clarification, with examples.]

Playing with blocks and drawing them

Groups of similar cubes can be arranged in space to form various shapes. How can a robot think about, and draw pictures of, different spatial configurations of nine cubes in 3-D space, without ever making use of cartesian coordinates, or polar coordinates or other arithmetical representations of spatial structures and relationships, as used by current robots and machine vision systems? I suspect a child pushing a pencil through a hole uses brain mechanisms that more directly represent spatial structures and processes (using what were called "analogical representations" in Sloman 1971.

How could it discover that there are some configurations that can be drawn, but could not possibly exist -- e.g. the configuration discovered by Swedish artist Oscar Reutersvard in 1934



discussed in this file (still under construction/revision):
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/impossible.html#impossible
(Skip to section: “Pictures of possible and impossible object configurations”.)

Clarifying requirements for visual perception and learning

There are many other aspects of human/animal visual perception that I think AI systems are not yet close to replicating, partly because the requirements tend to go unnoticed by most researchers.

Here’s an example that needs a lot more discussion than could fit into this document. I have some videos taken with a (cheap) camera moving around a fairly rich and varied garden with occasional gusts of wind making petals, leaves, etc. move. What do our visual systems achieve when looking at those videos or moving round the garden looking at the bushes, shrubs, trees, flowers, etc. (without being familiar with the species there)?
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vision/plants

I’ve posed a task for AI: not to design a system that can do what we do, but to design a set of *requirements* for such a system! How could such a set of requirements be evaluated?

I think most vision research is tested against very limited sets of requirements, and often the wrong ones.

E.g. 3-D stereo vision systems (or visual slam systems) are often tested by their ability to generate different views of a scene, when perceived from different locations, or even the ability to produce fly-through videos. But my brain can’t do that, except for very simple views. Expert artists are much better, but that’s a specialised application of some powerful mechanisms shared with non-artists -- and birds, squirrels, hunting mammals, and others perhaps??

So what do normal human visual systems do during walks around a botanical garden full of previously unknown (to the viewer) plant forms?

I don't think anyone at present (and that includes me) can specify the requirements to be met by an AI vision system that can do what we do when looking at complex, varied, changing, scenes.

But I've been collecting many fragments of the competences, e.g. telling whether two flowers never seen before are likely to be members of the same, previously unknown, species; or whether an unfamiliar object seen from one viewpoint is also one of the objects visible from another viewpoint, where its 2-D projection is quite different.

[Unfamiliarity rules out use of previous training on that shape.]

What does a nest-building crow need to see in order to select a location for the next twig it brings to the unfinished nest?

What does it need to see in the part-built nest in order to control its search for the next twig? Or does it just fetch any available twig then see how it can be used.

Does anyone still remember Betty the hook-making crow from New Caledonia, in Oxford 2002? (Alex Kacelnik and Jackie Chappell, etc.)
http://news.bbc.co.uk/1/hi/sci/tech/2178920.stm
https://www.youtube.com/watch?v=UDg0AKfM8EY

Not all humans have the same perceptual, learning, and problem-solving capabilities.

Some young autistic-spectrum people can spontaneously draw complex pictures of a 3-D scene that most humans cannot, though they may improve with training. So, a general theory of human-like intelligence must enable us to be able to specify *generic* designs that accommodate various kinds of exceptional *more specific* designs, and perhaps explain why such sophisticated capabilities are abnormal?

(Perhaps related to how resources are deployed during normal and abnormal development?)

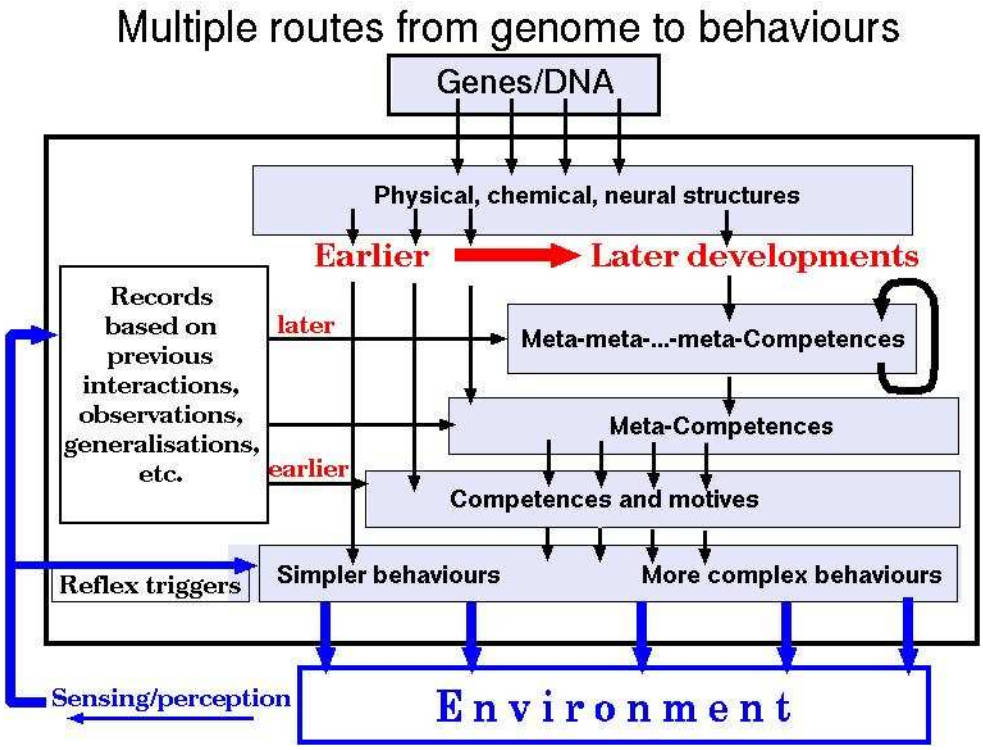
At present AI theories partly specify mechanisms that some neuroscientists seek in brains. And vice versa. But I think most of the research in visual neuroscience is based on false, or at least seriously incomplete, specifications of what needs to be explained.

[This is also true of the widely admired Perceptual Control Theory of William T Powers, developed in parallel with a lot of AI work, but with mutual ignorance, mostly.
http://www.pctweb.org/psy/psychology.html]

Some features of epigenetic patterns in intelligent animals

Jackie Chappell and I published an invited paper in 2007 suggesting that attempts to divide **species** into *precocial* (fully formed and competent at or soon after birth/hatching) and *altricial* (born/hatched relatively helpless or incompetent but able to develop sophisticated competences later on) should be replaced by a distinction between **competences** that are "pre-configured" i.e. mostly specified in the genome and possibly also operational before or soon after birth), and those that are "meta-configured" i.e. developed later, on the basis of a combination of information in the

genome and information acquired at various stages during development. There are not just two cases, but a spectrum of cases illustrated by different trajectories through this development architecture:



Routes from the genome on the left develop very early, those to the right develop later and make more use of what has been learnt in the environment and what evolution has timed to grow/develop later though based partly on the genome. These ideas were first presented in Jackie Chappell and Aaron Sloman, (2007,) Natural and artificial meta-configured altricial information-processing systems, in *International Journal of Unconventional Computing,* 3, 3, pp. 211--239, http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#717

In effect, this replaces Waddington's idea of a fixed epigenetic landscape with a fitness landscape whose specification is constantly being modified during the life of an individual by interactions between newly expressed features of the genome and results of earlier environmental influences.

A QUESTION ABOUT PROCESSING POWER NEEDED

I suspect that when we have adequate specifications of what needs to be explained we may realise that the computational powers of brains vastly exceed the powers assumed by current theories and models of how neurons function.

If that's right, e.g. if there's a huge amount of complex computation going on within each neuron (using chemistry, or special properties of microtubules?) then current estimates of when AI systems will match the computational power of brains may be *grossly* underestimating how far we still have to go in order to produce adequate hardware.

John von Neumann anticipated this possibility in the 1958 book written while he was dying of cancer:
The Computer and the Brain
(Silliman Memorial Lectures) (3rd Edition, with Foreword by Ray Kurzweil. 2012)

There's lots more to be said, but I have gone on too long already. I am steadily accumulating examples and theoretical discussion on the Meta-Morphogenesis web site.
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html

I'll be very happy to add counter-arguments, new proposals or new examples there, or links to online materials challenging or enriching what's already there.

Aaron

http://www.cs.bham.ac.uk/~axs
Aaron Sloman,
Honorary Professor of Artificial Intelligence and Cognitive Science
(Retired, but still working full time)
School of Computer Science,
The University of Birmingham Edgbaston Birmingham B15 2TT UK

References

- A. Sloman and B.S. Logan, 1999,
  Building cognitively rich agents using the Sim_Agent toolkit (invited paper).
  Communications of the Association for Computing Machinery, 42, 3, pp. 71--77, March,
  http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#49
- A. Sloman, 2012 -- on, The Meta-Morphogenesis Web site
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html
-
-
-
-