Transcript of "Aaron Sloman interviewed by Adam Ford"
Artificial Intelligence - Psychology
Oxford December 2012

Aaron Sloman -- Adam Ford,
(Based on a draft transcript by Dylan Holmes [*])

What is this document?

This a transcript of a 57 minute video on YouTube, of Aaron Sloman being interviewed by Adam Ford, at the Artificial General Intelligence (AGI) Winter Conference, St Anne's College, Oxford University, December 2012.

The interview was on 9th December 2012: Adam Ford posed a number of questions to direct the interview, as indicated by the main section headings below.

A slightly reduced quality downloadable version of the video is available here:
Interview (120 MB WEBM):
http://www.cs.bham.ac.uk/research/projects/cogaff/movies/AaronSloman-Adam-Ford-agi-dec-2012-oxford.webm

Revised extended transcript of interview (this file!) (HTML):
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/transcript-agi-interview.html

This transcript is also available as PDF (derived from the html version: may be slightly out of date):
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/transcript-agi-interview.pdf

An earlier version of this HTML transcript was copied to H+ Magazine by Adam Ford on 29th Sept 2013
http://hplusmagazine.com/2013/09/29/aaron-sloman-on-psychology-and-artificial-intelligence-oxford-interview/

An expanded PDF version of the interview, with additional comments and references, is here:
http://www.cs.bham.ac.uk/research/projects/cogaff/14.html#1401

NOTE: Related videos
A separate 2hr 30 min video, by Adam Ford, of Aaron Sloman's tutorial on The Meta-Morphogenesis project, mentioned in this interview, was recorded the next day and made available here:
Aaron Sloman - Meta-Morphogenesis - How a Planet can produce Minds, Mathematics and Music
27 Nov 2015
Now also at:
http://www.scifuture.org/metamorphogenesis-how-a-planet-can-produce-minds-mathematics-and-music-aaron-sloman/
(No transcript of the tutorial exists.)

In June 2014 Adam Ford recorded an impromptu interview with Aaron Sloman following his experiences as a "judge" in the 2014 "Turing test" event at the Royal Society of Londaon. That interview is here:
https://www.youtube.com/watch?v=ACaJlJcsvL8
The Turing Test - Did Eugene Goostman pass the Turing Test?
See also http://www.cs.bham.ac.uk/research/projects/cogaff/misc/turing-test-2014.html

---

[*]History of this transcript

On 2013-09-07 Dylan Holmes posted a first draft transcript of the tutorial, at http://aurellem.org/thoughts/html/sloman.html

The transcript was then corrected and edited by Aaron Sloman, who also inserted some additional text and links, and changed the original text in a few places where the words uttered in the interview could be improved on. A short **Further Reading** section was added, near the end (which may be expanded later). The result was this document, initially located in a different place, and moved here on 26 Jul 2015.

Because of those changes, the spoken words are different from the corresponding text below, in some places. Usually the changes have merely added information.

In December 2013, an expanded PDF version of this document (a possible book chapter), with some historical background and additional information was made available here. Perhaps that will one day be re-written and published.

I am extremely grateful to Dylan for the amount of effort he put into producing the first draft of this transcript, which was remarkably accurate.

This document is liable to change. So, if you use or comment on this text please include a URL if possible, so that readers can see the original (or the latest version thereof). Please report errors, corrections, improvements to A.Sloman @ cs.bham.ac.uk

---

Original Editor's note (Dylan Holmes):

This is a working draft transcript which I made of this nice interview of Aaron Sloman. Having just finished one iteration of transcription, I still need to go in and clean up the formatting and fix the parts that I misheard, so you can expect the text to improve significantly in the near future.

To the extent that this is my work, you have my permission to make copies of this transcript for your own purposes. Also, feel free to e-mail me with comments or corrections.

You can send mail to Dylan at `transcript@aurellem.org`.
His current web site (26 Jul 2015): https://logical.ai/

Table of Contents

1 Introduction: Background, from mathematics through philosophy to AI

1.1 Aaron Sloman evolves into a philosopher of AI

[0:09] My name is Aaron Sloman. My first degree many years ago in Cape Town University was in Physics and Mathematics, and I intended to go on and be a mathematician. I came to Oxford and encountered philosophers — I had started reading philosophy and discussing philosophy before then, and then I found that there were philosophers who said things about mathematics that I thought were wrong, so I gradually got more and more involved in philosophical discussions and switched to doing a philosophy DPhil.

Then I became a philosophy lecturer and about six years later, in 1969, I was introduced to artificial intelligence (by Max Clowes) when I was a lecturer at Sussex University in philosophy.

I very soon became convinced that the best way to make progress in areas of philosophy, including philosophy of mathematics, which I felt I hadn't dealt with adequately in my DPhil, philosophy of mind, philosophy of language and other things—the best way was to try to design and test working fragments of mind and maybe eventually put them all together, but initially just working fragments that would do various things.

[1:12] I learned to program and ~ with various other people including Margaret Boden whom you've interviewed (here) ~ —helped develop an undergraduate degree in AI and other things and also began to do research in AI, which I thought of as doing philosophy, primarily.

[1:29] I later moved to the University of Birmingham — I came in 1991 — and I've been retired for a while (since 2002), but I'm not interested in golf or gardening so I just go on doing full time research, and my department is happy to keep me on without paying me, and provides space and resources so that I can continue meeting bright people including at conferences and I try to learn and make progress if I can.

1.2 AI is hard, in part because there are tempting non-problems.

One of the things I've learnt and understood more and more over the many years — forty years or so since I first encountered AI — is how hard the problems are, and in part that's because it's very often tempting to *think* the problem is something different from what it actually is. And then people design solutions to the non-problems, and I think of most of my work now as just helping to clarify what the problems are: what is it that we're trying to explain — and maybe this is leading into what you wanted to talk about:

I now think that one of the ways of getting a deep understanding of that is to find out what were the problems that biological evolution solved, because we are a product of *many* solutions to *many* problems, and if we just try to go in and work out what the whole system is doing, we may get it all wrong — badly wrong!

2 What problems of intelligence did evolution solve?

2.1 Intelligence consists of solutions to many evolutionary problems; no single development (e.g. communication) was key to human-level intelligence.

[2:57] Well, first I would challenge the assumption that we are the dominant species (as suggested in the question not recorded in the video.). I know it looks like that but actually if you count biomass, if you count number of species, if you count number of individuals, the dominant species are microbes — maybe not one of them but anyway they're the ones who dominate in that sense, and furthermore we are mostly — we are largely composed of microbes, without which we wouldn't survive.

[3:27] There are things that make humans (you could say) superior to other animals in some respects, and inferior in others. This resulted from a collection of developments of which there isn't any single one. For instance, some people suggest that human language changed everything. By our human language, they mean human communication in words, but I think that was a later development from what must have started as the use of richly structured *internal* forms of representation — which are there in nest-building birds, in pre-verbal children, in hunting mammals — because you can't take in information about a complex structured environment in which things can change, where you may have to be able to work out what's possible and what isn't possible, without having some way of

representing the components of the environment, their relationships, the kinds of things they can and can't do, the kinds of things you might or might not be able to do — and *that* kind of capability needs internal languages.

I and colleagues at Birmingham have been referring to them as "generalized languages" (GLs) because some people object to .... using the term "language" to refer to something that isn't used for communication. But, from my viewpoint, not only humans but many other animals developed abilities to do things to their environment to make them more friendly to themselves, which depended on being able to represent possible futures, possible actions, and work out what's the best thing to do. (So they also used internal languages -- GLs.)

[5:13] And nest-building in corvids for instance—crows, magpies, rooks, and so on — are way beyond what current robots can do, and in fact I think most humans would be challenged if they had to go and find a collection of twigs, one at a time, maybe bring them with just one hand — or with your mouth — and assemble them into a structure that is shaped like a nest, and is fairly rigid, so that you could trust your eggs in them when wind blows. But they (those birds) are doing it.

They're not our evolutionary ancestors, but they're an indication — that example is an indication — of what must have evolved in order to provide control over the environment in *that* species.

> **Insert:** Movies and photos available at the University of Oxford Behavioural Ecology Research Group provide examples of what New Caledonian crows can do.

2.2 Speculation about how communication might have evolved from internal languages.

[5:56] So I think hunting mammals, fruit-picking mammals, mammals that can rearrange parts of the environment, provide shelters, …. also needed to have ways of representing possible futures, not just what's there in the environment. I think at a later stage, that developed into a form of communication — or rather the *internal* forms of representation became usable as a basis for providing content to be communicated. That happened, I think, initially through performing actions that expressed intentions, and probably led to situations where an action (for instance, moving some large object) was performed more easily, or more successfully, or more accurately if it was done collaboratively. So someone who had worked out what to do might start doing it, and then a conspecific might be able to work out what the intention is, because that person has the *same* forms of representation and can build theories about what's going on (in the actor's mind), and might then be able to help.

[7:11] You can imagine that if that started happening more (a lot of collaboration based on inferred intentions and plans) then sometimes the inferences might be obscure and difficult, so the *actions* might be enhanced to provide signals as to what the intention is, and what the best way is to help, and so on.

> **Insert:** (So actions enhanced to provide communication during collaboration may have been precursors to separately signed communications.)

[7:35] So, this is all hand-waving and wild speculation, but I think it's consistent with a large collection of facts which one can look at — and find if one looks for them, but [facts which] one won't notice if one doesn't look for them — about the way children, for instance, who can't yet talk, communicate, and the things they'll do, like going to the mother and turning the face to point in the direction where the child wants it to look and so on; that's an extreme version of action indicating intention.

Insert: A slide presentation elaborating this idea is here.

[8:03] Anyway. That's a very long roundabout answer to one conjecture that the use of communicative language is what gave humans their unique power to create and destroy and whatever, and I'm saying that if by that you mean *communicative* language, then there was something before that which was *non*-communicative language, and I suspect that non-communicative languages continue to play a deep role in *all* human perception —in mathematical and scientific reasoning, in problem solving — and we don't understand very much about it.

[8:48] I'm sure there's a lot more to be said about the development of different kinds of senses, the development of brain structures and mechanisms to support all that, but perhaps I've droned on long enough on that question.

3 How do language and internal states relate to AI?

[9:09] Well, I think most of the human and animal capabilities that I've been referring to are not yet to be found in current robots or computing systems, and I think there are two reasons for that: one is that it's intrinsically very difficult; I think that in particular it may turn out that the forms of information processing that one can implement on digital computers as we currently know them may not be as well suited to performing some of these tasks as other kinds of computing about which we don't know so much — for example, I think there may be important special features about *chemical* computers which we may talk about later.

3.1 In AI, false assumptions can lead investigators astray.

[9:57] So, one of the problems then is that the tasks are hard … but there's a deeper problem as to why AI hasn't made a great deal of progress on these problems that I'm talking about, and that is that most AI researchers assume things—and this is not just AI researchers, but [also] philosophers, and psychologists, and people studying animal behavior—they all make assumptions about what it is that animals or humans do, for instance they make assumptions about what vision is for, or assumptions about what motivation is and how motivation works, or assumptions about how learning works, and then they try — the AI people try — to model [or] build systems that perform those assumed functions. So if you get the *functions* wrong, then even if you implement some of the functions that you're trying to implement, they won't necessarily perform the tasks that the initial objective was to imitate, for instance the tasks that humans, and nest-building birds, and monkeys and so on can perform.

3.2 Example: Vision is not just about finding surfaces, but about finding affordances.

[11:09] I'll give you an simple example. It is often assumed that the function of vision in humans (and in other animals with good eyesight and so on) is to take in optical information that hits the retina, forming often changing patterns of illumination, where there are sensory receptors that detect those patterns, and then somehow from that information (plus maybe other information gained from head movement or from comparisons between two eyes) to work out what there was in the environment that produced those retinal patterns. And that is often taken to mean "Where were the surfaces off which the light bounced before it came to me?". So you essentially think of the task of the visual system as being to reverse the image formation process. The 3D structure is out there, the lens causes the image to form on the retina, and then the brain tries use the (possibly changing) image contents to build models of the 3D structures and processes out there.

Insert: This sort of theory is often attributed to David Marr. Compare the conjecture about brains as building models of the environment, in the final chapter of Kenneth Craik's 1943 book *The Nature of Explanation*

That's a very plausible theory about vision, and it may be that that's a *subset* of what human vision does, but I think James Gibson pointed out that that kind of thing is not necessarily going to be very useful for an organism, and it's very unlikely that that's the main function of perception in general, namely to produce some physical description of what's out there.

Insert: J. J. Gibson, *The Ecological Approach to Visual Perception,* 1979,

I think there are far more kinds of "affordance" than Gibson noticed, and have presented some ideas about how to extend and generalise his work in this slide presentation: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#gibson What's vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond

[12:37] What does an animal *need*? It needs to know what it can do, what it can't do, what the consequences of its actions will be …. He introduced the word *affordance*: from his point of view, the functions of vision, or more generally perception, are to inform the organism of what the *affordances* are for action, where that would mean what the animal, *given* its morphology (what it can do with its mouth, its limbs, and so on, and the ways it can move) what it can do, what its needs are, what the obstacles are, and how the environment supports or obstructs those possible actions.

[13:15] And that's a very different collection of information structures that you need from, say, "where are all the surfaces?": if you've got all the surfaces, *deriving* the affordances would still be a major task.

So, if you think of the perceptual system as primarily (for biological organisms) being devices that provide information about affordances and so on, then the tasks look very different. And I think most of the people working on.., doing research on.., computer vision in robots, haven't taken all that on board, so they're trying to get machines to do things which, even if they were successful, would not make the robots very intelligent (and in fact, even the tasks they're trying to get robots to do are not really easy to do, and they don't succeed very well— although, there is progress: I shouldn't disparage it too much!)

3.3 Online and offline intelligence

[14:10] It gets more complex as animals get more sophisticated. So, I like to make a distinction between online intelligence and offline intelligence. So, for example, if I want to pick something up — like this leaf <he plucks a leaf from the table> — I was able to select it from all the others in there, and while moving my hand towards it, I was able to guide its trajectory, making sure it was going roughly in the right direction — as opposed to going out there, which wouldn't have been able to pick it up — and these two fingers ended up with a portion of the leaf between them, so that I was able to tell when I'm ready to do that <he clamps the leaf between two fingers> and at that point, I clamped my fingers and then I could pick up the leaf.

[14:54] (Whereas, —) That's an example of online intelligence: during the performance of an action (both from the stage where it's initiated, and during the intermediate stages, and where it's completed) I'm taking in information relevant to controlling all those stages, and that relevant information keeps changing. That means I need stores of transient information which gets discarded almost immediately and is then replaced or modified. That's online intelligence. And there are many forms; that's just one

example, and Gibson discussed quite a lot of examples which I won't try to replicate now.

[15:30] But in offline intelligence, you're not necessarily actually *performing* the actions when you're using your intelligence; you're thinking about *possible* actions. So, for instance, I could think about how fast or by what route I would get back to the lecture room if I wanted to get to the next talk or something. And I know where the door is, roughly speaking, and I know roughly which route I would take, when I go out, I should go to the left rather than to the right, because I've stored information about where the spaces are, where the buildings are, where the door was that we came out — but in using that information to think about that route, I'm not actually performing the action. I'm not even *simulating* it in detail: the precise details of direction and speed and when to clamp my fingers, or when to contract my leg muscles when walking, are all irrelevant to thinking about a good route, or thinking about the potential things that might happen on the way. Or what would be a good place to meet someone who I think frequents a particular — bar or something — I don't necessarily have to work out exactly *where* the person's going to stand, or from what angle I would recognize them, and so on.

[16:46] [So,] Offline intelligence — [which] I think became not just a human competence. I think there are other animals that have aspects of it: Squirrels are very impressive. Gray squirrels, at least, as you watch them defeating squirrel-proof bird feeders, seem to have a lot of that [offline intelligence], as well as the online intelligence used when they eventually perform the action they've worked out would get them to the nuts.

[17:16] I think that what happened during our evolution was that mechanisms for acquiring and processing and storing and manipulating information that was more and more remote from the performance of actions, developed. An example is taking in information about where locations are that you might need to go to infrequently: There's a store of a particular type of material that's good for building or putting on roofs of houses (or something); out around there in some direction. There's a good place to get water somewhere in another direction. There are people that you'd like to go and visit in another place, and so on.

[17:59] So taking in information about an extended environment and building it into a structure that you can make use of for different purposes is another example of offline intelligence. And when we do that, we sometimes use only our brains, but in modern times, we also learned how to make maps on paper and walls and so on. And it's not clear whether the stuff inside our heads has the same structures as the maps we make on paper: the maps on paper have a different function; they may be used to communicate with others, or meant for *looking* at, whereas the stuff in your head you don't *look* at; you use it in some other way.

[18:46] So, what I'm getting at is that there's a great deal of human intelligence (and animal intelligence) which is involved in what's possible in the future, what exists in distant places, what might have happened in the past (sometimes you need to know why something is as it is, because that might be relevant to what you should or shouldn't do in the future, and so on), and I think there was something about human evolution that extended that offline intelligence way beyond that of other animals. And I don't think it was *just* human language, (but human language had something to do with it) but I think there was something else that came earlier than language (used for communication) which involves the ability to use your offline intelligence to discover something that has a rich mathematical structure.

3.4 Example: Even toddlers use sophisticated geometric knowledge

I'll give a simple example. If you look through a gap, you can see something that's on the other side of the gap. Now, you *might* see what you want to see, or you might see only part of it. If you want to see more of it, which way would you move?

Well, you could either move *sideways*, and see through the gap—and see roughly the same amount but a different part of it, if it's a room or whatever, or you could move *towards* the gap and then your view will widen as you approach the gap. Now, there's a bit of mathematics in there, insofar as you are implicitly assuming that information travels in straight lines, and, as you go closer to a gap, the straight lines that you can draw from where you are through the gap, widen as you approach that gap. Now, there's a kind of theorem of Euclidean geometry in there which I'm not going to try to state very precisely (and as far as I know, wasn't stated explicitly in Euclidean geometry) but it's something every human toddler learns. (Maybe other animals also learnt it? I don't know.) [21:00]

But there are many more things: more actions to perform, to get you more information about things, actions to perform to conceal information from other people, actions that will enable you, to operate, to act on, a rigid object in one place in order to produce an effect on another place.

So, there's a lot of stuff that involves lines and rotations and angles and speeds, and so on, that I think humans (and, to a lesser extent, other animals) developed the ability to think about in a generic way. [Using a combination of biological evolution and individual learning].

That meant that you could take the generalizations out of the particular contexts and then re-use them in a new contexts in ways that I think are not yet represented at all in AI and in theories of human learning in any explicit way — although some people are trying to study learning of mathematics.

> **Insert:** There has been a vast amount of research on how to give robots the ability to accumulate observational evidence and derive useful high probability generalisations. But the re-usable mathematical generalisations do not take the form of high probability generalisations based on large amounts of evidence: mathematical reasoning about geometric relationships is not concerned with probabilities, any more than theorems of arithmetic such as "3+5=8" or "There are infinitely many prime numbers" summarise statistical evidence.

4 Animal intelligence

4.1 The priority is *cataloguing* what competences have evolved, not ranking them.

[22:03] I wasn't going to challenge the claim that humans can do more sophisticated forms of tracking, just to mention that there are some things that other animals can do which are in some ways comparable, and some ways superior to things that humans can do. In particular, there are species of birds and also, I think, some rodents — squirrels, or something else — that can hide nuts and remember where they've hidden them, and go back to them. And there have been tests which show that some birds are able to hide tens — you know, eighty or something nuts — and to remember which ones have been taken, which ones haven't, and so on. And I suspect most humans can't do that. I wouldn't want to say categorically that maybe we couldn't, because humans are very varied, and also a few people can develop particular competences through training. But it's certainly not something I can do.

4.2 AI can be used to test philosophical theories

[23:01] But I also would like to say that I am not myself particularly interested in trying to align animal intelligences according to any kind of scale of superiority; I'm just trying to understand what it was that biological evolution produced, and how it works, and I'm interested in AI *mainly* because I think that when one comes up with theories about how these things work, one needs to have some way of testing the theory.

And AI provides ways of implementing and testing theories that were not previously available: Immanuel Kant [e.g. in his *Critique of Pure Reason* (1781)] was trying to come up with theories about how minds work, but he didn't have any kind of a mechanism that he could build to test his theory about the nature of mathematical knowledge, for instance, or how concepts were developed from babyhood onward. Whereas now, if we do develop a theory, we have a criterion of adequacy, namely it should be precise enough and rich enough and detailed to enable a model to be built. And then we can see if it works.

[24:07] If it works, it doesn't mean we've proved that the theory is correct; it just shows it's a candidate. And if it doesn't work, then it's not a candidate as it stands; it would need to be modified in some way.

5 Is Artificial General Intelligence (AGI) feasible?

5.1 It's misleading to compare the brain and its neurons to a computer made of transistors

[24:27] I think there's a lot of optimism based on false clues: for example, one of the false clues is to count the number of neurons in the brain, and then talk about the number of transistors you can fit into a computer or something, and then compare them. This comparison may be undermined by finding out more about how synapses work. For example, I once heard a lecture in which a neuroscientist claimed that a typical synapse in the human brain has computational power comparable to the Internet a few years ago, because of the number of different molecules that are doing things, the variety of types of things that are being done in those molecular interactions, and the speed at which they happen. If you somehow count up the number of operations per second or something, then you get these comparison figures.

5.2 For example, brains may rely heavily on chemical information processing

Now even if the details aren't right, there may just be a lot of information processing that's going on in brains at the *molecular* level, not the neural level. Then, if that's the case, the processing units will be orders of magnitude larger in number than the number of neurons. And it's certainly the case that all the original biological forms of information processing were chemical; there weren't brains around, and still aren't in most microbes. And even when humans grow their brains, the process of starting from a fertilized egg and producing this rich and complex structure is, for much of the time, under the control of chemical computations, chemical information processing— of course constrained by sources of physical materials and energy, also.

[26:25] So it would seem very strange if all that capability was something thrown away when you've got a brain, and all the information processing, the challenges that were handled in making a brain, [were totally disconnected from the mechanisms of a complete functioning brain …] This is hand-waving on my part; I'm just saying that we *might* learn that what brains do is not what we think they do, and that problems of replicating them are not what we think they are, solely in terms of numerical estimate of time scales, the number of components, and so on.

5.3 Brain algorithms may be optimized for certain kinds of information processing other than bit manipulations

[26:56] But apart from that, the other basis of skepticism [about the imminence of AGI] concerns how well we understand what the problems are. I think there are many people who try to formalize the problems of designing an intelligent system in terms of streams of information thought of as bit streams, or collections of bit streams, and they think of as the problems of intelligence as being the construction or detection of patterns in those streams, and perhaps not just detection of patterns, but detection of patterns that are usable for sending *out* streams to control motors, and so on, in order to achieve various goals.

And that way of conceptualizing the problem may lead on the one hand to oversimplification, so that the things that *would* be achieved, if those goals were achieved, may be much simpler, and in some ways inadequate for the replication of human intelligence, or the matching of human intelligence— or, for that matter, squirrel intelligence—but in another way, it may also make the problems harder: it may be that some of the kinds of things that biological evolution has achieved can't be done that way. And one of the ways that might turn out to be the case is not because it's impossible *in principle* to do some of the information processing on artificial computers, based on transistors and other bit-manipulating mechanisms; but it may just be that the *computational complexities* of solving problems, i.e. the complexities of processes or finding solutions to complex problems [using bit manipulations], are much greater, and therefore you might need a much larger universe than we have available in order to do things -- than if the underlying mechanisms were different. Other [non bit-manipulating] information processing mechanisms might be better tailored to particular sorts of computation.

5.4 Example: find the shortest path by dangling strings

[29:07] There's a very well-known example, which is finding the shortest route if you've got a collection of roads, and they may be curved roads, and lots of tangled routes from A to B to C, and so on. If you start at A and you want to get to Z, a place somewhere on that map, the process of finding the shortest route will involve searching through all these different possibilities and rejecting some that are longer than others and so on. But if you make a model of that map out of strings, where the strings are all laid out on the maps and so have the lengths of the routes, then, if you hold the two knots in the network of strings which correspond to the start point and the end point, and just *pull them apart*, then the bits of string that you're left with in a straight line will give you the shortest route, and that process of pulling just gets you the solution very rapidly in a parallel computation, where all the others just hang by the wayside, so to speak. [**Note Added 1 Oct 2013:** This is an old idea, mentioned for example by Hubert Dreyfus and John Haugeland, "The computer as a mistaken model of the mind", in *Philosophy of Psychology*, Ed. S.C.Brown, London, Macmillan, 1974, pp. 247--258. A partly similar computer model of finding the shortest route in a network of roads could store at every junction a list of names of all other junctions, and for each other junction an indicating of which route goes to that target junction. Pre-computing the best direction from node X to each of the other nodes, for every X would be worthwhile if the network is to he used often, without any searching. But the string version of the roadmap does not need such pre-computing.]

5.5 In sum, we know surprisingly little about the kinds of problems that evolution solved, and the manner in which they were solved.

[30:15] Now, I'm not saying brains can build networks of string and pull them or anything like that; that's just an illustration of how, if you have the right representation, correctly implemented—or suitably implemented—for a problem, then you can avoid very combinatorially complex searches, which will maybe grow exponentially with the number of components in your map, whereas with this thing, the time it takes won't depend on how many strings you've got on the map; you just pull, and time required will depend only on the shortest route that exists in there, even if that shortest route wasn't obvious from the original map.

[30:59] So, that's a rather long-winded way of formulating the conjecture — a roundabout way of supporting the conjecture — that there may be something about the way molecules perform computations where they have the combination of continuous change as things move through space and come together and move apart, and whatever — and also snap into states that then persist, so, as a result of quantum mechanisms, you can have stable molecular structures which are quite hard to separate. They may need catalytic processes to separate them, or extreme temperatures, or strong forces, but they may nevertheless be able to move very rapidly in some conditions in order to perform computations.

[31:49] Now there may be things about that kind of structure that enable searching for solutions to *certain* classes of problems to be done much more efficiently (by brains) than anything we could do with computers. It's just an open question.

[32:04] So it *might* turn out that we need new kinds of technology that aren't on the horizon in order to replicate the functions that animal brains perform —or, it might not. I just don't know. I'm not claiming that there's strong evidence for that; I'm just saying that it might turn out that way, partly because I think we know less than many people think we know about what biological evolution achieved.

[32:28] There are some other possibilities: we may just find out that there are shortcuts no one ever thought of, and it will all happen much more quickly—I have an open mind; I'd be surprised, but it could turn up.

6 A singularity of cognitive catch-up

6.1 What if it will take a lifetime to learn enough to make something new?

[32:59] There *is* something that worries me much more than the singularity that most people talk about, which is machines achieving human-level intelligence and perhaps taking over [the] planet or something. There's also what I call the *singularity of cognitive catch-up*, the SOCC, singularity of cognitive catch-up, which I think we're close to, or maybe have already reached—I'll explain what I mean by that.

One of the products of biological evolution—and this is one of the answers to your earlier questions which I didn't get on to—is that humans have not only the ability to make discoveries that none of their ancestors have ever made, but to shorten the time required for similar achievements to be reached by their offspring and their descendants. So once we've, for instance, worked out ways of doing complex computations, or ways of building houses, or ways of finding our way around,…our children

don’t need to work it out for themselves by the same lengthy trial and error procedure; we can help them get there much faster.

Okay, well, what I’ve been referring to as the "Singularity of Cognitive Catch-up" depends on the fact that’s fairly obvious—and has often been commented on—that in case of humans, it’s not necessary for each generation to learn what previous generations learned *in the same way*. And we can speed up learning once something has been learned -- speed up the learning by new people. And that has meant that the social processes that support that kind of education of the young can enormously accelerate what would have taken…perhaps thousands [or] millions of years for evolution to produce: it can happen in a much shorter time.

[34:54] But here’s the catch: in order for a new advance to happen — e.g. for something new to be discovered that wasn’t known before, like Newtonian mechanics, or the theory of relativity, or Beethoven’s musical style, or whatever — the individuals have to have traversed a significant amount of what their ancestors have learned, even if they do it much faster than their ancestors, to get to the point where they can see the gaps, the possibilities for going further than their ancestors, or their contemporaries, have done.

[35:27] Now in the case of knowledge of science, mathematics, philosophy, engineering and so on, there’s been a lot of accumulated knowledge. And humans are living a *bit* longer than they used to, but they’re still living for [whatever it is], a hundred years, or for most people, less than that. So you can imagine that there might come a time when in a normal human lifespan, it’s not possible for anyone to learn enough to understand the scope and limits of what’s already been achieved in order to see the potential for going beyond it and to build on what’s already been done to make that…those future steps.

[36:10] So if we reach that stage, we will have reached the singularity of cognitive catch-up because the process of education that enables individuals to learn faster than their ancestors did, is the catching-up process, and it may just be that we at some point reach a point where catching up can only happen within a [whole] lifetime of an individual, and after that they’re dead and they can’t go beyond. And I have some evidence that there’s a lot of that around, because I see a lot of people coming up with what *they* think of as new ideas which they’ve struggled to come up with, but actually they just haven’t taken in some of what was done by other people, in other places before them.

I think that’s [the case] *despite* the availability of search engines, which now make it easier for people to get information.

For instance, when I was a student, if I wanted to find out what other people had done in the field, it was a laborious process of going to the library, getting books, and so on, whereas now, I can often do things in seconds that would have taken hours. So that means that if [only] seconds are needed for that kind of work, my lifespan has been extended by a factor of ten or something

So maybe that *delays* the singularity, but it may not delay it enough. But that’s an open question; I don’t know. And it may just be that in some areas, this is more of a problem than others. For instance, it may be that in some kinds of engineering, we’re handing over more and more of the work to machines anyway, and they can go on doing it. So for instance, most of the production of computers now is done by a computer-controlled machinery. Although some of the design work is done by humans, a lot of *detail* of the design is done by computers, and they produce the next generation, which then produces the next generation, and so on.

[37:57] I don’t know if humans can go on having major advances, so it’ll be kind of sad if we can’t.

7 Spatial reasoning: a difficult problem

[38:15] Okay, well, there are different branches of mathematics, and they have different properties. So, for instance, a lot of mathematics can be expressed in terms of logical structures or algebraic structures and those are pretty well suited for manipulation on computers, and if a problem can be specified using the logical/algebraic notation, and the solution method requires creating something in that sort of notation, then computers are pretty good.

There are lots of mathematical tools around—there are theorem provers and proof checkers, and all kinds of things, which couldn’t have existed fifty, sixty years ago, and they will continue getting better.

But there was something that I was alluding to earlier when I gave the example of how you can reason about what you will see by changing your position in relation to a door, where what you are doing is using your grasp of spatial structures; and how, as one spatial relationship changes, namely you come closer to the door or move sideways and parallel to the wall or whatever, then other spatial relationships change in parallel, so the lines from your eyes through to other parts of the room on the other side of the doorway change: they spread out more as you go towards the doorway, but as you move sideways, they don’t spread out differently, but focus on different parts of the internal … they access different parts of the other room.

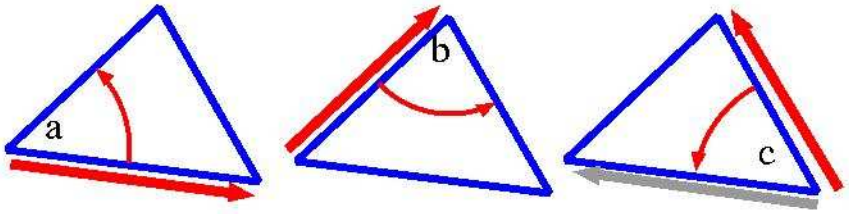
Now, those are examples of ways of thinking about relationships and changing relationships which are not the same as thinking about what happens if I replace this symbol with that symbol, or if I substitute this expression in that expression in a logical formula. And at the moment, I do not believe that there is anything in AI amongst the mathematical reasoning community, the theorem-proving community, that can model the processes that go on when a young child starts learning to do Euclidean geometry and is taught things about—for instance, I can give you a proof that the angles of any triangle add up to a straight line, 180 degrees.

7.1 Example: Spatial proof that the angles of any triangle add up to a half-circle

There are standard proofs which involve starting with one triangle, then adding a line parallel to the base.

One of my former students, Mary Pardoe, came up with a proof when she was teaching mathematics in a school, which I shall demonstrate with this <he holds up a pen> — can you see it?

Suppose I have a triangle here that’s got three sides, if I put this pen on it, on one side — let’s say the bottom — I can rotate it through one of the angles until it lies along …another side, and then if necessary move it up to the other end of that side. Then I can rotate it again, until it lies on the third side. And then I’ll rotate it again and it must eventually end up on the *original* side, but it will have changed the direction it is pointing in — and it won’t have crossed over itself, so it must have gone through a half-circle, and that says that the three angles of a triangle add up to a rotation of half a circle, which is a beautiful kind of proof, and almost anyone can understand it.



Mary Pardoe's proof of the triangle sum theorem.

Some mathematicians don't like it, because they say it hides some of the assumptions, but nevertheless, as far as I'm concerned, it's an example of a human ability to do reasoning which, once you've understood it, you can see will apply to any triangle.

> **Insert (24 Dec 2013):** Understanding that that is a proof requires understanding why the possibility of the process I have described does not depend on the location, size, orientation, colour, or even perfect depiction of the triangle. A mathematical learner has to grasp invariant features of a structure or process.

It's got to be a planar triangle — not a triangle on a globe, because on a globe the angles can add up to more than half a rotation; you can have three *right* angles if you have an equator…a line on the equator, and a line going up to to the north pole of the earth, and then you have a right angle and then another line going down to the equator, and you have a right angle, right angle, right angle, and they add up to more than a straight line.

But that's because the triangle isn't in the plane, it's on a curved surface. In fact, that's one of the differences…definitional differences you can take between planar and curved surfaces: how much the angles of a triangle add up to.

[42:28] But our ability to *visualize* and notice the *generality* in that process, and see that you're going to be able to do the same thing using triangles that stretch in all sorts of ways, or if it's a million times as large, or if it's made of something different; if it's drawn in different colors or whatever — none of that's going to make any difference to the essence of that process. And that ability to see the commonality in a spatial structure which enables you to draw some conclusions with complete certainty—subject to the possibility that sometimes you make mistakes, but when you make mistakes, you can discover them, as has happened in the history of geometrical theorem proving. Imre Lakatos had a wonderful book called *Proofs and Refutations* — which I won't try to summarize — but he has examples: mistakes were made; that was because people didn't always realize there were subtle sub-cases which had slightly different properties, and they didn't take account of that. But once they're noticed, you can rectify that. [43:25]

7.2 Geometric results are fundamentally different from experimental results in chemistry or physics.

[43:28] But it's not the same as doing experiments in chemistry and physics, where you can't be sure it'll be the same on Mars or at a high temperature, or in a very strong magnetic field. With geometric reasoning, in some sense you've got the full information in front of you; even if you don't always notice an important part of it. So, that kind of reasoning [geometrical reasoning], as far as I know, is not currently implemented anywhere in a computer. And most people who do research on trying to model mathematical reasoning, don't pay any attention to that, because of … they just don't think about it. They start from somewhere else, maybe because of how they were educated. I was taught Euclidean geometry at school. Were you?

(Adam ford: Yeah)

Many people are not now. Instead they're taught set theory, and logic, and arithmetic, and algebra, and so on. And so they don't use that bit of their brains, without which we wouldn't have buildings and cathedrals, and all sorts of things we now depend on.

NOTE ADDED: There is a great deal of research on geometrical theorem proving, but normally the axioms and theorems are translated into formalisms based on Cartesian coordinate representations of geometry. So the machines prove theorems about sets of numbers and equations or inequalities relating numbers, not theorems about geometry such as Euclid proved, even if there's a strong structural relationship between the two domains. See also http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-theorem.html

8 Is near-term artificial general intelligence (AGI) likely?

8.1 Two interpretations: a single mechanism for all problems, or many mechanisms unified in one program.

[44:35] Well, this relates to what's meant by "general". When I first encountered the AGI community, I thought that what they all meant by *general* intelligence was *uniform* intelligence — intelligence based on some common simple (maybe not so simple, but) single powerful mechanism or principle of inference.

And there are some people in the community who are trying to produce things like that, often in connection with algorithmic information theory and compressibility of information, and so on. But there's another sense of "general" which means that a system with general intelligence can do lots of different things, like perceive things, understand language, move around, make things, and so on — perhaps even enjoy a joke. That's something that's not nearly on the horizon, as far as I know. Enjoying a joke isn't the same as being able to make laughing noises.
NOTE ADDED:
[Nor is it the same as being able to recognize jokes, or generate jokes, as some AI programs can do. See http://en.wikipedia.org/wiki/Computational_humor]

Given, then, that there are these two notions of general intelligence—there's one that looks for one uniform, possibly simple, mechanism or collection of ideas and notations and algorithms, that will deal with any problem that's solvable — and the other that's general in the sense that it can do lots of different things that are combined into an integrated architecture (which raises lots of questions about how you combine these things and make them work together) and we humans, certainly, are of the second kind: we do all sorts of different things, and other animals also seem to be of the second kind, perhaps not as general as humans.

[46:33] Now, it may turn out that in some near future time, who knows—decades, a few decades—you'll be able to get machines that are capable of solving in a time that will depend on the nature of the problem, but any problem that is solvable, and they will be able to do it in some sort of tractable time — of course, there are some problems that are solvable that would require a larger universe and a longer history than the history of the universe, but apart from that constraint, these machines will be able to do anything and they will have general intelligence.

But to be able to do some of the kinds of things that humans can do -- like the kinds of geometrical reasoning where you look at the shape and you abstract away from the precise angles and sizes and shapes and so on, and realize there's something general there, as must have happened when our ancestors first made the discoveries that were eventually put together in Euclidean geometry -- may require mechanisms of a kind that we don't know about at the moment.

Maybe brains are using molecules and rearranging molecules in some way that supports that kind of reasoning. I'm not saying they are — I don't know, I just don't see any simple…any obvious way to map that kind of reasoning capability onto what we currently do on computers.

[47:57] There is—and I'll just mention this briefly before finishing—there is a kind of thing that's sometimes thought of as a major step in that direction, namely you can build a machine (or a software system) that can represent some geometrical structure, and then be told about some change that's going to happen to it, and it can predict in great detail what will happen. This happens for instance in game engines, where you say "We have all these blocks on the table and I'll drop one other block", and then the program uses Newton's laws and properties of rigidity of the parts and the elasticity and also stuff about geometry and space and so on, to give you a very accurate representation of what will happen when this brick lands on this pile of things: it will bounce and go off, and so on. And you just, with more memory and more CPU power, you can increase the accuracy. [Added: and the program can compute trajectories following a very wide variety of initial configurations, with more or less accuracy depending on the complexity and the duration of the predicted changes.]

But that's totally different from looking at *one* example, and working out what will happen in a whole *range* of cases at a higher level of abstraction, whereas the game engine does it in great detail for *just* this case, with *just* those precise things, and it won't even know what the generalizations are that it's using that would apply to other similar cases. So, in that sense, you may get AGI — artificial general intelligence — pretty soon, but it will be limited in what it can do. And the other kind of general intelligence which combines all sorts of different things, including human spatial geometrical reasoning, and maybe other things, like the ability to find things funny, and to appreciate artistic features and other things may need forms of, types of, mechanism that we don't know about, and I have an open mind about that.

9 Artificial General Intelligence (AGI) impacts

9.1Implications of the two types of general intelligence.

[49:53] Well, as far as the first type [of artificial general intelligence] is concerned, it could be useful for all kinds of applications — there are people who worry about whether a system that has that type of intelligence, might in some sense take over control of the planet. Well, humans often do stupid things, and the machines might do something stupid that would lead to disaster, but I think it's more likely that there would be other things [done by humans[that will lead to disaster— population problems, using up all the resources, destroying ecosystems, and whatever. But certainly it would go on being useful to have these calculating devices.

Now, as for the second kind of artificial general intelligence [combining many different capabilities in one system], I don't know—. If we succeeded at putting together all the parts that we find in humans, we might just make an artificial human, and then we might have some of them as our friends, and some of them we might not like, and some of them might become teachers or whatever, ..., composers, etc.

But that raises a question: could they, in some sense, be superior to us, in their learning capabilities, their understanding of human nature, or maybe their wickedness or whatever? These are all issues on which I expect the best science fiction writers would give much better answers than anything I could do. But I did once fantasize when I wrote a book in 1978 [The Computer Revolution in Philosophy: Philosophy, science and models of mind in the Epilogue], that perhaps if we achieved that kind of thing, that they [the intelligent machines] would be wise, and gentle and kind, and realize that humans are an inferior species, but they have some good features, so they'd keep us in some kind of secluded…restrictive kind of environment, but keep us away from dangerous weapons, and so on. And find ways of cohabiting with us. But that's just fantasy.

Adam Ford: Awesome. Yeah, there's an interesting story *With Folded Hands* where the computers want to take care of us and want to reduce suffering and end up lobotomizing everybody, but keeping them alive so as to reduce the suffering.

Aaron Sloman: Not all that different from *Brave New World* [by Aldous Huxley], where it was done with drugs and so on, but different humans are given different roles in that system.

There's also *The Time Machine*, by H.G. Wells, where, in the distant future, humans have split/evolved into two types: the Eloi, who lived on the surface of the earth, and the Morlocks who lived underground, and were intelligent and unattractive, whereas the Eloi lived on the surface of the planet. They were pleasant and pretty but not very bright, and they were eaten by the Morlocks!

Adam Ford: Yeah ... that's strange, ... in the future.

Aaron Sloman: As I was saying, if you ask science fiction writers, you'll probably come up with a wide variety of interesting answers.

Adam Ford: I certainly have; I have spoken to Kim Stanley Robinson, Sean Williams, David Brin, and, ... … who else?

Aaron Sloman: Did you ever read a story by E.M. Forster called *The Machine Stops* — a very short story, now available on the Internet, written around 1909, about a future time when people sit in their rooms, in front of screens, and they type things, to communicate with one another: and they don't meet. They have debates, and they give lectures to their audiences that way, using the internet.

One of those people is a woman whose son says "I'd like to see you" and she says something like "What's the point? You can already talk to me", but he wants to come and talk to her face to face — I won't tell you how it ends, but .... this was written in about [1909] so it's over a hundred years ago … people are in their rooms, they sit in front of screens, and they type things, and they communicate with one another that way, and they don't meet; they have debates, and they give lectures to their audiences that way.

Adam Ford: Reminds me of the Internet.

Aaron Sloman: Well, yes; he invented it! It was extraordinary that he was able to do that, before most of the components that we now need for it existed.

Adam Ford: [Another person who did that] was Vernor Vinge, a novella called *True Names*.

Aaron Sloman: When was that written?

Adam Ford: The seventies.

Aaron Sloman: Okay, well a lot of the technology was already around then. The original bits of internet were working, in about 1973, I was sitting … 1974, I was sitting at Sussex University trying to use…learn LOGO, the programming language, to decide whether it was going to be useful for teaching AI, and I was sitting at a paper teletype. There was [no screen only] paper coming out. ...[The machine transmitted] ten characters a second from Sussex to UCL [University College London] computer lab by telegraph cable, from there to somewhere in Norway via another cable, from there by satellite to California to a computer in Xerox Palo Alto Research Center [Xerox PARC] where they had implemented a computer with a LOGO system on it, with someone I had met previously in Edinburgh, Danny Bobrow, and he allowed me to have access to this system.

So there I was typing. And furthermore, it was duplex typing, so every character I typed didn't show up on my terminal until it had gone all the way there and echoed back, so I would type, and the characters would come back four seconds later.

[55:26] But that was the Internet, and I think Vernor Vinge was writing after that kind of thing had already started, but I don't know. Anyway.

[55:41] Another…I mentioned H.G. Wells, *The Time Machine*. I recently discovered, because David Lodge had written a sort of semi-novel about him [A Man of Parts], that he [H.G. Wells] had invented Wikipedia, in advance— he had this notion of an encyclopedia that was free to everybody, and everybody could contribute, in a collaborative effort. So, go to the science fiction writers to find out the future — well, a range of possible futures.

Adam Ford: Well the thing is with science fiction writers, they have to maintain some sort of interest in their readers, after all the science fiction which reaches us is the stuff that publishers want to sell, and so there's a little bit of a … a bias towards making a plot device there, and so the dramatic sort of appeals to our amygdala, our lizard brain; will sort of be there obviously, will be mixed in. But I think that they do come up with sort of amazing ideas; I think it's worth trying to make these predictions; I think that we should focus more time on strategic forecasting, I mean take that seriously.

Aaron Sloman: Well, I'm happy to leave that to others; I just want to try to understand these problems that bother me about how things work. And it may be that some would say that's irresponsible if I don't think about what the implications will be. Well, understanding how humans work *might* enable us to make surrogate humans — I suspect it wont happen in this century; I think it's going to be too difficult.

Further Reading

- The Meta-Morphogenesis Project
- Aaron Sloman - Meta-morphogenesis - How a Planet can produce Minds, Mathematics and Music Another video by Adam Ford: Tutorial presentation on the Meta-Morphogenesis Project at the AGI 2012 conference, Oxford.
- Meta-Morphogenesis and Toddler Theorems: Case Studies
- A DRAFT list of types of transitions in biological information-processing

- A new idea developed November 2014:
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/construction-kits.html
  Construction kits required for biological evolution
  (Including evolution of minds and mathematical abilities.)
  The scientific/metaphysical explanatory role of construction kits:

-- fundamental or
-- -- derived;
-- concrete or
-- -- abstract or
-- -- hybrid (e.g. physical configurations and rules constraining the configurations).

Original Date: 2013-08-22T16:07-0400

Authors:
Video produced by Adam Ford (interviewing Aaron Sloman)
Original Transcript of video produced by Dylan Holmes,
later revised by Aaron Sloman

Last Updated:
28 Sep 2013; 19 Oct 2013; 25 Dec 2013; 26 Jul 2015 (moved to new location)

Org version 7.9.3f with Emacs version 24

Revisions produced by A. Sloman using the Poplog Ved Editor