[Preprint]

A Philosophically Motivated View on AI and Robotics

Lars Kunze (Oxford Robotics Institute) interviews
Aaron Sloman (Honorary Professor of Artificial Intelligence and Cognitive Science
University of Birmingham, UK

September 2019

Aaron Sloman holds a BSc (maths and physics at Cape Town (1956). He was a Rhodes Scholar at Balliol College, Oxford 1957-60, initially studying mathematics, but switched to logic then philosophy of mathematics defending Immanuel Kant (DPhil 1962). He taught philosophy at Hull University, then Sussex University 1964-91, apart from a year (1972-3) in Edinburgh converting to AI, as a result of a paper criticising Logicist AI at IJCAI 1971. At Sussex, he helped to start undergraduate teaching in programming and AI and later the School of Cognitive and Computing Sciences (COGS). He helped with the development of the Pop-11 programming language, the Poplog development environment and a lot of AI teaching material. He was a GEC Research Professor 1984-6. In 1991, he moved to Computer Science, University of Birmingham, UK. He retired in 2002, but continued as Honorary Professor of AI and Cognitive Science. He is an elected Fellow of AAAI (1991), a Fellow of AISB (1997), and a Fellow of ECCAI (1999). He was awarded an Honorary DSc from Sussex University in 2006 and is a Fellow of the Alan Turing Institute since 2018. He is currently working full time on the unfunded Turing-inspired Meta-Morphogenesis project[1].

Figure 1: Aaron Sloman, Honorary Professor of Artificial Intelligence and Cognitive Science, University of Birmingham, United Kingdom.

In December 2018, KI talked to Prof. Aaron Sloman about his views on artificial intelligence (AI) and robotics. The interview covers his multidisciplinary background; the development of robots; the integration of robotics and AI as well as the testability of robot intelligence. Finally, the interview addresses real-world applications and future challenges.

Note: This is a pre-print. The final version of the interview is at `https://link.springer.com/article/10.1007/s13218-019-00621-1`.

[1] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`

Background: AI, Robotics & Philosophy

KI: Let's start with your background. When and how did you get in contact with artificial intelligence?

I should have learnt about it in 1960, or earlier, because of the work I was doing. But in fact, I knew nothing until about seven years after I had completed a DPhil defending Immanuel Kant's *Philosophy of Mathematics* finished in 1962. I became a Philosophy Lecturer, first at Hull University, then at Sussex University.

While I was at Sussex University, one of the leading, highly charismatic, AI researchers in vision, Max Clowes (pronounced "Clews"), joined Experimental Psychology and we became friends. I learnt about attempts to go from images to descriptions of what is in a scene, which at that time was unusual. Research on pattern recognition involved attaching labels to images or parts of images, but Clowes and a few others had begun to develop programs that could produce *structural descriptions* of complex objects depicted in complex pictures: e.g. *"Image regions A, B and C, represent surfaces meeting at convex edges, between A and B, between B and C, and between C and A, and all three edges meet at a common vertex V"*. (A formal notation was used, not English expressions.) That ability to describe 3D scenes depended on a prior ability to produce descriptions of the 2D structure of images, e.g. in terms of lines, junctions and regions bounded by lines. Work in linguistics had shown how *syntactic* structures of verbal expressions related to *semantic* structures of expressed meanings. In Clowes' work, syntactic structures in linguistic expressions were replaced by points, lines and regions defining the structures of (certain classes of) images and the semantic information conveyed by sentences and their parts corresponded to the 3D information derived from images and their parts, unlike the *pattern recognition* paradigm that simply segmented a complex image into parts and attached labels to the parts. I found his work very interesting, and we had discussions alone and with other AI researchers and psychologists. I attended his classes in AI and programming, and learnt to write some toy programs, initially using Algol60, I think, around 1969.

I read books and papers under Max's direction and discussed them with him, including Minsky's mind-blowing progress report on AI (Minsky, 1963), and an important new article by McCarthy and Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence" (McCarthy & Hayes, 1969)[2]. It was a deep and influential contribution to the methodology of artificial intelligence and, as it claimed, had philosophical implications. I agreed with some of it but not all. In particular, they claimed that logic, was, amongst other things, a *heuristically adequate* form of representation for an intelligent reasoner, *i.e.* it provides a framework supporting efficient searches for solutions to problems. However, from my personal experience of solving problems in geometry—one of my favourite subjects in school and at university (my first degree was maths and physics before I switched from mathematics to philosophy as a graduate student)—I knew that if I had been presented problems in Euclidean geometry in a logical form and if I was allowed to think about them only by manipulating axioms and logical rules, then I would never have been able to solve the problems in the sort of time frame that I did when I was using diagrams either on paper or in my mind, when reasoning about spatial structures. So I wrote a paper acknowledging that despite the great power of logic, for heuristic power one often needs to supplement it with other things, e.g. diagrams, and I tried to characterise the differences. That short paper was presented at the International Joint Conference in AI (IJCAI) in 1971 (Sloman, 1971)[3]. Bernard Meltzer, the founding editor of the new journal *Artificial Intelligence*, asked if he could publish it, and I agreed. It was also re-published in a book (Nicholas, 1977).

He then invited me to spend a year at Edinburgh University, one of the leading AI centres in the world at that time, with four AI groups straddling robotics, vision, language, AI in education, computational neuroscience, and computational logic. I learnt from all of them, and had my brain re-wired! I was mainly located in the Department of Computational Logic, headed by Meltzer, where I read books and papers on computer science and AI, learnt to program in POP-2, Lisp, and Prolog, and met outstanding visiting and local AI researchers (including Geoffrey Hinton, then a PhD student, who later worked with me at Sussex). After I returned to Sussex, in a highly interdisciplinary group including Margaret Boden, I was

[2]Now available at `http://jmc.stanford.edu/articles/mcchay69.html`
[3]Online at `https://www.ijcai.org/proceedings/1971` pages 270-278

able, to teach and do research in AI and its links to philosophy[4] and collaborative research on vision (Sloman, Owen, Hinton, & O'Gorman, 1978). The main driver for these developments was Max Clowes. Unfortunately, he died quite young, around 1980.[5]

KI: You mentioned that you looked at geometry problems in the context of scene understanding and that those got you very interested in AI research. Did these spatial reasoning problems also get you interested in robotics?

My interest in AI was not at first specifically related to robots or applied artificial intelligence, although I interacted with the robotics group during my year in Edinburgh. My original interest in AI came from an attempt in my 1962 DPhil thesis (Sloman, 1962) to defend Immanuel Kant's *Philosophy of Mathematics* (Kant, 1781). Kant had claimed that mathematical knowledge did not fit into the two categories that earlier philosophers, in particular David Hume, another great philosopher, had used for types of knowledge. One of the categories (according to Kant) is *apriori* (but not necessarily innate) knowledge, which you can get by thinking about what's impossible or necessarily the case. That is, you can be sure of some truths without examining things in the world, e.g. searching for counter-examples, whereas *empirical* knowledge requires investigation of things in the world, which normally cannot rule out the possibility of counter-examples turning up later, which frequently happens in science. My thesis defended Kant's use of three separate distinctions: *apriori/empirical*, *analytic/synthetic* and *necessary/contingent* (summarised in (Sloman, 1965)), arguing that mathematical knowledge was in the first category in each distinction. By that time most philosophers thought that Kant had been refuted by Einstein's work on general relativity, arguing that physical space is not Euclidean, as confirmed by Eddington's observation of the solar eclipse in 1919. Hempel's critique of Kant was especially clear and forceful (Hempel, 1945). My thesis defended Kant against those criticisms, (a) by pointing out that Kant's claims had not implied that human mathematical reasoning was infallible (the work of Lakatos in (Lakatos, 1976) documents mistakes made even by great mathematicians) and (b) by pointing out that Einstein's work had not undermined the use of non-logical forms of geometrical reasoning by great ancient mathematicians.

Before Kant, David Hume claimed (a) that all *apriori* knowledge of *necessary* truths was (in Kant's terminology) *analytic*, i.e. merely concerned with defining relations between our ideas and their logical consequences: From definitions of our concepts we can use logic to derive consequences, without inspecting the world, and, according to the view Kant opposed, that's the only way to acquire *apriori* knowledge of *necessary* truth. Hume's second claim, (b) was that all other truths must be empirical: if you can't prove a proposition using only definitions and logic, you have to study the world, observe and measure things, and check that anything you discover also works at high altitude, or high speed, and on different planets and so on. (Apologies to Hume-experts: I am simplifying in order to explain Kant's views.)

Kant responded: No, there is something distinct from Hume's cases (a) and (b)—sometimes referred to as *Hume's fork*. I.e. you discover some things not by using definitions and logical reasoning from definitions, nor by inspecting the world in sufficient detail to cover all cases, but by doing something else—which he wasn't able to describe very precisely, though he gave examples from arithmetic and geometry and some examples of causal reasoning, e.g. if you reverse your direction of motion through a fixed environment the order in which you encounter objects or locations must be reversed and spatial discoveries, e.g. two straight lines in a plane surface cannot enclose a finite space. (Think about how many planes are required bound a finite volume.)

Kant used words like *intuition* in contrast with *perception*, which can provide only empirical knowledge of contingent truths. Although Kant used the label *"apriori"* he was not claiming that we are born with apriori knowledge: it isn't necessarily innate, since you obviously need experience in order to develop the ability to make (non-empirical) mathematical discoveries.

According to Kant, those discoveries are not *derived from* experience though the insights are *awakened by* experience, using unknown innate mechanisms whose powers develop through interaction with the world. Kant thought that the mechanisms were very difficult to explain though I suspect he would have

[4]Discussed at length in *The Computer Revolution in Philosophy: Philosophy, science and models of mind* `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

[5]http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-clowestribute.html

tried to use AI, if he had been born two centuries later! By some process that is very hard to understand, by interacting with the world, humans use complex (still unknown) information processing mechanisms to make non-empirical discoveries about space and spatial relations of the kind that were expressed as axioms by Euclid. Further discoveries could be made by deriving consequences from those axioms, or discovering aspects of space not derivable from Euclid's axioms, e.g. properties of origami geometry. Such reasoning often used diagrams, not logic. Moreover the diagrams could be imagined: they did not have to be physical objects. (Kant is mentioned again below.)

After encountering AI, I hoped to support claims in my DPhil thesis by showing—with working robots—how Kant's claims might be demonstrated in a "baby" robot that grows up to be a mathematician, starting with abilities that nobody really understands yet, and gradually bootstrapping itself, and eventually discovering truths not in Hume's two categories of empirical and essentially logical or definitional (analytic) truths. That was, for a while, my main motivation in using AI, though I later became interested in using it to explain a wider class of phenomena, e.g. feelings, desires, values, emotions, consciousness, etc., based on information processing *architectures* for human-like minds. That later led to the *CogAff*—Cognition and Affect—project.[6]

Most AI researchers had other motivations, including building robots to solve practical problems, and I became interested in those problems also. In particular, during the 1970s, and beyond, there were AI projects in Edinburgh, Sussex and other places (including Birmingham to which I moved in 1991). Some of those projects aimed at getting robots to be able to manipulate things. I thought that that work could contribute to the problem of explaining how ancient mathematical discoveries in geometry were possible, although that was not the motivation of most other AI researchers working on giving robots spatial competences.

Robot Development & Design Principles

KI: When designing and developing a robot that gains experience and accumulates empirical knowledge by discovering things in the world, as you have described it, how important do you think is the actual physical embodiment of that robot? And to what extent can simulations be used for this?

I used to think all the real problems could be dealt with using computer simulation, since even continuous processes can be simulated as closely as required on digital computers—illustrated by many stunning computer-generated videos or musical performances. But understanding a class of processes is not the same as being able to simulate all members. Understanding includes knowing that certain describable structures and processes cannot exist, for example a non-self crossing polygon with more corners than sides. I now suspect humans and other animals have insights into spatial impossibilities and necessary connections that modern mathematics treats as logical consequences of axioms for geometry discoverable by AI theorem provers, but which ancient brains discovered using very different mechanisms that are not available in current theorem provers, but are used in development of young humans, chimpanzees, squirrels, crows and other intelligent animals.

Clearly, some aspects of human intelligence, e.g. logical and algebraic competences, can be directly implemented in working computer programs—e.g. competence at using truth tables to determine necessity or impossibility in propositional logic. But different, currently unknown, mechanisms are required for understanding and modelling the parallel but interacting processes of physical development and development of intelligent control in young animals, including control of changing combinations of forces, sizes and weights of body parts; or the bootstrapping of new layers of sophistication that relate parts of the individual to increasingly complex and remote parts of the environment (e.g. when climbing a tree or flying from the nest for the first time) where some changes in problem features are continuous (e.g. size) and others are discrete, e.g. reachability, fitting into.

[6] `http://www.cs.bham.ac.uk/research/projects/cogaff/`

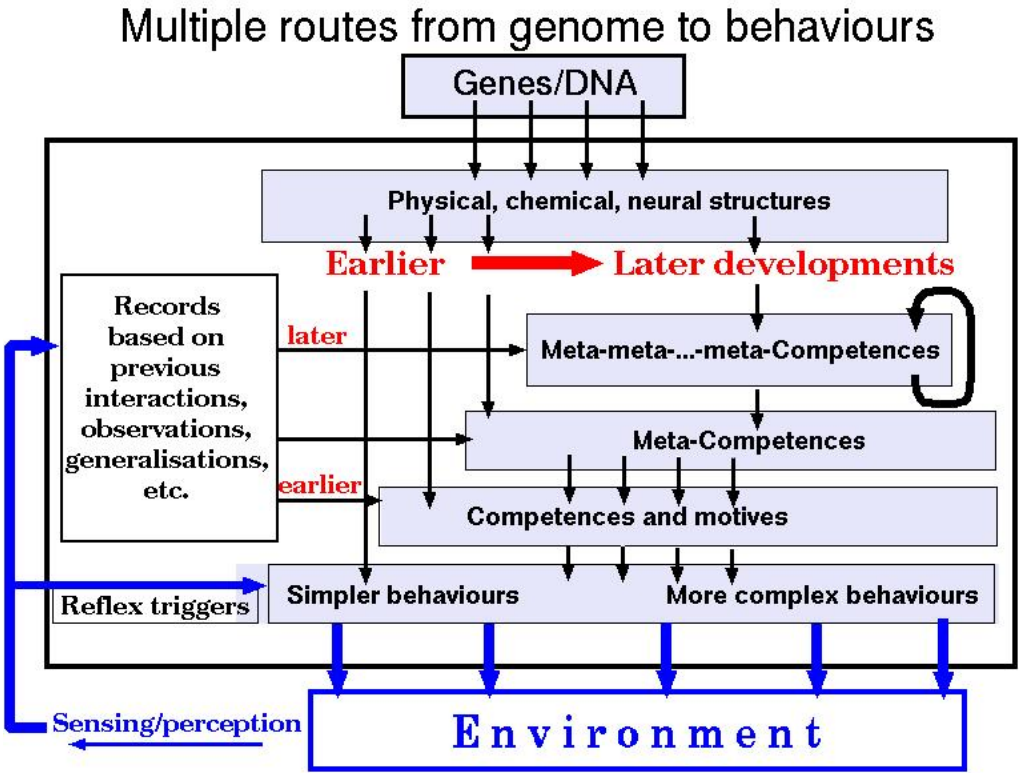

Figure 2: A Meta-Configured genome expresses different genes at different stages of development, where later gene-expression can be deeply influenced by results of earlier gene-expression.

KI: Could that be related to Kant's notion of how experience "awakens" knowledge?

Yes, I now think a (roughly) "Kant-like" theory is provided by the Meta-Configured Genome theory, developed with Jackie Chappell, summarised in Fig.2, showing multiple routes from the genome to competences, where later routes make use of information gained from earlier routes, so that "later" genes can have very different consequences in different environments which affect "earlier genes" differently.[7] Abstract, *parametrised*, competences based on previous evolutionary achievements are made available at various stages of development by a meta-configured genome. How such competences develop can depend on information acquired from the environment, e.g. during earlier stages of development. Instantiating a previously discovered (or inherited) powerful high-level (abstract) pattern is a very different process from searching for completely new detailed patterns in statistical data—like the difference between identifying a verb/object relation common to many languages and identifying the particular manifestations of that pattern in English, French, Urdu, etc. Insofar as the more general pattern is somehow in the human genome, it may have taken many lifetimes to get there: but individuals don't need to go through a process of discovering that pattern.

Moreover if a control pattern is parametrised, constant re-learning from scratch will not be necessary as body parts change size and muscles develop in strength: instead parameters may be adjusted in a principled way. However, continuous changes can also produce discontinuous changes, e.g. when growth makes some gaps impossible to pass through and makes some previously impossible reaching or stepping tasks possible. I suspect further research will show that that processes based on a Meta-Configured genome (Fig.2) can avoid huge amounts of trial and error learning (e.g. reinforcement learning) if at various stages previously dormant genetic mechanisms are triggered, deploying previously *evolved* abstract patterns

[7] More details available here `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-configured-genome.html`

instantiated using detailed information acquired by individuals, which may vary across individuals or across cultures, as illustrated by later stages of grammatical development in children. Similar processes of pattern *instantiation* rather than pattern *discovery* may affect many aspects of development in young humans, and other intelligent species, enormously speeding up development compared with purely bottom up pattern learning. (Readers familiar with the work of theoretical biologist Brian Goodwin (1931-2009) may recognise echoes of his ideas here.) All of this illustrates a surprising point: evolution can make mathematical discoveries used in meta-configured genomes, that products of evolution can make use of without being mathematicians.

Recognition and use of instances of genetically acquired abstract process-patterns is very different from searching for statistical relationships, as proposed in current neurally inspired theories of learning. There are also differences between (a) using abstractions based on relations between numerical measures and (b) using abstractions based on geometric and topological relations, e.g. containment, connectedness, topological equivalence. Further, the practical power of some patterns depends on the *necessity* of certain features that cannot be established empirically, for example the necessary transitivity of 1-1 correspondence, which is central to the use of counting and fundamental for understanding the natural numbers. If the importance of that feature is somehow "discovered" by evolutionary processes, genetic mechanisms can transmit abilities to use those features, even though no form of empirical learning could establish their necessary truth (as Kant pointed out). Partial orderings can also have mathematical consequences.

KI: Do you think that this kind of information processing could be realised using modern computers or do we need other forms of computation?

It is not clear that all the important details of such processes can be modelled in programs running on digital computers, because many of the patterns understood by human-like (or ape-like, squirrel-like, crow-like) learners are spatial (geometrical and topological) not numerical or logical. There may be hybrid discrete+ continuous brain mechanisms (e.g. chemical mechanisms) that are required for types of reasoning and discovery that cannot be implemented in digital circuits. Although digital computers have been used for geometry theorem provers based on Hilbert's logicist version of Euclid's axioms, those theorem provers don't *discover* the axioms as ancient mathematicians did. They have to be provided by human programmers. Moreover, not all ancient geometrical discoveries were derivable from Euclid's axioms, e.g. trisection of an arbitrary angle is impossible using Euclid's constructions, yet the *neusis* construction, known to ancient mathematicians, makes it easy.[8] Human visual systems don't seem to be capable of measuring exact sizes and distances, although they can become more precise with practice. I suggest that most of the time they use *partial orderings* (e.g. X is closer/wider/taller... etc. than Y) and changes in partial orderings and visibility produced by motion of viewer or other objects. (Gibson (1979) made a similar point in relation to texture gradients and optical flow gradients.)

Perhaps the brain mechanisms underlying ancient discoveries in geometry and topology start from spatial representations using only topological relations (e.g. containment, overlap, continuity) and partial orderings of sizes, distances and angles, not absolute values, and derive increasing precision by using sequences of comparisons. But even without high precision many practical control problems can be solved in topological and semi-metrical spaces, e.g. ruling an option out as impossible (getting an elephant through a doorway), analogous to ancient mathematical uses of diagrams that did not require absolute precision — and without the kinds of searches in spaces of logical and equational structures required by logic-based theorem provers. At the very least, the use of "analogical" forms of representation (Sloman, 1971) can provide important advantages in *heuristic* power, but they may do much more than that.

Modern computers are often used to *depict* complex continuous spatial processes on a screen, but *understanding* them (e.g. deciding whether a certain type of transition is possible or not) is harder to implement—e.g. grasping that no kind of rotation or translation of the elephant can make a difference if it is too big to fit through the doorway. Trying to join three rods to make a triangle with ends of rods meeting at the corners is easy in some cases and impossible in others (e.g. think of the length of the longest rod). What sort of machine can detect the impossibility independently of trying vast numbers of sizes

[8] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/trisect.html`

and locations of the rods, and without presuming the representation of Euclidean geometry in terms of numerical coordinates?[9]

KI: How could these discovery mechanisms be implemented in future robots?

If currently emerging ideas about the importance of sub-neural chemistry, mentioned below, turn out to be correct then it may be impossible to build an electronic digital computer with sufficient computing power to emulate a mammal or bird brain, e.g. because (a) chemical processes combine discrete and continuous change in ways that cannot be replaced (for all purposes) by purely digital mechanisms, (b) vast numbers of interacting molecules require far less space and energy than digital simulations of such interactions, and possibly (c): even if (a) and (b) can be overcome in principle, vastly more energy may be required for electrically powered digital equivalents to the sub-neural chemical mechanisms. Future research may show that the *only* way to implement a mechanism with the computational power and energy requirements of a typical bird or mammal brain is to replicate the biological mechanisms (Newport, 2015). The long term implications for AI may be unsettling.

Although Kant could not have thought of all these points, I think his notion of knowledge "awakened by" but not "derived from" experience can be reinterpreted as referring to evolutionary mechanisms that produce "gappy" (parameterised) information structures in the genome, at different levels of abstraction, expressed at different stages during development, which combine with information acquired during earlier interactions with the environment produced by previous stages of gene expression—a process depicted schematically in Fig.2. So genes expressed later—awakened by a combination of cumulative experiences and timing mechanisms, but not derived from experiences—use "parameters" previously acquired from the environment during earlier gene expression. Various stages of language development seem to be the most obvious examples of such multi-layered gene expression, starting with "babbling" shaped by the environment and used by genes expressed later that employ babbling competences to form a toolkit for producing and recognising short (proto-)verbal communications... and so on.

KI: If much information acquired from early behaviours is primarily of use during later gene expression, how will young learners be motivated to perform those behaviours?

That's a very important question, and I think the answer contradicts the widely held belief that all motivation must be based on expected reward (including avoidance of expected harm). An alternative explanation is that some behaviour triggering mechanisms were selected by evolution simply because they can sometimes produce information that is not immediately relevant but can be stored for possible later use, without the agent knowing that this is happening. Examples may be "playful" reflex behaviours, including "mock fighting" behaviours and exploratory behaviours in young animals. Information required for use by later genes can be obtained from the environment if earlier genes produce mechanisms that trigger reflex explorations and activities in the environment that collect information, or which trigger motives whose function is to acquire information from the environment and to develop competences that can be used *as parameters* during later stages of development, i.e. as components of more complex actions. So an infant's motivation to grasp something or put it in its mouth or throw it away need not be based on any expected reward. Instead, if the brain has appropriate mechanisms, the detection of an opportunity (of the right type) may directly trigger an unmotivated reflex action that has no further purpose as far as the child is concerned, but whose effects sometimes provide important new information that can be stored and used at some future date, or can develop a skill, or strengthen muscles, without that being a motive. The motivation to perform the action is *architecture*-based, not *reward-based*.[10] When the action is performed and information is acquired the child cannot know what future rewards can result. The left-to-right arrows in Fig.2 show crudely how information acquired by products of earlier gene-expression

[9] Many examples are discussed in this paper and papers it references: `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/impossible.html`

[10] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/architecture-based-motivation.html`

can provide parameters for genetic mechanisms expressed later. Some "reflexes" can simultaneously serve immediate practical needs (e.g. getting food, or attention from parents) and also provide information that is stored and later used to fill gaps (i.e. provide parameters) in later gene expression—something close to this was demonstrated by Ugur(2010).

When a new genome-layer is expressed in development the new mechanisms can "inspect" previously created structures and select some for use as parameters in new more complex utterances, theories, reasoning processes, actions, ... etc., combining previous discoveries to create something new and more powerful by instantiating previously evolved more abstract patterns, using products of late gene expression.[11] This is very different from multi-layered learning mechanisms that repeatedly look for statistical regularities, compute probabilities and attempt to maximise expected rewards.

KI: You said earlier that language development is an obvious example for multi-layered information processing that you are describing. Can you elaborate on this?

Language development clearly uses something like the Meta-configured genome mechanisms. Unless I've misunderstood him, Noam Chomsky seems to have thought that this sort of multi-layer genome-driven development was unique to evolution and development of communicative language, where the features of the mechanism are most evident, partly because young children go through well-documented layered phases of language development that are visible to non-linguists.[12] E.g., at an early stage, before they talk, infants do a lot of babbling. And when they get their babbling more or less under control, the sounds produced vary according to which language is in use in the child's environment. Deaf children may babble with sign language. The babbling competence is used by later stages of gene expression, where the babbling competences are recruited to form more complex uttering and recognising competences in producing and understanding verbal utterances that are structured, usable, and systematic. Moreover they can also be used as a basis for further development—during later gene expression, e.g. when infants start using previously acquired competences to construct words and simple sentences, followed later by phrases and clauses in more complex sentences with more varied functions.

For example, the ability to understand and use counterfactual conditionals develops later than abilities to understand and use much simpler constructs (Beck, Robinson, Carroll, & Apperly, 2006). Story generation, and theory construction, build on that. Likewise simple process-controlling competences e.g. moving hands, mouth, tongue, etc. are followed at later stages of development by mechanical disassembly and assembly process requiring far greater cognitive powers.

Parallel developments in sensing and perception mechanisms allow individuals to understand longer, more complex, utterances than the ones they produce. That usually feeds into development of more complex linguistic production competences, enabling further linguistic and cognitive development.

These are all aspects of the functions of a meta-configured genome, as a result of which layered competences in brains develop under the control of the genome, influenced by the environment, as indicated in Fig.2. Some genetic mechanisms lie dormant, waiting for earlier abilities to achieve a certain level of competence. Then the genome produces several staggered layers of new mechanisms where each new layer checks what has been achieved by earlier developments, and starts to re-organise what has been learnt, imposing a new structure. One symptom of that is that children start using grammatical constructions in a more regular way. (Unfortunately, I can give English speaker examples only.) Previously a child will learn that you can say "Johnny hit me and I ran away". After a later stage of gene-expression, the same child will say "Johnny hitted me and I runned away" because the higher-order mechanism that examines what has been learnt previously has found a regularity about past tense, namely, add the *-ed* sound on the end of a verb, and that dominates older habits. At first that discovery overrides previously learnt constructs. Later on the grasp of syntactic constructions becomes more sophisticated and the child will (unwittingly) learn that there are special cases, e.g. verbs that are exceptions to the previously learnt rules: e.g. "hit" is an exception to the general rule about past tense formation. These layers of re-organisation are not simply

[11] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/compositionality.html` is an incomplete attempt to develop further aspects of this idea.

[12] Many years ago I wrote to him suggesting that what he called "Cartesian Linguistics" should be called Kantian linguistics ... but to no avail.

based on general-purpose learning mechanisms: they are specific products of the evolutionary history of the species.

When a newly constructed layer of mechanism takes control, children may start saying things that are ungrammatical; including things they have never heard anyone else say (e.g. "... hitted me"), and no amount of verbal correction will change them—for a time. After a while, yet another brain mechanisms seems to grow, or get switched on, able to ask *How do the things developed so far fit all the previously and recently acquired empirical information?* It does an amazing software engineering job: equivalent to examining the most important regularities that it has found so far and somehow finding that some of those regularities have regular exceptions (as opposed to occasional slips). On that basis it creates *exception handlers*, and the child goes back to saying, "Yesterday Johnny hit me and I ran away" instead of "Yesterday Johnny hitted me and I runned away", as well as things like "Johnny kicked me so I punched him", i.e. coping with both regular and irregular verbs. I don't know whether this is done by having some genetic specification for forms of exception to linguistic rules, or by using a generic mechanism that allows empirically well established regularities to have exceptions, provided that the exceptions occur with some minimal frequency and regularity, e.g. strange outliers like "slay", "slew", "slain" in English. However, I doubt that mere frequency is the trigger. (Linguists may know more about this than I do.)

In effect, these mechanisms involve several layers of *software engineering* achieved by a mixture of evolutionary and developmental processes. They could not simply be products of a fixed general learning process that does all the work. In humans different genetic layers perform different tasks. The grammatical "over- generalisation and correction" mechanism needs to work differently in different languages, so the corrections cannot all be pre-specified in the genome: instead, later processes of gene expression produce changes that make use of information acquired during earlier gene expression processes, providing required "parameters". This is closely related to proposals about "representational redescription" by Karmiloff-Smith (1990, 1992).

Later developments in linguistic competence are not the same in all languages: the more sophisticated competences build on language-specific products of earlier development, in language-specific ways. So the "meta-configured" genome mechanisms allow later layers of gene-expression to modify products of earlier layers in a context-sensitive (e.g. culture sensitive) manner.[13]

As Karmiloff-Smith points out, these multi-layered context-sensitive gene-expression mechanisms are not restricted to mechanisms for language: her 1990 study involved picture generation competences. I suggest that similar layered genetic mechanisms underpin a great deal of human learning and development, including growing abilities to synthesise complex actions (e.g. screw motions of a hand), but many of the layers of competence produced are not as visible as those in spoken language. Perceptual competences, reasoning competences, learning competences, and action-preparation competences are much harder to observe, as their immediate products are mainly internal. She wrote (1990) "In my view, situating sequential constraints within this broader context of multiple representational levels, and linking redescription processes to cognitive flexibility and ultimate conscious access, offers a new, deeper account of developmental change."

I think her mentor Jean Piaget noticed related features of human development. He called himself a *genetic epistemologist*. He had read Kant, Frege and Russell, and knew about logic, but he didn't know anything about computing until very late in life. So, in trying to develop theories that capture some of these ideas, he collected a lot of useful data, but he lacked adequate theories about mechanisms. Late in life, he recognised that he should have learnt to program, but he was in his 80s by then, and he died soon after.

Returning to the genome: my ideas about its "multi-layered" structure and multi-stage gene expression increasingly influenced by "parameters" of varying types (i.e. mostly non-numerical) came partly from research on language development and, to a lesser extent, from psychological studies of non-linguistic development. Some of the key ideas emerged in discussions with biologist Jackie Chappell (begun 2004, leading to (Chappell & Sloman, 2007)) and are still being developed. Her work is concerned to a large extent with non-human cognition, e.g. in birds and apes.

My own observations of intelligence in young children, in squirrels, in nest-building birds, as well as first-hand experience of mathematical discovery, led me to think that future AI systems will also need to have different genetic layers and developmental processes, but I don't think any current AI systems have

[13] Please note: these ideas are still under development—corrections and suggestions welcome.

them, and I don't think neuroscientists or psychologists know how they work. My attempts to persuade robotics colleagues to help in development and testing of these ideas failed partly because the change would have disrupted their existing research too much.

Integration of AI & Robotics

KI: That brings me to my next question. Do you think then that we have the right approach to develop robots? I mean, we often take the most advanced methods in sensing, planning, reasoning, acting, language processing and so on and we try to integrate them into a single system. Should we instead rather adopt a more developmental approach and learn things incrementally?

Yes, but that will not guarantee progress. Ancient astronomers took the most advanced methods available at the time and came up with the Ptolemaic theory, according to which stars and planets rotate around the Earth. Collection of increasingly varied kinds and amounts of information over many centuries produced a succession of changes (a) regarding what needed to be explained and (b) regarding good conceptual frameworks and tools for formulating explanatory theories.

I think the currently most popular tools in AI have much in common with the modes of thinking that produced the Ptolemaic theory: they share the goal of trying to characterise observed patterns in a general way. But some great scientists and philosophers have recognised the need for deeper modes of thinking that go beyond finding economical ways to summarise observations, including Immanuel Kant, though he too was limited by his time. The study of intelligence requires deep explanatory mechanisms with generative powers, as much as physics, chemistry and biology do.

Since its birth around 1956 AI has involved a number of different approaches and different subsets have been fashionable at different times. Approaches explored by about 1962 were summarised by Minsky in (Minsky, 1963). As mentioned above, by 1969, when I met Clowes, there were two main competing strands of research on vision, one called *AI*, the other called *Pattern Recognition*. The first was mostly based on logic, logic-like reasoning, symbols, rules, operations on rules, and in some cases grammars (Kaneff, 1970). In contrast, research in pattern recognition attempted to use statistical analysis of sample data to train programs to attach labels to parts of images, including groups of labelled parts. For instance, Azriel Rosenfeld at Maryland University, was one of the leaders around that time. There was a deep clash between the sort of thing he was doing and the sort of thing that Max Clowes was doing; which tried to derive *descriptive relational structures* from images, not merely segmentation and labelling. Clowes (and some others) used *reasoning* about the structures and relationships to select between alternative *descriptions* of complex objects (e.g. "Regions A, B and C, represent surfaces separated by convex edges meeting at a common point") not just labels (Clowes, 1971, 1973). That research strand attempted to find structures and relationships using algebraic, geometric, and logical reasoning, while the pattern recognition approach collected lots of data, trying to find correlations relating labelled parts of images, suggesting labels for larger parts sharing features. Those two strands developed in parallel with different groups and subgroups dominating AI research activities at different times.

At present, probabilistic and statistics-based learning mechanisms have produced surprisingly many and varied useful results due to advances in the algorithms and both advances in computer power and the vast amounts of data that have been collected and made rapidly available. Available computer power has expanded enormously over the last 20 years. Using that power, we can make limited but sometimes very useful progress using current techniques for solving circumscribed problems. For example, if you want just to be able to answer questions that are no longer than some number of words and the syntactic structures are fairly shallow, then it is often possible to do useful things by training the system because the space of possibilities is not too large to be fairly well covered by available data.

On the other hand, if someone produces a sentence that starts by referring to ancient mathematicians and the theorems that they discovered and proved using geometrical and topological reasoning, and then discusses the scope and prospects for artificial intelligence, including comparing some current limitations of AI and neuroscience with what can be achieved by those earlier discovery processes, as in the sentence

you are reading now, then I do not think that any feasible amount of statistical training of an AI understander will be able to cover the space of structures that can be produced or encountered because that space explodes exponentially as significant sentence fragments become longer and more nested, with cross-references, as illustrated by this sentence.[14]

We need to combine different AI research strands. In particular, in support of *scientific* AI research into modelling/replicating natural intelligence, we need to find out what has to be in our machines initially. E.g., depending on the type of natural intelligence under investigation that will require trying to understand what evolution has put into human genomes, squirrel genomes, crow genomes, elephant genomes (etc.) that we don't yet have in our robots. I suggest that that will require understanding what the billions of sub-neural molecules are doing in each synapse! Their functions are largely ignored by most neuroscientists and AI researchers. One of the exceptions is the neuroscientist Seth Grant (Grant, 2018). See also (Trettenbrein, 2016). (Molecular computations are mentioned again below).

KI: What would be a good way to integrate these different methods? On one side, we have data-driven approaches, mainly driven by deep learning algorithms, big data, and also lots of computing power. And on the other side, we have model-based approaches, mainly based on logic. So how can we bring these two strands together?

Answering this requires a deep analysis of what each research strand can and cannot achieve. I now think something different from both is needed. For a while, I thought integration could be achieved by bringing people with expertise in the two dominant AI strands to work together on practical problems, including trying to design robots with a mixture of capabilities, for example, some of the sorts of things that young children and other animals, e.g. squirrels, can do. By analysing the problems and trying to explain the inadequacies in current models and theories, I hoped to provide new ideas about how programs, programming languages, and physical computation mechanisms need to be extended.

Analysing difficult cases led to the conjecture that bridging the deepest gaps between natural and artificial intelligence required new ideas about computation. That was triggered by being asked to comment on one of Alan Turing's papers during the Turing Centenary year (2012)[15]. In 2013 Elsevier published a collection of papers by Turing with invited commentaries (Cooper & van Leeuwen, 2013). Barry Cooper, the main editor, invited me to write short comments on various aspects of Turing's work, including the *wrongly labelled* "Turing test", and also Turing's long 1952 paper on the chemical basis of morphogenesis (Turing, 1952). (The best introduction for non-mathematicians is Philip Ball's summary (Ball, 2015).) Hardly anyone working in AI, computer science, philosophy, psychology or neuroscience has read it, but it is now his most cited paper—e.g. by physicists, chemists, and mathematicians. It discusses how two chemicals diffusing through a liquid at different rates, can, under certain conditions produce a wide variety of 2D patterns, for example, surface features like blobs, lines, spirals, dots and groups of dots. (I have not taken in all the mathematical details!)

I wondered why he was writing about that two years after publication of his Imitation Game proposal (which I think is widely misconstrued as a *test* for intelligence, but that's another topic). This led me to the conjecture that Turing was looking for ways to extend ideas about computation e.g. using mechanisms with a combination of continuous and discrete state changes. In his 1950 paper he wrote: *"In the nervous system chemical phenomena are at least as important as electrical"*, without explaining why. I suspected he was already thinking about the 1952 paper on chemistry-based morphogenesis although he did not explicitly link the two. Modern computers make use of discrete changes in bit patterns, implemented in transistors, though use has been made of analog (continuous) computation since antiquity, including sundials and mechanical models of the solar system (orreries). If brains combine discrete and continuous mechanisms in a deeply integrated way, that could involve chemistry because chemical processes intrinsically include continuous changes, e.g. when molecules fold, twist, come together or move apart, and discrete changes, e.g. when chemical bonds are formed or broken, using catalytic mechanisms that require very little energy

[14] Giving the above sentence to google-translate and asking it to translate into a few other languages and back to English, produces evidence of incomprehension that seems to vary between target languages. The Winograd Schema challenge is closely related to this point: `https://en.wikipedia.org/wiki/Winograd_Schema_Challenge`

[15] Which included the Alan Turing Centenary Conference, Manchester UK, June 22–25, 2012

and occur very fast.

In "What is life?" in 1944 (Schrödinger, 1944), Schrödinger drew attention to ways in which quantum physics provides types of discrete state-change (in chemical bonds) that Newtonian physics could not support, and showed how this could explain otherwise mysterious aspects of biological reproduction, including reliable inheritance of features across multiple generations. I don't know whether Alan Turing ever read Schrödinger's book, but it was highly influential and he was thinking about closely related topics in the 1950s; so it is possible that he read it, but he didn't mention Schrödinger in the 1952 paper.

Anyway, I suspected Turing was trying to come up with some new form of chemistry-based computation combining continuous and discrete processes so I began trying to extend his ideas, and continued collecting examples of spatial reasoning, in the hope of finding important clues about required mechanisms. It wasn't till 2018 that I was given a clue supporting my guess about Turing's motivation, namely: in his thesis he had distinguished mathematical *intuition* and, mathematical *ingenuity* and claimed that computers (e.g. Turing machines) were not capable of mathematical intuition, only mathematical ingenuity. I don't know whether he ever attempted to define those terms or defend his claim. I have not yet found out whether he knew about Kant's claims regarding mathematical (e.g. geometrical or topological) intuition, or whether he came up with related ideas completely independently, which is to be expected of great thinkers!

Thinking about how Turing might have extended the ideas led me to propose a project seeking evolutionary transitions in types of biological information processing, using many kinds of evidence about the varieties of organism on this planet, from the very simplest forms of proto-life to what exists now. Ideas about very early uses of information processing must involve guesswork though for later organisms we may find hints in fossil records. Instead of looking only at physical structures, morphology, and behaviours, which we can sometimes observe and sometimes not, we can also try to find out what their *information requirements* were and what mechanisms could have performed those functions: very difficult tasks because information processing doesn't leave fossils, though many of its effects do.

That was the genesis of the project of trying to use *intelligent guesswork* (disciplined conjectural thinking) to understand intermediate forms of information processing, by investigating information processing requirements of different sizes and shapes of life forms in different environments, with different physical abilities to move, manipulate, avoid, consume, etc., and different sensing apparatus and so on, and different likely needs. I call this the *Meta-Morphogenesis (or M-M) project*[16], which is what I suspect Turing might have worked on if he had lived another 40 years. I am not as clever as Turing, but a lot has already come out of that idea since 2012. It led to a totally different approach from trying to test ideas simply by building human-like machines. The ideas need to be tested in terms of what they are capable of explaining, which must be far more general than the competences of any one human, and many of them will be equally relevant to information processing needs of other organisms, e.g. all vertebrates, or even older evolutionary precursors of humans, or other biological branches (e.g. spiders, octopuses, ...).

Some (perhaps most) of the discoveries should be testable using evidence of previously undiscovered evolved virtual machinery in various organisms. An important feature of many running virtual machines is that although they happen to be implemented in physical processes there is no way of translating descriptions of their information processing states or processes into descriptions of states and processes of the physical machinery in which they happen to be implemented.

For example, *seeking an explanation for a murder* is a process that can involve many physical objects in different locations, evidence gained from many people, various kinds of forensic testing (some of which may be invented during or because of this investigation), and all sorts of historical records; and which could have involved many different details if some of the background facts had been different. There is no hope of translating "X is seeking an explanation for Y" into a well defined disjunction of descriptions of physical states and processes going on in X or X and X's environment. Likewise descriptions of the computational processes in a desktop PC cannot be translated into a description of physical processes underlying those computational processes. A sequence of software upgrades while the system is running could significantly alter what memory locations are used when and what they are used for, without altering the virtual machine states, only their physical implementations.

Moreover, many of the processes, for example email processes, or internet web searches, will use a

[16] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`

massive, complex, dynamically changing network of external computational and physical resources. The terms used to describe what's going on are not *definable* in terms of the physical resources, for many of them could have changed (e.g. network routes switched) while the system was running, even if they did not actually change. In some cases the physical changes will introduce new technology never considered by the designers of the original system—as has constantly happened to the internet, and even to much of the software on personal computers. The set of actual email accounts, their names, their physical implications will not be fixed. The numbers have been steadily growing for decades, as numbers of users increase.

A strand of the M-M project is collecting direct or indirect evidence of past transitions in information processing *requirements* and corresponding transitions in information processing *mechanisms*. For very early (proto-)organisms this is unavoidably speculative—yet the speculation is disciplined if based on evidence about varieties of extinct species and environments in which they lived, as well as evidence about the huge variety of species and biological environments and mechanisms that still exist on this planet, on all scales of size. There is also much direct and indirect evidence regarding changes in types of information required, and types of "construction kit"[17] required for creation and maintenance of organisms or their parts, or their construction kits, including construction-kits for building information processing mechanisms of many kinds—not just brains!

Testability of Robot Systems

KI: I see how your M-M project relates more to the developmental approach that we were talking about earlier.

However, coming back to what you said about creating robots for solving practical problems. Today, robots are often built for specific purposes and their performance, capabilities, and intelligence can be evaluated, for example, in international robot competitions. How do you think your approach can be evaluated? And what role, if any, does Turing's Imitation Game, that you mentioned, play for evaluating robot systems in general?

I don't recommend focusing only on testing a *particular* working system. We need to evaluate various *abstract designs* for *generic types* of "baby" robot capable of developing in many different ways, depending to a considerable extent on the opportunities provided by the environment during an extended process of development—as is the case with every new human, ape, crow, elephant, etc. (I've discovered that that idea is hard for some people to grasp.)

So the generative power of one design of that sort could include a wide variety of instances with very many differences of detail, partly dependent on environments, which can vary enormously, like the environments of humans across centuries, since they first evolved. So a test for a human-like genome, should not focus on a particular combination of competences but the genome's ability to produce a huge variety of different collections of competences in different environments, including different language competences, such as sign language competence in many intelligent humans born deaf, as well as powerful spoken and written languages, and technical extensions of those languages developed to meet new requirements, like the languages of science including chemistry, physics, biology and mathematics. But no individual human has all those competences!

I have argued elsewhere (as has the mathematician David Mumford[18]) that long before languages were used for communication there must have been powerful *internal*, structured, but not necessarily linear, languages for expressing the contents of perceptions, recorded memories, information about conspecifics and the local environment, intentions, plans, hypotheses, and questions. On the basis of those internal languages, sign languages later evolved for communication, initially to meet requirements for collaborative actions (e.g. moving or lifting heavy objects, or collaborative hunting). That's why every normal human is born with the ability to learn a rich sign language, even if that ability is never used. Spoken languages may

[17] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/construction-kits.html`
[18] `http://www.dam.brown.edu/people/mumford/blog/2016/grammar.html`

have taken over gradually because of their usefulness in more contexts, e.g. communicating in the dark, communicating around corners, or communicating while hands are occupied (e.g. in a shared lifting task).

There are lots of things that young children now do that no previous generation could ever do, e.g. playing with (electronic) tablets, moving things around on screens with their fingers and having conversations with remote individuals that they cannot see or hear directly, e.g. using email, phone, or video links.[19] Something in the human genome provides enormous powers to enable and enhance collaborative action, in combination with features of whatever environment the human develops in, which may itself include products of earlier human engineering. This may continue in the distant future, in ways that we cannot now imagine, just as people living 1000 years ago would have no conception of video conversations across oceans.

The point is: producing a robot that can perform a fixed set of tasks cannot answer the question *What enables humans to do that?*, because when humans do it they use mechanisms able to go far beyond what a typical current robot does. Those abilities are potentially there in the genome at birth, but the realisation process interacting with the individual's environment can both enormously extend each individual's abilities and also narrow down some of the original potential. For example, a normal human is born with the ability to develop any of thousands of spoken, signed, and written languages used on this planet, now and in the past. But, as the individual develops, that potential is narrowed by the fact that many of the abstract genes expressed at various stages after birth are instantiated to form specific collections of competences, thereby losing much of their potential to be instantiated in other ways. A human child can develop competence in a few languages in parallel, but not thousands.

Turing's "Imitation Game" focuses on a very narrow, technically convenient kind of process that you can use to test some machines. Even if we extend this to robots not just talking but also manipulating things, carrying out commands, or even teaching other robots —whatever we do— it may be possible to produce success in a particular type of task, but without replicating the full depth of the process by which humans arrive at similar capabilities. At every stage of learning, humans have the ability to go in many different directions, whereas current programmed robots can only do things in a restricted space. Those that learn can only learn in very limited ways—partly because of limitations of their physical designs, but mainly because of limitations of the theories behind their learning mechanisms.

KI: To follow up on this, is there then any test that would allow us to conclude that we have developed a system with capabilities that exceed those of humans, or which are at least comparable?

We already have machines that vastly exceed many of the capabilities of humans, including high precision numerical calculators, and search engines that can retrieve items from a collection of billions of items on the basis of a fragmentary key, which humans can do on a much smaller scale (e.g. if I give some people two words "Armstrong moon", or "turkey christmas"). But that associative ability is not the sort of spatial intelligence that enables a two year old child to play with toys, or a crow to build a nest that can safely hold eggs, or an adult elephant to help her baby out of a mud-hole.

Instead of testing a *particular design* for an intelligent robot, we need to test a *theory* characterising mechanisms that combine genome-driven development with multi-stage learning to produce intelligent machines in very varied (but not totally arbitrary) environments, with the results that we see in humans and many other intelligent animals. The meta-configured genome theory (Fig.2) illustrates this. Instead of proposing a general learning mechanism used from birth onwards, it postulates layers of increasingly abstract mechanism specifications, evolved at different times, and expressed at different stages of development, supporting new varieties of learning, exploration, and creativity, where each layer adds new ways of combining information acquired at an earlier stage of development with new information acquired while using the new genome layer — instantiating multiple schematic design patterns produced by evolutionary changes in ancestors, including ancestors in earlier species.

The requirement for a generic explanatory theory is fairly obvious as regards *physical* design patterns, e.g. designs shared across all vertebrates, or all vertebrates with limbs, or all mammals as noted by

[19] This possibility was anticipated in a short story "The machine stops" by the novelist E.M.Forster in 1909, `http://archive.ncsa.illinois.edu/prajlich/forster.html`.

D'Arcy Thompson (Thompson, 1917). We can generalise from physical design to evolved types of information processing. Variants of old design patterns may be shared across different species, e.g. early musculoskeletal control mechanisms, where control details vary both within individuals during growth and development of behaviours, and also across related species — implying an ancient requirement for parameterised specifications in the genome. Some parameter changes produce new species, whereas others support different stages of development within a species. The meta-configured genome idea assumes that multiple uses of parameters (or more precisely *parameter-gaps*) evolved both across generations and also for use during stages of individual development.

Since the basic mechanisms of reproduction and development are chemical, the natural implementation for parameters is a separation between fixed and variable components of complex molecules, where the variable components may at first be individual particles then structured molecules in later species. Molecular parameters allow far richer forms of variation than numerical parameters. So in a sense evolution anticipated Frege's discovery that arguments of functions are not necessarily only numbers but can also be functions and more complex structures (e.g. trees, grammars, axioms, etc.).

This requires a complex type of "cooperation" between evolution and development. Whereas standard forms of deep learning repeatedly use the same type of mechanism (assumed to be a product of evolution?) to add new levels of competence based on statistical records of sensor and motor signals acquired while performing actions, our proposal is that in more intelligent species later stages of development use different kinds of learning, evolved at different stages in the history of the species.

Note: The use of "whereas" in the previous sentence illustrates a type of thought content, or communicative content, that could only be produced at a relatively late stage of development, because it uses a new kind of ability (signalled by "whereas") to compare two previously understood thought contents. That new ability may not have existed in the earliest language users.[20] Perhaps it initially evolved as a variant of an ability to compare perceived structures, somehow encoded in the genome at a level of abstraction not restricted to any particular human language.

Another example of a capability that must have evolved relatively late is the ability to recognise varieties of *impossibility* and *necessity* in a space of useful structures: an essential feature of both intelligent choice of actions and mathematical discovery (as I believe Kant understood).

For any particular environment, there is a collection of competences and behaviours appropriate in that environment, which may be far less useful in another environment, as can be demonstrated by transplanting someone who has grown up in a modern city environment to a jungle environment with dangerous predators, or a steep rocky mountainside where a misplaced foot can be fatal, though local humans cope well (or *vice versa*). Likewise an expert speaker in one language community may be an incompetent communicator in another.

Such human forms of expertise are not simply products of a *general* learning mechanism provided in the individual from birth, if the meta-configured genome theory is correct. Instead there are many processes of learning in different generations, i.e. in *ancestors* of current individuals, repeatedly followed in later generations by genome changes combined with abstraction (provision for parameters).

In contrast, a typical current robot design that works well in a certain range of environments (like a "passive walker" robot able to walk down a variety of planar, and not too steep, slopes) does not start with the deeper potential that allows humans and many other intelligent animals to evolve and develop successfully in a much wider variety of physical (and in some cases social) environments. Compare the enormous variety of *manufactured* environments produced by humans over millennia.

Over time, humans constantly change their environments in ways that enable (and sometimes require) new developmental trajectories in children, whose novelty arises from combinations of old abstractions (in genomes) with new details during individual development. (This is closely related to Karl Popper's notion of knowledge stored in the environment: his "Third World" (Popper, 1972).)

A robot design that develops well in a particular modern factory environment using current AI learning and teaching mechanisms might fail miserably if transplanted (at any stage) to a stone-age environment, whereas a normal human child, transplanted early enough will do as well as other stone-age individuals.

My suggestion is not that robots with human-like intelligence should start with a monolithic totally

[20] Try giving the sentence "I like meat with gravy whereas I prefer fish with sauce" to a translation program, and then translate the result back to English.

general learning mechanism (as sought by many current AI researchers), but with something like the *layered* bootstrapping mechanism evolved for human life on this planet in the distant past, where each new layer, when activated, combines previously evolved abstract features that it provides, with parameters acquired during earlier layers of individual development and learning that can vary enormously across individuals and environments. The effects of this are most obvious in language development, but that's because the effects of external language use are so public. Parts of the multi-stage mechanism seem to be shared by other intelligent vertebrates, especially those with extended dependency on parents, e.g. orangutans. I don't know of invertebrate examples, but there may be some (octopuses?).

A multi-layered collection of genetic mechanisms for producing information processing capabilities could no more be replaced by a single general purpose genetic mechanism operative from birth (or earlier) than the multi-layered collection of mechanisms for producing and controlling various *physical* features during development could be replaced by a single general-purpose growth genome combined with an initial physical form.

Our scientific/philosophical task is to explain and model the huge potential to learn and develop that is produced by a human genome (or even a squirrel, crow, or ape genome). Explaining deep and varied context-sensitive developmental power is a much harder task than replicating a particular product of that power (which is also very hard in the case of human spatial reasoning capabilities). Simpler monolithic (i.e. not multi-stage) designs may be useful and interesting in various ways, but do not answer the deep scientific questions about the nature of human intelligence, or more generally, biological intelligence.

Real-world Applications and Future Challenges

KI: For many real-world applications it would be sufficient to develop a system that can just perform a certain set of tasks such as, for example, in autonomous driving. Or do you think that this problem is already too complex for any robotic system today?

We've learnt over several decades that there are some tasks at which robots excel. For instance, robots have been assembling cars, and other machines, where all the parts required and exactly what should be done with them is known in advance and where the precise requirements for assembly are the same for every instance of a particular part in a particular model. Robots have been doing such repetitive assembly far better than humans, i.e. more reliably, more quickly, and more cheaply since several decades ago, e.g. in car factories. I suspect robots functioning in highly controlled environments on specific tasks will continue to get better, performing more reliably, more quickly and more cheaply than humans.

But not all tasks have such controlled environments and unchanging requirements. Car-driving, for example, includes many different kinds of sub-problems that vary enormously across environments and change over time, e.g. as other road users change. Problems encountered on motorways, are much simpler and less varied than those likely to be encountered when driving in an old part of a congested European city with narrow roads. Even on a motorway, difficult situations can arise, e.g. in a car approaching a situation where there has been a crash, whose details may be unusual, or when an unexpected large animal turns up.

I expect we'll have vehicles that are computer-based and highly trained which, on motorways, will be more reliable than most (but not the most expert) humans. I say that from personal experience of having fallen asleep on a motorway and nearly killed myself. I was really lucky—a slight difference in the environment and I would have died in 2012. A fairly simple AI system, could have taken over after noticing that I was not in control, for example, by looking at my eyes, while detecting that the road ahead of me was curving around to the left whereas I was driving straight. Making a car follow the curve of the road (and its crash barriers), while gently waking the driver could be achieved in the near future. So, on motorways, I suspect there is lots of scope for AI-controlled vehicles, although there will always be situations that haven't been taken account of in the training, which the system cannot reason about and deal with adequately as a human would. But those cases may be sufficiently rare for the benefits on the whole to be worthwhile.

In congested urban environments, however, all sorts of things can happen that generate an explosive combination of possibilities involving things like wheelbarrows, bicycles, prams, rickshaws, animals and animal-drawn vehicles, someone playing a game with a new toy that happens to roll or fly across the road,

and so on. Human drivers don't always cope well in novel situations but their deep understanding of space, shape and motion can be used to meet novel challenges.

I don't think the ability to deal with a much richer variety of cases can be achieved simply by generalising from examples provided in training. The universe has lower level generative power that can always, except in tightly controlled contexts, provide novel configurations. Human drivers will also fail if facing a sufficiently fast moving novelty, e.g. a sink hole suddenly opening up, but the generative brain mechanisms that constantly allow humans to come up with new machines and gadgets, new works of art or musical compositions, new jokes, new stories, and new theories can also in many cases (but not all) generate new appropriate response in a dangerous novel encounter. There are individual differences however.

The generative power of the universe is not matched by the generative power of individual brains, but biological evolution has clearly produced brains whose creativity regularly copes with situations or problems not previously encountered by those brains or any other brains: the powers of a species like ours can keep growing, apparently indefinitely, even if each individual has limits.

Explaining the creativity and ingenuity of humans well enough to provide comparable capabilities in our robots may require mechanisms based on deep theoretical advances well beyond current AI and current neuroscience. Until then, human creativity and ingenuity will repeatedly outperform the best AI systems, although clever designers or trainers may enable some of the easier novel targets to be reached.

I hope it's clear why I won't trust my life to an AI-driven car in urban environments in the foreseeable future—not this century probably, especially if replicating brain functions requires understanding sub-neural chemistry as well as understanding the combinatorics of molecular processes that are vastly more combinatorially explosive than neuronal level computations in which all learning derives from changes in scalar values (measures, frequencies), not structures, e.g. grammatical or chemical structures. I believe, that von Neumann anticipated some of these ideas in his last book (von Neumann, 2012).

KI: So, when developing robots in these very complex scenarios, what do you think are the greatest challenges that we are facing? And what do you think should be the role of universities as companies are now taking a lead role in the development of AI and robotic systems; simply because they have more manpower . . . and more financial resources!

Perhaps my answer will surprise you. I think the main challenge is our educational system. It is not producing thinkers of the kind that are required for some of the hardest problems. There are various reasons for this. One is that the teaching of geometry as I learnt it as a child has changed. Bright school-children used to learn how to solve construction problems and prove theorems in Euclidean geometry using diagrams: it was a standard part of academic education. It was part of Immanuel Kant's education. Some of his views about the nature of human minds were based on that sort of background.

Now many school-leavers may have learnt a little logic, set theory and algebra, and perhaps learnt to reason formally from axioms expressed using logic, but they have learnt only very shallow subsets of geometry and topology. E.g., I meet graduates who tell me that at school they simply memorised facts, such as the triangle sum theorem, or Pythagoras' theorem, but have never learnt to find and check proofs. And if they do learn to prove theorems in geometry, for instance, at a university, they may only learn it in a logical framework, e.g. starting from something like Hilbert's axioms.

I did not learn logic and abstract set theory until I was a graduate student, but I don't think that hampered my early mathematical development, any more than not knowing such things hampered Archimedes, or Zeno. At school, before going to university, I benefited enormously from learning to find and check geometric proofs or constructions.

There are now impressive geometry theorem provers that start from logicised versions (or variants) of Euclid's axioms, and then use formal reasoning to produce conclusions, e.g. (Chou, Gao, & Zhang, 1994). But that's not what the ancient mathematicians did. For example, ancient mathematicians discovered the concepts and axioms for Euclidean geometry without deriving them from axioms! They also made discoveries that went beyond Euclid's axioms, such as the *neusis* construction that enables arbitrary angles to be trisected—impossible in Euclidean geometry.

However, no current AI geometry theorem provers that I know of can make such discoveries

because they can only start from (possibly extended versions of) Euclid's axioms and work out logical consequences. They can do logical and arithmetic reasoning but not spatial reasoning or make discoveries of the sorts that originally led to axiomatised geometry, including discoveries like the *neusis* construction,[21] which was known to ancient mathematicians (e.g. Archimedes) but was excluded from the teaching of Euclidean geometry, apparently on the grounds that it combines properties of straight edges with properties of compasses, which reduced the purity of geometry. But excluding it reduced the *power* of geometrical reasoners!

The point of all this is to indicate how the creative mathematical power of biological minds (a) exceeds the power of statistics/probability based learning systems, which cannot discover, represent or reason about impossibility and necessity, as Kant seems to have understood, long ago, and (b) exceeds the heuristic power of logic+algebra based formal systems in certain domains of reasoning and problem solving concerned with impossibility and necessity in spatial structures and processes — despite the fact that sophisticated logic-based theorem provers outperform all humans on certain tasks, just as computers have outperformed humans on arithmetic tasks, sorting tasks, searching tasks, and others, for several decades. But I am not aware of any computer based machine that can start with something like the knowledge at birth of a baby human and achieve the understanding of numbers of a six year old child.

The fact that the old powerful ways of thinking are no longer a standard part of education, will inevitably restrict the abilities of future AI researchers attempting to find ways to replicate human mathematical creativity. And that is likely to restrict the AI systems they develop.

Another factor relevant to progress on hard research problems, is the enormous growth of human knowledge that is now available to be learnt by potential researchers, who need a much extended education to provide the breadth and depth of understanding required for important advances. We still expect students to leave school at about 18, spend three or four years getting a first degree, and then after five or six years of post-graduate education to be ready to become successful researchers, as demonstrated by ability to produce highly cited publications and win grants. Instead, the most able potential researchers need to spend at least another decade after their first degree broadening and deepening their knowledge, including knowledge of the history and philosophy of science and mathematics to help develop their judgement. Highly accelerated pressures on young researchers to publish, get grants and attract citations (pressures I did not encounter as a young university lecturer) *seriously* interfere with the continued learning and development required to produce ground-breaking thinkers who can significantly advance human knowledge, though the current system may train humans to be machines for generating conference and journal papers in restricted domains, often based on groups that form efficient mutual citing communities.

I believe this educational system is inadequate to produce the kinds of researchers needed for the deepest and most difficult ground-breaking advances in knowledge, as opposed to fairly shallow extensions of current knowledge flooding journals and conferences. In the UK, the problem was hugely exacerbated by the decision around 1990 to abolish polytechnics, which were performing important educational and industrial/ commercial training functions by turning them all into universities, thereby seriously diluting resources for funding university-level research, and depriving the nation of an important post-school educational resource: its polytechnics!

On the whole, current educational systems tend not to produce graduate researchers and university lecturers with the kind of broad and deep education needed for them to perform the future-oriented functions of universities, including inspiring and guiding future ground-breaking researchers. In particular, research on understanding cognition in all its forms requires an education encompassing mathematics, chemistry, physics, biology, neuroscience, psychology, philosophy (e.g. philosophy of mathematics, of science, of language, of mind) as well as a broad and deep understanding of varieties of forms of computation and their strengths and weaknesses. A lot of the education should be project-based, but without pressure to get publications or high citation counts, as opposed to critical and constructive reviews by supervisors and peers.

We need more bright learners leaving school to be exposed to a variety of additional disciplines, learning to combine information of very different kinds when appropriate, and working on deep and difficult projects without pressure to publish and attract funds. That may be slowly happening to a very

[21] See `https://en.wikipedia.org/wiki/Neusis_construction` and `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/trisect.html`

small (lucky) subset of researchers. But it's not happening on a sufficiently wide scale, and I suspect it is not happening to enough people to generate the new thinkers who can come up with ideas that will enable us to make deep new advances and also educate the next generation to continue the process.

I was very lucky as a young graduate, because I was allowed to switch from mathematics to philosophy, and later, as a young lecturer, to switch from philosophy to AI, without anybody chasing me to get grants or to produce publications. It took me a long time. My undergraduate degree (in mathematics and physics) lasted from 1953 to 1956. As a graduate student (1957-62) I moved from mathematics to logic, to philosophy of mathematics, then became a lecturer in philosophy. Later, thanks to psychology seminars where I met Max Clowes, I encountered AI in 1969 and began to do theoretical work in AI in 1972. My first substantial AI development project between 1975 and 1978 (with David Owen, Frank O'Gorman and Geoffrey Hinton) tested ideas about vision, reported in (Sloman et al., 1978). So, between 1953 and 1978 I was lucky to experience an extended educational process in which I was mostly learning, including learning about philosophy, biology, psychology, and linguistics, then programming and AI. The learning continued long after that, and accelerated after I (formally) retired, around 2002. Far more young researchers should have that kind of breadth of education without pressures to produce anything in particular, except to go on learning, teaching, and demonstrating progress to peers and mentors, with cross-institutional reviews (but no league tables) to maintain standards. Such a culture could encourage experienced researchers to share very hard unsolved problems with younger colleagues (as Max Clowes did with me), with the possibility of triggering something new and deep, even if it takes far longer than the duration of a typical grant or a temporary research fellowship.

Deep new advances in knowledge may emerge that, despite astounding advances in technology and physical and biological sciences, our current system does not encourage, as indicated by the widespread neglect of Kant's deep ideas among researchers in AI, psychology and neuroscience. I wonder how many other cross-disciplinary bridges are waiting to be built that can support deep new advances. Is this already happening, without my knowing about it?

KI: Aaron, thanks again for taking the time and sharing your ideas and views with us. It was a great pleasure talking to you!

References

Ball, P. (2015). Forging patterns and making waves from biology to geology: a commentary on Turing (1952) 'The chemical basis of morphogenesis'. *Royal Society Philosophical Transactions B*. Retrieved from http://dx.doi.org/10.1098/rstb.2014.0218 (Interview: https://www.youtube.com/watch?v=6ed54_95kP4)

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006, March/April). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, *77*(2), 413–426.

Chappell, J., & Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. *International Journal of Unconventional Computing*, *3*(3), 211–239. Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#717

Chou, S., Gao, X., & Zhang, J. (1994). *Machine Proofs In Geometry: Automated Production of Readable Proofs for Geometry Theorems*. Singapore: World Scientific. Retrieved from http://www.mmrc.iss.ac.cn/~xgao/paper/book-area.pdf

Clowes, M. B. (1971). On seeing things. *Artificial Intelligence*, *2*(1), 79–116. Retrieved from http://dx.doi.org/10.1016/0004-3702(71)90005-1

Clowes, M. B. (1973). Man the creative machine: A perspective from Artificial Intelligence research. In J. Benthall (Ed.), *The Limits of Human Nature*. London: Allen Lane.

Cooper, S. B., & van Leeuwen, J. (Eds.). (2013). *Alan Turing - His Work and Impact*. Amsterdam: Elsevier.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Grant, S. G. N. (2018). Synapse molecular complexity and the plasticity behaviour problem. *Brain and Neuroscience Advances*, *2*, 1–7. Retrieved from https://doi.org/10.1177/2398212818810685

Hempel, C. G. (1945). Geometry and Empirical Science. *American Mathematical Monthly*, *52*. Retrieved from http://www.ditext.com/hempel/geo.html (Repr in Readings in Philosophical Analysis, ed. H. Feigl and W. Sellars, New York: Appleton-Century-Crofts, 1949)

Kaneff, S. (Ed.). (1970). *Picture language machines*. New York: Academic Press.

Kant, I. (1781). *Critique of pure reason*. London: Macmillan. (Translated (1929) by Norman Kemp Smith)

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*, 57–83. Retrieved from https://www.academia.edu/25128022/Constraints_on_representational_change_Evidence_from_childrens_drawing

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Lakatos, I. (1976). *Proofs and Refutations*. Cambridge, UK: Cambridge University Press.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 4* (pp. 463–502). Edinburgh University Press. (reprinted in McC90)

Minsky, M. L. (1963). Steps toward artificial intelligence. In E. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 406–450). New York: McGraw-Hill.

Newport, T. (2015). *Brains and Computers: Amino Acids versus Transistors*. Kindle. Retrieved from https://www.amazon.com/dp/B00OQFN6LA

Nicholas, J. M. (Ed.). (1977). *Images, Perception, and Knowledge*. Dordrecht-Holland: Reidel.

Popper, K. R. (1972). *Objective Knowledge*. Oxford: Oxford University Press.

Schrödinger, E. (1944). *What is life? the physical aspect of the living cell*. Cambridge University Press.

Sloman, A. (1962). *Knowing and Understanding: Relations between meaning and truth, meaning and necessary truth, meaning and synthetic necessary truth (DPhil Thesis)* (Doctoral dissertation, Oxford University). Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/62-80.html#1962

Sloman, A. (1965). 'Necessary', 'A Priori' and 'Analytic'. *Analysis*, *26*(1), 12–16. Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/62-80.html#1965-02

Sloman, A. (1971). Interactions between philosophy and artificial intelligence: The role of intuition and non-logical reasoning in intelligence. *Artificial Intelligence*, *2*(3), 209 - 225. Retrieved from http://www.sciencedirect.com/science/article/pii/0004370271900117 doi: https://doi.org/10.1016/0004-3702(71)90011-7

Sloman, A., Owen, D., Hinton, G., & O'Gorman, F. (1978, July 18-20th). Representation and Control in Vision. In D. Sleeman (Ed.), *Proc. AISB/GI Conference* (pp. 309–315). Hamburg, Germany. Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/62-80.html#1978-03

Thompson, D. W. (1917). *On Growth and Form*. Cambridge: Cambridge University Press. ((Revised Edition 1948))

Trettenbrein, P. C. (2016, Oct). The Demise of the Synapse As the Locus of Memory: A Looming Paradigm Shift? *Frontiers in Systems Neuroscience*, *10*(88). Retrieved from http://doi.org/10.3389/fnsys.2016.00088

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *237*(641), 37-72. Retrieved from https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1952.0012 doi: 10.1098/rstb.1952.0012

Ugur, E. (2010). *A Developmental Framework for Learning Affordances* (Doctoral dissertation, The Graduate School of Natural and Applied Sciences, Middle East Technical University, Ankara, Turkey). Retrieved from http://www.cns.atr.jp/~emre/papers/PhDThesis.pdf

von Neumann, J. (2012). *The Computer and the Brain (Silliman Memorial Lectures)*. Yale University Press. ((3rd Edition, with Foreword by Ray Kurzweill. Originally published 1958.))