Minds & Machines (2007) 17:249–259
DOI 10.1007/s11023-007-9065-3

Self-improving AI: an Analysis

John Storrs Hall

Received: 12 October 2006 / Accepted: 8 May 2007 / Published online: 7 September 2007

Abstract Self-improvement was one of the aspects of AI proposed for study in the 1956 Dartmouth conference. Turing proposed a "child machine" which could be taught in the human manner to attain adult human-level intelligence. In latter days, the contention that an AI system could be built to learn and improve itself indefinitely has acquired the label of the *bootstrap fallacy*. Attempts in AI to implement such a system have met with consistent failure for half a century. Technological optimists, however, have maintained that a such system is possible, producing, if implemented, a feedback loop that would lead to a rapid exponential increase in intelligence. We examine the arguments for both positions and draw some conclusions.

Keywords Artificial intelligence · Learning · Self-improving · Autogeny · Complexity barrier · Bootstrap fallacy

Introduction

In his classic paper on computing machinery and intelligence, Turing (1950) notes the difficulty of writing an intelligent program in the conventional way:

> At my present rate of working I produce about a thousand digits of programme a day, so that about sixty workers, working steadily through the fifty years might accomplish the job, if nothing went into the wastepaper basket. Some more expeditious method seems desirable.

This paper is based on Chapter 7 of the author's forthcoming book *Beyond AI: Creating the Conscience of the Machine* (Amherst, NY: Prometheus, May 2007), which was in turn based on the paper delivered at AI@50.

J. S. Hall (✉)
Storrmont, Laporte, PA, USA
e-mail: josh@autogeny.org

In view of later experience, he probably underestimated it at a billion bits; however, given modern software development environments and knowledge engineering tools, the estimate in terms of developer-days doesn't seem too far-fetched.

But he did propose a more expeditious method: program an imitation of a child mind and then educate it into adulthood. He assumes that the amount of effort in bringing up the child machine would be similar to that of educating a human. This would amount to some fraction of an adult's time for 20 years—less than 1% of the direct programming approach.

Quite aside from the difficulty of finding 60 programmers with the brilliance of Alan Turing, the problem with writing an AI program directly in Lisp (or whatever) is that the adult human mind remains perfectly capable of learning, inference, abduction, concept formation, metaphorical modes of thought, and so forth. If we can program these things at all, we might as well program them first and let the machine learn the rest for itself.

Turing's estimate of the complexity of the "child machine" is quite optimistic:

> Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed.

Having taken essentially the direct programming option for the indicated half-century, standard practice in AI stands now at the somewhat embarrassing brink of producing the equivalent of a mind with everything *except* the key abilities of being able to learn and grow. The number of "digits of program" necessary for those abilities remains to be seen; perhaps it is comparatively small, as Turing implied. What cannot be denied is that it appears to involve a level of cleverness that has eluded us all these years.

An extrapolation of this fact is built into a dictum that has a fairly strong following in the AI and cognitive science community. This is called the "bootstrap fallacy;" it is solidly based on decades of trying to create a learning system that could improve itself, only to have it bog down or "run out of steam" at some disappointingly low level. The dictum says that the notion of an unboundedly self-improving machine is a fallacy. This is unequivocally true of many specific learning machines; the interesting question is the generalization to all possible ones. So far, the experimental evidence supports it.

Universal Intelligence

The crucial question is whether there is some structure or level of capability in a learning machine that is analogous to universality in a formal model of computation; that is to say, a level of capability at which a program can improve itself without limit. Hereinafter we will use "universality" to refer to this hypothetical learning capability, and fully qualify any references to formal computational universality if they are necessary.

It is valuable to make a distinction between two closely-related concepts. Let us assume, for example, that human intelligence is competent, in the fullness of time, to create an artificial intellect the equal of a human.[1] That AI would be by assumption capable, in its turn, of creating an intellect of the same level; but there is no logical implication that either the AI or the human is thus capable of arbitrary unbounded self-improvement. The possibility remains that human intelligence is not (quite) universal, on this definition. Thus while some of the arguments below resemble ones against (or for) human-level AI, they are distinct; in particular, many of the common arguments about AI per se do not address the issue of universality.

Another distinction to be made is simply the question of processing speed. At least some of the difference in IQ in humans is simply a difference in speed; clearly, for Turing equivalent computers at least, a difference in speed does not imply a difference in ultimate function. The question is problematic, because we can straightforwardly predict, from Moore's Law,[2] that 10 years after the advent of a (mere, non-self-improving) human-level AI, the same software running on machinery of the same cost would do the same human-level tasks 1,000 times as fast as we. It could:

- read an average book in one second with full comprehension;
- take a college course, with all due homework and research, in 10 min;
- write a book, again with ample research, in 2 or 3 h;
- produce a human's lifetime intellectual output, complete with all the learning, growth, and experience involved, in a couple of weeks.

It is difficult to contemplate such an intellect and not consider it strictly more powerful than human; but in the following we will at least attempt to keep the notion of boundedness of learning ability separate from processor speed.

There remain two separable questions: whether there is a universal level or form of intelligence, and whether human intelligence qualifies. We consider first the arguments against universality; clearly, if it is impossible in general, then humans are not universal.

The Argument from Animals

The world abounds with species that have been called intelligent: crows, gray parrots, dolphins, whales, and chimpanzees are typical examples. All of these animals learn, many in ways that are as yet not well understood. Most mammalian predators teach their young hunting skills; housecats even house-train their kittens. The learned skills of the primates are widely remarked.

However: chimpanzees will never create even a chimp-level AI, no matter how long they work on it. Learning in all the animals is strictly limited; there is a clear ceiling on the kinds of concepts, or even associations, they can form.

[1] There is a similar informal concept often referred to as "AI-completeness" (see Wikipedia: "AI-complete"). However, it is always assumed that humans are AI-complete, which we do not assume, so we will avoid that usage here.

[2] The demise of Moore's law has been predicted so often since the 1970s that we feel confident in asserting that the burden of proof lies completely on any critiquer.

The same argument can be made for the human animal. We learn rapidly and copiously, especially through childhood, but in well-defined phases, as in language acquisition. There is clearly much that we are programmed to learn, again in many ways that are not understood; but like language, these programs run their course and we attain a maturity with a stable-and static-understanding of the world.

Humans do not experience runaway positive feedback or exponential self-improvement.

No known example of actual intelligence has the proposed universality; therefore it is incumbent on its proponents to show how it might actually work, and no one has done so.

The Argument from Experience

The AI community has not failed to follow Turing's advice for lack of trying. In the original Dartmouth conference proposal, one of the seven "aspects of the artificial intelligence problem" to be studied was

> 5. Self-Improvement
>
> Probably a truly intelligent machine will carry out activities which may best be described as self-improvement. Some schemes for doing this have been proposed and are worth further study. It seems likely that this question can be studied abstractly as well. (McCarthy et al. 1955)

In (1969) McCarthy wrote, "Our ultimate objective is to make programs that learn from experience as effectively as humans do. It may not be realized how far we are presently from this objective." He had first proposed the Advice Taker in 1958. It was seminal to substantial work in situation logics, planning, knowledge representation, and so forth (including the classic "monkey and bananas" problem) but never exhibited the kind of general learning capability that was its motivation. Throughout the early history of AI, there was a steady stream of attempts at a general cognitive architecture, ranging from GPS to SOAR, intended to be capable of learning in a human-like way. None was. Systems seemed to fall on two sides of a divide: those that learned from experience couldn't learn outside a predefined area; those that were general "learned" by being programmed.

The later efforts most often mentioned in connection with the bootstrap fallacy are Lenat's AM and Eurisko. This is ironically because they came the closest to exhibiting the kind of learning that was desired. Lenat et al. (1983) famously explains why, after appearing to work well, they ran out of steam.

The most obvious way for an intelligent program to extend its capabilities is by writing more code. For the first quarter-century of AI, automatic programming was a strong subfield, well represented for example in the *Handbook of Artificial Intelligence*[3] that summarizes the state of the field ca. 1980. Soon thereafter,

[3] Feigenbaum et al. (1982) Vol II, pp. 295–379.

however, automatic programming languished and has largely disappeared.[4] Its only strong remnant, genetic programming, exhibits exactly the same search-limited ceiling on complexity that haunted AM and Eurisko.

The Argument from Inductive Bias

Any learning system must have an inductive bias; the faster and more effective the learning, the stronger the bias. The stronger the bias, the more restricted the generality of learning and the more likely the process is to lodge in a local maximum.[5] This argument can be seen as a special case of the comparison of weak and strong methods in representation and search: strong methods are brittle, weak ones don't work for problems of non-trivial size.

The conclusion is that any usable learning system must be of limited generality, and thus universal systems are impossible.

The Case for Universality

Unlike the negative case, the mere fact that universality was possible would not establish that human intelligence is in fact universal. Some of the arguments below leave this question open; others proceed by arguing that humans are universal, implying universality is possible.

Algorithmic Probability

Solomonoff inductive inference is *provably complete*.[6] That means that given any string of symbols whatsoever, algorithmic probability reveals any regularities it may contain and provides a method for predicting its continuation. Because algorithmic probability is incomputable, this result is only suggestive; but it is in fact its incomputability that makes a proof of completeness possible.[7]

The result is suggestive is because there appears to be no barrier to increasingly better approximations to the incomputable ideal. This in some sense reflects what our instincts tell us about an open-ended learning machine—we could hardly expect, from our current limited knowledge, to be able to say much more about it. At the same time, it precludes any impossibility theorem that would cut us off short at any fixed level.

[4] Lou Steinberg, in a private communication, analyses the problem as that understanding program specifications substantially simpler than a higher-level-language implementation of the program requires arbitrary real-world knowledge.

[5] Baum (2004) p. 316, for example, espouses the view that inductive bias limits the possibility of a general learning mechanism.

[6] R. J. Solomonoff, "Complexity-Based Induction Systems: Comparisons and Convergence Theorems," *IEEE Trans. Information Theory* **IT-24**(4):422–432.

[7] If it were computable, we could disprove its completeness by a construction reminiscent of the ones used in Godel's Theorem and the Halting Problem. The construction essentially forces the machine to try to outwit itself.

The Argument from Biological Self-reproduction

This is an argument by analogy. We observe that no existing engineered manufacturing system is capable of self-reproduction, and that therefore arguments similar to ones against universality could be made on this basis. However, biological cells are self-reproducing, acting as an existence proof for this somewhat counter-intuitive capability.

The core of the analogy is the notion that the self-referential structures at the heart of a cell, or in von Neumann's (1966) architecture, overcome the difficulties of self-reproduction by a form of universality (in the constructor). It was in fact the search for open-ended self-improvement, machines which could transcend what he called the "complexity barrier," that led von Neumann to his studies of self-replication. We can imagine that a similar self-referential and universal ability might be possible for a "concept constructor" in the mind.

We suppose that a universal mind, composed of software, might have the capability to create more software that augments its capabilities in a way analogous to how a young animal grows into a large one, without limit in some cases such as reptiles. The small animal takes its nourishment in little nibbles; the large one in huge gulps; but the small one, given time, could have eaten all the same food.

The Argument from Evolution

Darwin didn't understand genes; even today evolution is not completely understood. We can quote McCarthy (1969) again:

> It is very difficult to see how the genetic representation scheme manages to be general enough to represent the great variety of animals observed and yet be such that so many interesting changes in the organism are represented by small genetic changes. The problem of how such a representation controls the development of a fertilized egg into a mature animal is even more difficult.

And yet somehow the "genetic representation scheme" does it. Much remains to be learned, but progress has been substantial. One of the missing pieces appears to be very similar to the same organizing principle needed for universality of mind. Evolution faces exactly the same problem as AM (or AI in general): it does some search, and builds a structure. Having the structure to work with makes the space bigger-evolution should have run out of steam, like AM, with unicellular life. That it didn't implies there is a technique or techniques that continue to reduce the search space as complexity is created.

If this is true, it seems likely that the technique could be adapted for software search; perhaps a parallel could also be drawn to the brain's search algorithm in general learning. This argument does not directly imply that humans are universal learners, but if we can discover the mechanism, it will have essentially the same effect as far as AI is concerned.

The Subjective Argument

In examining our experience with learning, many of us do not feel that there is a set limit to what we could learn, given enough time and attention. There are clearly limits to the rate, and given our finite lifetimes to the total volume we can learn. We feel we are like slow Turing machines with short tapes, which but for these limitations, could be universal.

One possible objection to this argument is by analogy to the "flashlight theory of consciousness:" that consciousness seems pervasive because we're not conscious of what we're not conscious of. It's clear, for example, that our built-in representation modules have strong limits: we can't visualize four-dimensional objects. Yet the general learning ability can find ways to represent them such that we can manipulate them mentally, predict their properties, and so forth, and this gets easier the more we work with them.

It's also clear that there are plenty of individuals who hit a wall with math, or programming, or thermodynamics, or economics, or art, or music, or whatever. Yet there are also other individuals who are extreme polymaths and have no trouble with any field: da Vinci springs to mind, as do Leibniz, Franklin, Maxwell, and von Neumann.

The Argument from Human Uniqueness

Chimpanzees, our nearest relatives, are an endangered species. There are about as many chimps in the world as the population of Green Bay, Wisconsin.[8] Chimps have a few behaviors that have been described as tool use, but these do not significantly extend their range or enhance their viability.

Humans, starting from a common ancestor with chimps about 5 million years ago, are found everywhere on the planet's surface. The human birthrate could replace the chimp population in about 8 h. Human technology extends our range by an order of magnitude or more. Two of the three classic "necessities of life," food, clothing, and shelter, are *inventions*.

Humans come with almost the same genetic endowment as chimps, yet our evolutionary results are about as different as possible. Given that the difference in effect has been the innovation and use of technology, it is hard to escape the conclusion that the difference in cause is that we crossed some mental watershed as the homonid brain developed from the simian.

We can draw an even closer line with *homo neanderthalensis*. Neanderthals had brains *bigger* than modern humans by 10%.[9] They were more heavily built and muscled. Their technology—stone tools, clothing, shelters, and fire—was good enough to let them live in north-central Europe during an ice age.

[8] 100,000. Note that some estimates of wild chimp population range up to twice that.

[9] Jared Diamond, *The Third Chimpanzee*, Harper, New York, 1992, p. 35.

About 46,000 years ago, *homo sapiens* (the Cro-Magnons) appeared in Europe, and within 6,000 years, the Neanderthals had disappeared.[10] Although the cause is a subject of debate in the paleoanthropological community, it seems inescapable that the humans out-competed the Neanderthals for game and habitat. As the wave of replacement swept across Eurasia, the two species never coexisted at any given spot for more than 1,000 years—and the estimates for the overlap keep shrinking as the evidence comes in.[11]

It isn't known whether Neanderthals had language. What is known, however, is that they innovated at a much slower rate than homo sapiens.[12] Neanderthal technology remained essentially constant for 100 millenia.

Again, it is hard to escape the conclusion that Neanderthals had more specialized mental organs, but we came with the one general-purpose one that made the difference. The human result is so outrageously different from that of any similar animal that it stands in need of a qualitative, not merely quantitative, explanation. A general, unlimited learning ability fits the bill.

The Argument from the Scientific Community

We can draw another analogy to the scientific community. Any human individual is limited by our dismally low bandwith to long-term memory, probably limited by total memory, and limited increasingly by advancing senescence and ultimate mortality on both counts. The scientific community as a whole is not. There does not appear to be any upper bound on what science can learn in the long run. At any given time, the reigning paradigms limit vision in a way consonant with the limits of individuals; but the course of science as a whole can be charted as a sequence of paradigm shifts. From the right perspective, this is the normal operation of science, on an historical timescale.

Concepts common in science, and indeed throughout the rest of technological civilization, are often wildly foreign to the environment of ancestral adaptation. An example immediately at hand is the metamathematical theory underlying the concept of the universal computer. Much of scientific understanding, indeed, requires the suppression of native intuitions; our inborn understanding of up, down, and solid ground must give way to Newtonian physics and planets spinning through space. We have the ability not only to extend our built-in special-purpose learning but to surpass it completely.

If we can build a machine that exhibits human-level intelligence, we can build one that emulates the scientific community. The universal level of learning capability reduces to the "mere" human level after all. Historically, science and technology *have* exhibited a positive feedback self-improvement along an exponential trend line. The universality hypothesis is true.

[10] Mellars, Paul, et al., "A new radiocarbon revolution and the dispersal of modern humans in Eurasia", *Nature*, 439. 931–935 (2006).

[11] "Neanderthals in Europe Killed Off Earlier", Associated Press, Feb 23, 2006.

[12] Diamond, p. 44. Note, however, that though from as recently as 1992, Diamond's information on Neanderthals is already significantly out of date.

Conclusions

The argument from the scientific community seems the most compelling. The arguments for the fallacy of bootstrapping are primarily the generalization of experience with brains and systems below the level of universality.

It may well be that the individual average human mind is just below the level of universality. A research scientist needs an IQ of about 140;[13] this agrees with the fact that roughly 1% of the population become scientists or engineers.[14] There was no need, in the ancestral environment, for everyone to innovate. One percent was plenty.[15]

We know that humans with IQs of up to 200 or so can exist, and thus such levels of intelligence are possible. People like da Vinci, Newton and Einstein are generally thought to have seen things ordinary people couldn't, no matter how long they took. Perhaps Turing really did understand how the "child machine" should have been programmed—or at least it may be that we must wait for his like to appear again. Might it be the case that a sped-up AI of IQ 100 would simply get a lot of work done, while a sped-up AI of IQ 140, or 200, would improve itself in an exponential takeoff?

The other phenomenon at work is the fact that innovation often requires the maverick individual working against the orthodoxy of the field as a whole, to induce a Kuhnian paradigm shift—a typical example being Wegener's theory of continental drift. Might not a single individual, even though relieved of limitations on learning bandwidth and lifespan, become so set in his ways as to effectively lodge in some local maximum of knowledge?

This forms something of a reply to the inductive bias argument. The human mind has many modules which learn each with its own bias. The general learning capability also involves a bias, but one built through experience and not constrained in any particular direction. Individuals often get the bias wrong and are unable to gain the proper insights; but the community as a whole can search the space of possible biases and ultimately find one that matches the structure of some part of reality.

A theoretical ultimately rational machine would predict the results of its actions and pursue the optimal course constantly. It would be endlessly creative, never having to rely on habit or arduously learned skills. Our mental equipment is much more limited, and we have to rely on remembered and copied skills almost all the time. Our creativity is the tiniest sliver, the first peek of the sun over the horizon as sunrise begins. It is just enough, combined with our imitating ability, to build up the stock of human knowledge over the centuries. But that's a world of difference from darkness.

[13] see http://www.audiblox.com/iq_scores.htm

[14] NSF Science Resources Statistics Infobrief, http://www.nsf.gov/statistics/infbrief/nsf02325/ While it is tempting to believe that scientists and engineers form the brightest one percent of the population, we can make the slightly less dubious assumption that S&E below the 99th percentile and others above it cancel out.

[15] IQ tests, of course, don't actually measure innovative ability. They measure cognitive skills that have been found to correlate with innovative ability, so the implications are purely statistical.

We conclude that universality is possible, and that we have at least one example; evolution may constitute another one.

Implications

Various writers, notably Vinge (1993) and latterly Bostrom (2003), have expressed concern that a rapidly self-improving artificial intellect could emerge on such a short time scale as to render human attempts at guidance and control ineffective. There are several reasons for a lack of alarm.[16]

Although hardware for running a human-level AI exists it is currently represented by the top 10 or so supercomputers in the world, in the author's best estimate. These are multimillion dollar installations, and have strong previous calls on their time. Even if someone were to pay to dedicate, say, Blue Gene to running an AI full time, it would only approximate a normal human intelligence.

The hardware may be available in principle, but the software is not; and while it seems likely that a properly structured AI could be a better programmer than a human of otherwise comparable cognitive abilities, automatic programming is currently one of the most poorly developed of AI's subfields. Any reasonable extrapolation of current practice predicts that early human-level AIs will be secretaries and truck drivers, not programmers.

Even humans, with the processing power of a thousand supercomputers at our disposal, take years to grow up, plus about a decade to become really expert in any given field—including AI programming. More to the point, it takes the scientific community some extended period to develop a theory, and the engineering community some further time to put it into practice. Even if we had a complete and valid theory of mind, which we do not, putting it into software would take years; early versions would be incomplete and full of bugs. The completely understood, tightly coded, highly optimized software of mature AI may run a human equivalent in real time with 10–100 teraops, but early versions will not.

On the present analysis, human-level AIs must be pushed up the scale to the level of research scientists, and there must be a community of them, before there is scope for unlimited feedback in self-improvement. This means that any takeoff will happen more slowly than might have otherwise have been imagined.

Universality

The existence of universality is surely interesting in a theoretical sense. If universal learning can exist, perhaps a mechanism more elegant than a community of human-scale intellects can be found. Perhaps the essence of such a mechanism must be present in any human-level intellect, as the subjective argument suggests. In any

[16] Parts of this section adapted from Hall, J: "Nano-enabled AI: Some Philosophical Issues" forthcoming in *International Journal of Applied Philosophy*, Fall 2006 Special Issue on Nanoethics.

case, the implications for the possibilities for certain avenues of learning research are fascinating.

The final implication of universality is the most heartening. If human intelligence is in fact universal in our sense, we are in theory capable, given world enough and time, of learning and understanding anything that any other universal learner can. The machines may ultimately be faster and more capacious, but they will differ quantitatively and not qualitatively. We do not know what wonders they will discover in the coming decades—but we can learn.

References

Baum, E. (2004). *What is thought*. Cambridge: MIT.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. In I. Smit et al. (Eds.), *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence* (Vol. 2). Int. Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12–17.

Feigenbaum, E. A. et al. (Eds.) (1982). *The handbook of artificial intelligence*. Los Altos: Morgan Kaufman 1981–2 in 3 vols.

Lenat, D. B., & Brown Jon, S. (1983). Why AM and Eurisko appear to work. *Artificial Intelligence, 23*, 269–294.

McCarthy, J., Marvin, M., Nathaniel, R., & Claude, S. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. see http://www.formal.stanford.edu/jmc/history/dartmouth.html

McCarthy, J. (1969). Programs with common sense. In Marvin, M. (Ed.), *Semantic information processing* (pp. 403–418). Cambridge: MIT.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.

Vinge, V. (1993). The Coming Technological Singularity: How to Survive in the Post-Human Era. in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*. NASA Conf. Pub., 10129:11–22.

von Neumann, J. (1966). *Theory of self-reproducing Automata*. Urbana: U. of Illinois Press.