John Storrs Hall - Self-improving AI: an Analysis (2007)

Tom Rochette <tom.rochette@coreteks.org>

November 2, 2024 — 36c8eb68

0.1 Context

0.2 Learned in this study

0.3 Things to explore

1 Overview

2 Notes

2.1 Universal Intelligence

2.1.1 The Argument from Animals

- Chimpanzees will never create even a chimp-level AI, no matter how long they work on it
  - A counter-argument would be that given enough time, biology/nature succeeded at creating us, possibly capable of doing it, thus it may be possible for future generations of humans to be able to achieve something we were not
- Learning in all the animals is strictly limited; there is a clear ceiling on the kinds of concepts, or even associations, they can form
- There is clearly much that we are programmed to learn
- Human do not experience runaway positive feedback or exponential self-improvement

2.1.2 The Argument from Experience

- Systems seemed to fall on two side of a divide:
  - those that learned from experience couldn't learn outside a predefined area
  - those that were general "learned" by being programmed
- 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
- Soon thereafter, however, automatic programming languished and has largely disappeared. Its only strong remnant, genetic programming, exhibits exactly the same search-limited ceiling on complexity that haunted AM and Eurisko

2.1.3 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
- Strong methods are brittle, weak ones don't work for problems of non-trivial size
- Any usable learning system must be of limited generality, and thus universal systems are impossible

2.2 The Case for Universality

2.2.1 Algorithmic Probability

- Solomonoff inductive inference is provably complete. 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

2.2.2 The Argument from Biological Self-reproduction

- 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

2.2.3 The Argument from Evolution

- 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

2.2.4 The Subjective Argument

- There are clearly limits to the rate, and given our finite lifetimes to the total volume we can learn
- 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

2.2.5 The Argument from Human Uniqueness

- The human result (success vs others) 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

2.2.6 The Argument from the Scientific Community

- 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

2.3 Conclusions

- It may well be that the individual average human mind is just below the level of universality
- 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?
- We conclude that universality is possible, and that we have at least one example; evolution may constitute another one

3 See also

4 References