How Intelligible is Intelligence?

Anna Salamon, Stephen Rayhawk
Machine Intelligence Research Institute

János Kramár
Harvard University, MIRI Visiting Fellow

Abstract

If human-level AI is eventually created, it may have unprecedented positive or negative consequences for humanity. It is therefore worth constructing the best possible forecasts of policy-relevant aspects of AI development trajectories—even though, at this early stage, the unknowns must remain very large.

We propose that one factor that can usefully constrain models of AI development is the "intelligibility of intelligence"—the extent to which efficient algorithms for general intelligence follow from simple general principles, as in physics, as opposed to being necessarily a conglomerate of special case solutions. Specifically, we argue that estimates of the "intelligibility of intelligence" bear on:

- Whether human-level AI will come about through a conceptual breakthrough, rather than through either the gradual addition of hardware, or the gradual accumulation of special-case hacks;

Salamon, Anna, Stephen Rayhawk, and János Kramár. 2010. "How Intelligible is Intelligence?" In *ECAP10: VIII European Conference on Computing and Philosophy*, edited by Klaus Mainzer. Munich: Dr. Hut.

This version contains minor changes.

- Whether the invention of human-level AI will, therefore, come without much warning;
- Whether, if AI progress comes through neuroscience, neuroscientific knowledge will enable brain-inspired human-level intelligences (as researchers "see why the brain works") before it enables whole brain emulation;
- Whether superhuman AIs, once created, will have a large advantage over humans in designing still more powerful AI algorithms;
- Whether AI intelligence growth may therefore be rather sudden past the human level; and
- Whether it may be humanly possible to understand intelligence well enough, and to design powerful AI architectures that are sufficiently transparent, to create demonstrably safe AIs far above the human level.

The intelligibility of intelligence thus provides a means of constraining long-term AI forecasting by suggesting relationships between several unknowns in AI development trajectories. Also, we can improve our estimates of the intelligibility of intelligence, e.g. by examining the evolution of animal intelligence, and the track record of AI research to date.

1. Framing the Question

McCarthy famously stated that "human level AI might require 1.7 Einsteins, 2 Maxwells, 5 Faradays and .3 Manhattan Projects, the project coming after the conceptual breakthroughs." That is, McCarthy suggested that the critical aspects of human-level intelligence could be expressed in a compact intelligible theory, analogous to a physical theory like special relativity, which could then be applied to artificial intelligence (AI). Others have argued that human intelligence is a "swiss army knife" of incremental special case solutions adapting humans to our specific environment (Cosmides and Tooby 1994), without any single theoretically simple structure, suggesting that progress in AI will proceed by the gradual accumulation of fortuitous specialized techniques for more and more domains.

We can think of these views as differing with respect to the "intelligibility of intelligence." We propose that many disputes in forecasting the long-term development of AI reflect, at least in part, differing estimates of this intelligibility. Recognizing this common factor could help to improve our forecasts in the face of profound uncertainty. We discuss theoretical and empirical results bearing on the intelligibility of intelligence, and argue that they modestly favor a relatively high estimate.

2. Why the "Intelligibility of Intelligence" Matters for AI Forecasting

Roughly speaking, the more intelligible AI design is, the more we should expect:

- That human-level AI will come about through a conceptual breakthrough, rather than through either the gradual addition of hardware, or the gradual accumulation of special-case hacks;
- That the invention of human-level AI will, therefore, come without much warning;
- That, if AI progress comes through neuroscience, neuroscientific knowledge will enable brain-inspired human-level intelligences (as researchers "see why the brain works") before it enables whole brain emulation (Sandberg and Bostrom 2008);
- That superhuman AIs, once created, will have a large advantage over humans in designing still more powerful AI algorithms;
- That AI intelligence growth may therefore be rather sudden past the human level; and

- That it may be humanly possible to understand intelligence well enough, and to design powerful AI architectures that are sufficiently transparent, to create demonstrably safe AIs far above the human level.

Conversely, the less intelligible AI design is, the more one should expect that AI development will be gradual, will be based on black box technologies, will involve copying heuristics from the brain rather than duplicating evolution's computationally expensive brain-design search, and will be too complicated to allow proofs or advance control.

The intelligibility of intelligence thus provides a means of constraining long-term AI forecasting by suggesting relationships between several unknowns in AI development trajectories. Since human-level AI could have unprecedented positive or negative consequences for humanity (Yudkowsky 2008), such constraints are worth seeking.

3. Relevant Theory

Intelligible, General-Purpose Hardware Designs

Today's computers are intelligibly designed artifacts that can be repurposed relatively easily for many cognitive tasks. It is because such general-purpose hardware is practical that AI is plausible at all. The Church-Turing thesis, and the concept of Turing-universal computing machines, give a partial account of why such hardware is possible. The reason this account is only partial is that it ignores the constant differences between Turing machines in the memory requirements, CPU time, and the code length needed to run a given program; since these constraints are strong (Wolpert 2001), it is an interesting empirical finding that today's computers are in fact useful for many of the tasks we desire.

The existence of simple, intelligible, general-purpose hardware designs provides an update toward the intelligibility of intelligence insofar as: (1) efficient hardware design is part of the problem of AI design (since AIs need to run on some substrate), and (2) the track record of computer hardware design to date indicates that fairly simple hardware designs can work, even without specific rigging to the desired cognitive task.

Intelligible, General-Purpose Learning Algorithms

Algorithms can be designed that perform optimally on any data source, in the sense that, if the algorithm is given access to literally infinite computing power, then, in the limit as data goes to infinity, it will predict the data source at least as well as any other computable predictor (Legg 2008). These "general purpose learning algorithms" are simple and intelligible. They fit easily in 100 lines of Python, indicating at least some intelligible structure to AI design. Similarly, algorithms can be designed which automate searches for efficient algorithms: for any problem in a class, they can find and execute

a nearly optimal algorithm to solve that problem, using total run-time within a factor 5 of the shortest provable run-time bound of any finite algorithm for that class (relative to a fixed proof system), plus an impractically large, class-dependent additive constant (Hutter 2002).

However, just as the Church-Turing thesis does not in itself tell us that versatile computers can be designed with realistic hardware constraints, Solomonoff induction does not in itself tell us whether there exist similarly general algorithms that can be run within realistic hardware constraints, nor whether any such efficient general algorithms are simple and intelligible.

The No Free Lunch Theorems

Given finite datasets, learning algorithms must make guesses that will pay off in some but not all possible worlds. In this framework, one algorithm is only "better" than another insofar as its bets are tuned to the environment in which it finds itself, as formalized by the "No Free Lunch Theorems for Machine Learning" (Wolpert and Macready 1997). It is accordingly worth asking how much humans' success stems from such special-case tuning, rather than from a general ability to converge to any data source.

Perhaps the most interesting implication of the No Free Lunch Theorems is that if there is a sort of "general intelligence" which is useful for many learning tasks, there must be a complementary sort of "general intelligibility" which is present in all of those learning tasks. The meaningfulness and intelligibility of intelligence is thus closely tied to the existence of pervasive regularities in the outside world.

Blum's Speedup Theorem

Blum (1967) proves that there is a limit to the power of any effective procedure for accelerating other computations. Applied to a complex world with computationally general physics, this suggests that there may always be some classes of physically real processes for which the fastest known prediction procedure is the process itself. Wolfram (2002) makes nearly this argument in different terms, proposing a Principle of Computational Equivalence and a related notion of computational irreducibility. Intuitively, these ideas suggest that there is no finite intelligence which can gain predictive insight into every possible system.

4. Relevant Empirical Evidence

We can gather evidence concerning the intelligibility of intelligence by looking at: (1) the evolution of animal intelligence; (2) the creation of cultural patterns that boost the effective intelligence of human communities; and (3) the track record of AI research to

date. In all three cases, we can ask how much engineering effort was needed to create the observed intelligence, and to what extent that engineering effort made use of simple, domain-general methods (rather than an accretion of uninterpretable, situation-specific tricks).

For example, if intelligence is highly intelligible, one might expect that evolution could evolve a more intelligent creature by simply adding more copies of existing brain structures—much as one can create a faster parallel computer by adding more circuits of a fixed design. Conversely, if intelligence were relatively unintelligible, to increase intelligence one would need to painstakingly identify new special-case tricks. One should in this case see no simple mutations that substantially increase intelligence, much as there is no simple edit that, when performed on the first half of the Manhattan phone book, produces the many thousands of phone numbers needed for the remaining half.

Research suggests that brain size correlates with intelligence within humans (McDaniel 2005), within rats (Anderson 1993), and across species (Deaner et al. 2007). These findings are evidence for the intelligibility of intelligence.

Further evidence toward the relative intelligibility of intelligence is provided by the fact that complex learning evolved separately in widely divergent animal lineages, including cephalopods, corvids, parrots, elephants, and primates (Emery and Clayton 2004; Roth and Dicke 2005). This convergent evolution is evidence that, if one starts from the flatworm-like common ancestor of humans and cephalopods, complex learning is relatively easy to invent. On the other hand, preliminary evidence suggests that evolving nervous systems in the first place could have required a combinatorially unlikely coincidence that does not occur on most worlds with life, and that we observe only because of anthropic effects (Hanson 1998).

Human brains excel at ancestral tasks such as visual processing and navigation, far outperforming current AI programs in most tasks, but often struggle with evolutionarily novel challenges such as extended mental arithmetic. This pattern of strengths and weaknesses can be understood as a result of using developed competencies to emulate absent ones; e.g., mechanisms for processing visual imagery can be applied to manipulate complex abstract concepts through analogy and visual symbols (Pinker 2010). This suggests both that specific cognitive abilities evolve more easily than does general-purpose thinking skill, and, given the convergent evolution of complex learning in many lineages, that many different specific cognitive abilities can be used to approximate general intelligence.

Human culture arguably presents a similar picture. Simple, intelligible innovations such as the scientific method and market economics have substantially boosted the effective intelligence of human communities—but so has the gradual accretion of such special-case technologies as potato farming, plastics, or the Arabic numeral system. And

again, in AI, progress to date has come at times through general-purpose structural innovations (e.g., Bayesian decision networks), and at times through the accumulation of special case tricks (e.g., the Google PageRank algorithm).

5. Import

The above theoretical and empirical results weakly suggest that methods for designing artificially intelligent systems may be relatively intelligible, and thus that takeoff trajectories may be relatively abrupt. While the evidence favoring intelligibility is, for now, extremely preliminary, there are several routes open for obtaining better evidence, including animal intelligence research, research into the quantitative usefulness of different human cultural innovations, and research into practical approximations of "universal induction" algorithms such as AIXI.

References

Anderson, Britt. 1993. "Evidence from the Rat for a General Factor That Underlies Cognitive Performance and That Relates to Brain Size: Intelligence?" *Neuroscience Letters* 153 (1): 98–102. doi:10.1016/0304-3940(93)90086-Z.

Blum, Manuel. 1967. "A Machine-Independent Theory of the Complexity of Recursive Functions." *Journal of the ACM* 14 (2): 322–336. doi:10.1145/321386.321395.

Cosmides, Leda, and John Tooby. 1994. "Beyond Intuition and Instinct Blindness: Toward an Evolutionarily Rigorous Cognitive Science." *Cognition* 50 (1–3): 41–77. doi:10.1016/0010-0277(94)90020-5.

Deaner, Robert O., Karin Isler, Judith Burkart, and Carel van Schaik. 2007. "Overall Brain Size, and Not Encephalization Quotient, Best Predicts Cognitive Ability Across Non-Human Primates." *Brain, Behavior and Evolution* 70 (2): 115–124. doi:10.1159/000102973.

Emery, Nathan J., and Nicola S. Clayton. 2004. "The Mentality of Crows: Convergent Evolution of Intelligence in Corvids and Apes." *Science* 306 (5703): 1903–1907. doi:10.1126/science.1098410.

Hanson, Robin. 1998. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Unpublished manuscript, September 23. Accessed August 12, 2012. http://hanson.gmu.edu/hardstep.pdf.

Hutter, Marcus. 2002. "The Fastest and Shortest Algorithm for All Well-Defined Problems." *International Journal of Foundations of Computer Science* 13 (3): 431–443. doi:10.1142/S0129054102001199.

Legg, Shane. 2008. "Machine Super Intelligence." PhD diss., University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

McDaniel, Michael A. 2005. "Big-Brained People are Smarter: A Meta-Analysis of the Relationship between In Vivo Brain Volume and Intelligence." *Intelligence* 33 (4): 337–346. doi:10.1016/j.intell.2004.11.005.

Pinker, Steven. 2010. "The Cognitive Niche: Coevolution of Intelligence, Sociality, and Language." Supplement, *Proceedings of the National Academy of Sciences of the United States of America* 107 (S2): 8993–8999. doi:10.1073/pnas.0914630107.

Roth, Gerhard, and Ursula Dicke. 2005. "Evolution of the Brain and Intelligence." *TRENDS in Cognitive Sciences* 9 (5): 250–257. doi:10.1016/j.tics.2005.03.005.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

Wolfram, Stephen. 2002. *A New Kind of Science.* Champaign, IL: Wolfram Media. http://www.wolframscience.com/nksonline/toc.html.

Wolpert, David H. 2001. "Computational Capabilities of Physical Systems." *Physical Review E* 65 (1): 016128. doi:10.1103/PhysRevE.65.016128.

Wolpert, David H., and William G. Macready. 1997. "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82. doi:10.1109/4235.585893.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.