[10] Sam Harris and the Myth of Machine Intelligence

JOBST LANDGREBE AND BARRY SMITH

In 1959 C.P. Snow published an essay titled "The Two Cultures" in which he described how the cultures of the natural and of the human sciences were evolving away from each other. This, according to Snow, was leading to the creation of two separate intellectual worlds, and he predicted that the failure of interaction between the two would harm scientific progress. Today, his prediction is confirmed—a good example is the field of so-called Artificial Intelligence (AI).

The Two Cultures

In the seventeenth century, Isaac Newton still saw himself as both a philosopher and a mathematician-physicist. Leibniz, too, the co-inventor of the calculus, was an important mathematician and philosopher. But by the eighteenth century, mathematicians like Euler, Lagrange, and Gauss concentrated on mathematics and physics and rarely made statements of a philosophical nature. A conspicuous exception is Laplace, who thought that the universe could in theory be formalized into one huge set of differential equations. (Laplace was wrong.)

The first to detect the evolution of two separate branches of science was Wilhelm Dilthey, who introduced the terminology of "*Naturwissenschaften*" and "*Geisteswissenschaften*," referring, respectively to the sciences of nature and the sciences of the human mind (or soul). (Unfortunately the term 'humanities' does not convey the meaning of "*Geisteswissenschaften*" very well.) Indeed, by the 1920s the two cultures had split so far apart from each other that it had become difficult for a non-physicist to make philosophical statements of any value about the meaning of physics, while the physicists themselves—those who paid any attention to philosophy at all—were disposed to dismiss it as an object of ridicule. Good examples from the middle of last century are Popper's embarrassing statements about quantum mechanics and Richard Feynman's remarks on what he saw as the gibberish produced by philosophers. Each demonstrated a thorough lack of knowledge of the discipline they chose to write about. Nevertheless, Feynman was one of the greatest physicists of the second half of the twentieth century.

Enter Sam Harris

Sam Harris is a contemporary illustration of the difficulties standing in the way of coherent interdisciplinary thinking in an age where science and the humanities have drifted so far apart.

Harris is a neuroscientist by training. His PhD is about experiments using functional magnetic resonance imaging (fMRI) to measure signal changes in the brains of believers and nonbelievers as they evaluated the truth and falsity of religious and nonreligious propositions. His conclusion is that there is a region of the brain involved in emotional judgment that is also behind religious reasoning. This does not, unfortunately, reveal anything at all about the nature of human religious thinking as expressed, for example, in the writings of Luther or Bultmann.

Harris knows a lot about the theory of neuroscience but, according to some of his critics, he didn't himself perform any of the experiments discussed in his PhD dissertation (Peaceful Science 2018).

We're concerned here with Harris's views on AI, and specifically with his view according to which, with the advance of AI, there will evolve a machine superintelligence with powers that far exceed those of the human mind. This he sees as something that is not merely possible, but rather a matter of inevitability.

However, even though he is a self-described neuroscientist, he does not ask himself what *intelligence* is and, starting out from there, consider the question of how a superintelligence, or indeed any kind of intelligence, could be engineered inside a machine. He merely mentions scientists who claim that a superintelligent AI might come into being and then speculates, excitedly, about how a future superintelligence would treat human beings, namely, as he puts it, "like ants."

If, however, we look carefully at what intelligence is, and at how computers really work on the basis of mathematical models, then we can see that it is forever impossible to emulate inside a computer even the intelligence of crows or rabbits, let alone that of human beings.

What Is Intelligence?

Intelligence as it is exhibited in the behavior of organisms manifests in every case the following characteristics, as pointed out already in the 1920s by Max Scheler:

1. **It is a disposition (a capability) to adapt to new situations that is enabled by the organism's physical makeup.**
2. **It is a capability whose realization is sudden—springs suddenly forth—which means that it can happen at any time.**
3. **It is realized in actions which are**
   —meaningful, or in other words, appropriate to the situation, in the sense that the actions serve the achievement by the acting organism of its goals;
   —not primed by prior experiences; thus these actions are untrained, and not a product of repeated attempts involving trial and error;
   —novel from the perspective of the acting organism.

Intelligence as exhibited by non-human organisms (in particular by birds and by higher mammals) exhibits these features, but with the restriction that the goals mentioned under 3. are in every case instinctive, they relate to the organism's inborn need to survive and reproduce in a certain ancestral environment.

For humans, in contrast, the range of actions exhibiting intelligence extends far beyond what is instinctive and includes the ability to act intelligently even in environments which are entirely novel. In addition, human intelligence is capable of mental and linguistic acts which enable abstract, propositional thinking. It is this objectifying intelligence which gives us the ability to conceive, and then deliberately plan and build—often collectively—artifacts that can allow us to survive even where there is no life at all—in polar barrens in the high arctic, for example, or in submarines, or in outer space.

Why We Cannot Model Intelligence Mathematically

We do not know at all how the brains of vertebrates (reptiles, birds, and mammals) produce the capability we call intelligence. And we do not know how the human brain performs the impressive feats of objectifying intelligence.

Neuroscience is limited by the fact that even the most powerful neuroimaging technologies cannot penetrate to phenomena at the level of the atoms and ions making up the phospholipids, proteins, and other organized molecules of which neurons are comprised. Moreover, even if, *per impossibile*, data were available regarding the electrochemical and other events taking place in the organism at this level, we still could not determine any general laws governing how these events occur, of the sort which we could use to build the mathematical models we would need to program a computer. This is not only because there are trillions of biochemical events occurring every second in the billions of cells of the human organism, but also because the ways these events occur differ from one individual organism to the next. It is thus no accident that textbooks of neuroscience contain very few mathematical equations.

What we do know is that all vertebrates (like all living organisms) are animate complex systems made up of elements of many different types at different levels of granularity (from atoms and ions up to cells and organs). Such systems have the following properties:

1. **Change and evolutionary character—complex systems are marked by sudden and continuous changes of ele-**

ment types and element combinations, including changing behaviors on the part of instances of element types. The system as a whole has a creative character, which means that at any time new elements and new patterns of interaction between these elements can come into being. An example is the human language system, which reveals this sort of creativity every time a new word is coined.

2. **Element-dependent interactions—which lead to irregularity and non-repeatability. Irregularity means that the system does not behave in a way that can be formalized using equations. Non-repeatability signifies a behavior that cannot be reproduced experimentally. Both features are manifested by, for example, the stock market, or by the Earth's weather, climate, and geothermal systems.**

3. **Force overlay—complex systems involve several forces acting at the same time and potentially interacting, as for example when you are tempted by a chocolate éclair offered by your host at a party while reminding yourself that you need to lose weight. This property is often correlated with anisotropy (which means that the effect resulting from force overlay does not propagate with the same magnitude in all directions).**

4. **Non-ergodic phase spaces—logic systems have the property that, over sufficiently long periods of time, the time in which a system element occupies any given region of the system's phase space is proportional to the volume of this region. This holds for example in the case of molecules of gas in a sealed container. In complex systems, however, the accessible microstates of the system's phase space are not equiprobable over a long period of time. This in turn means that predictions of the sort which we use when we have an ergodic phase space—for example when we predict how the molecules of gas will behave when the container is heated—are impossible.**

5. **Drivenness—a driven system is one whose interactions involve use of some external or internal energy source, where the system then acts by dissipating this energy. Plants draw energy from the sun. The animals lower down the food chain draw energy from plants. Higher animals, including humans, draw energy from plants and animals. Humans in addition have furnished their environments**

with machines (engineered inanimate driven systems such as refrigerators or food processors), which they control by supplying them with energy. (The machines cease to operate when their energy supply is cut off.) A driven system, now, lacks any sort of equilibrium state towards which it would constantly be converging. It is, precisely, driven to move from one state to the next—something that we experience in our every waking moment. In engineered systems the drivenness (the fact that they dissipate energy) is not relevant for their main function. It is not relevant to the ways you use your computer that it is also—until you switch it off—constantly dissipating heat.

6. **Context-dependence—non-fixable boundary conditions and embeddedness in one or more wider environments. How you behave from one moment to the next depends on your (physical, social, . . .) environment. How the Moon behaves is determined by the simple force of gravity, which acts always in the same way to produce the very same sort of behavior.**
7. **Chaos—inability to predict system behavior due to inability to obtain exact measurements of starting conditions. We cannot predict how your brain will operate because the measurements we would need to make of the dispositions of your neurons at any given time would be orders of magnitude below the error threshold of our measuring instruments.**

The solar system, your toaster, your car radio, are logic systems—their behavior can be predicted using logic and laws of physics. But for complex systems with the seven just-mentioned properties—including human beings—we are unable to create mathematical models that can emulate anything more than consistently repeating patterns of their behavior (such as the sleep-wake cycle). This is because every AI system is an algorithm that must run inside a computer. And every algorithm is a piece of mathematics. More precisely, to be executable on a computer an algorithm must be a piece of mathematics of a certain highly restricted sort (it must be, in the jargon of the trade, Church-Turing computable). AI systems are, in spite of this limitation, able to achieve remarkable results by means of algorithms which chain together millions and sometimes billions of parameters, as in the case, for example, of machine translation. But such an algorithm works because its inventors have found a way to construct a logic system which is a sufficiently close approximation to a subset of outputs from a complex system—in this case from the human language system—to yield useful results. For the reasons given above, there is no way to produce a logic system model of the complex system itself.

Because of this limitation, we can never create artificial *intelligence*, where 'intelligence' means the capability that is possessed by humans and higher animals, described above. AI will never become intelligent in any sense of this term that can be applied to humans, let alone *more* intelligent than humans. Our inability to model properties of the mind also means that an AI system will never develop a will—because we cannot model the will mathematically. Nor can a 'machine will' evolve spontaneously from some 'machine evolution', because we are neither able to create an environment that would mimic the processes of biological evolution nor are we able to emulate those subjects of evolution (hominids) that led to biological intelligence. We have explained all this in somewhat greater detail in our book, *Why Machines Will Never Rule the World.*

What Sam Harris Knows

Sam Harris does not seem to know anything about all of this. For like so many others, including many putative AI experts, he has failed to do the interdisciplinary work that is required to understand the opportunities and risks of AI. Instead, he talks about things he does not understand.

This is damaging to the field in which he works, conveying aspects of science to a broader public. Instead of responsibly explaining the real issues around digitization and AI, he misleads his readers with exciting horror stories which have no basis in reality. This is irresponsible, especially given the fact that there are real dangers of digitization and AI, which include at least:

1. **Public and private surveillance of individual behavior, for example by media corporations.**

2. **Private systems designed to guide (manipulate) the perception, preferences, and acts of individuals so that they become optimized from the point of view of the manipulating entity. Examples are social media platforms such as Facebook, Twitter or YouTube.**
3. **Public social credit systems imparting rewards and punishments in order to enforce (for example) political norms.**

The *first trend* is quite advanced; our Internet usage behavior is being constantly recorded and supervised by corporations which increasingly play a role in deciding who gets to say what on social media. Even in the West, there is now a tendency on the part of the state to use data from social media platforms and other traces left by users of the Internet to drive the targeting of dissenters. A good example is the withdrawal of banking services from protesting truck drivers and their supporters in the winter of 2021–22 in Canada.

The *second trend*, also called ‘nudging’, is very advanced in the world of interactive digital media. Users are systematically influenced via selective perception, targeted advertisement and messaging as well as reinforcement of behavioral patterns.

The *third trend* is well advanced in the urban centers in China, and was until recently being rolled out in Italy.

These developments are some of the threats we're facing from digitization and AI, and Sam Harris has indeed described and criticized some of them. These, and not speculations belonging to poorly conceived science fiction about superintelligences that will never exist, are the trends which should be in the focus of a neuroscientist like Harris, working to popularize the understanding of science and philosophy.

References

Feynman, Richard P., Robert B. Leighton, and Matthew Sands. 2011 [1964]. *The Feynman Lectures on Physics*. Basic Books.

Harris, Sam. 2016. Can We Build AI Without Losing Control Over It? TED Talk (September 29th) <www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it?language=en>.

Harris, Sam. 2018. Superintelligence: AI Futures and Philosophy. (April 13th). <www.youtube.com/watch?v=rpsvcVWoC5s>.

Kandel, Eric R., James H. Schwartz, Thomas M. Jessell, Steven J. Siegelbaum, and A.J. Hudspeth. 2012 [1991]. *Principles of Neural Science.* Fifth edition. McGraw Hill.

Landgrebe, Jobst, and Barry Smith. 2022. *Why Machines Will Never Rule the World: AI Without Fear*. Routledge.

Peaceful Science. 2018. Is Sam Harris a Legitimate Neuroscientist? <discourse.peacefulscience.org/t/is-sam-harris-a-legitimate-neuroscientist/2458

Popper, Karl R. 1951. Indeterminism in Quantum-Mechanics and in Classical Physics. *The British Journal for the Philosophy of Science* 1:2.

Scheler, Max. 1961 [1928]. *Man's Place in Nature.* Noonday Press.

Snow, Charles P. 1993 [1959]. *The Two Cultures*. Cambridge University Press.

Wilson, Rhoda. 2022. Italy's Dystopian Social Credit System. *Exposé News* (April 27th) <https://expose-news.com/2022/04/27/italy-dystopian-social-credit-system>.