Deep Learning and the Sociology of Human-Level Artificial Intelligence

Seth D. Baum
Global Catastrophic Risk Institute
http://sethbaum.com * http://gcrinstitute.org

Metascience 29(2), 313-317, 2020, DOI 10.1007/s11016-020-00510-6.
This version 9 July 2020

Review of: Harry Collins, *Artifictional Intelligence: Against Humanity's Surrender to Computers*, Cambridge, UK: Polity, 2019. xi+239 pages, paperback, $22.95 US.

The field of artificial intelligence (AI) has long dreamed of building computers that can mimic human thought and behavior. The tremendous recent success of AI, driven largely by a technique called deep learning, may prompt one to think that this dream is close to becoming reality. Collins, a sociologist and outsider to the field of AI, argues to the contrary: deep learning is fundamentally incapable of producing human-level intelligence, especially with respect to language, which Collins views as the preeminent feat of human cognition. Instead, other approaches will be needed, in particular approaches in which the AI is embedded within human society and able to pick up on the nuances of human language by interaction with humans.

In my view, Collins's primary contribution is to provide an outside perspective on AI from the field of sociology. This is a worthy contribution. Most work on AI comes from computer science, cognitive science, and philosophy. Sociology and other social sciences are poorly represented. Collins shows that these fields have much to offer, in particular by illustrating the nuances of human interaction and discussing their implications for AI. In doing so, Collins provides fresh perspective on classic AI topics such as the Turing test, in which an AI tries to mimic human dialog, and the Chinese room, a thought experiment in which a human who cannot speak Chinese answers questions in Chinese by mechanistically consulting a giant lookup table. More on Collins's contributions below.

Unfortunately, Collins's book suffers from at least two major shortcomings. First, it is not well connected to the state of the art in AI. It has relatively little to say on the AI techniques themselves, especially the alternatives to deep learning. It also has a fairly short bibliography, about half of which consists of Collins's prior publications. For a more extensive and authoritative critique of deep learning and discussion of the future of the field, I recommend Marcus and Davis (2019). Collins's book should be read as a supplement to this for those who wish to learn more about sociology and language as they relate to AI.

Second, Collins treats the field of AI too much as a scientific endeavor and not enough as a domain with important ethical and societal implications. One might expect a sociologist to be more attuned to ethical and societal implications, but Collins's particular background is in the sociology of scientific communities, especially gravitational wave physicists. Collins would very much like to see the field of AI operate more like the field of gravitational physics, but this is bad advice. AI has traditionally paid relatively little attention to ethical and societal implications and now struggles to shift direction in light of its newfound impact. Collins pushes the field in the wrong direction.

A third concern that some may have is the book's emphasis on long-term prospects instead of more near-term matters. Whether to focus on the near-term or long-term is a hotly contested matter in AI. My own view is that much can be gained from downplaying that debate and instead focusing on more general issues like the challenge of promoting attention to ethical and societal implications (Baum 2018). I also agree with Collins that long-term AI merits near-term attention due to its profound potential importance, though I would emphasize its ethical and not its scientific importance. Nonetheless, those who favor attention to near-term AI will find little of interest in Collins's book.

One near-term theme in the book is its concept of "the Surrender". Humans "surrender" to AI by being overly deferential to it, even though current AI is cognitively inferior to humans. For example, Collins complains of customer service workers who are less helpful than they otherwise could be because they mindlessly follow the instructions of their computer system. The phenomenon of erroneously deferring to computers is known in psychology as automation bias or automation-induced complacency (Lyell and Coiera 2017). The psychology literature has learned a lot about the matter and what can be done about it, but Collins does not discuss any of it. Indeed, Collins does not discuss the Surrender in any substantial depth; it should not have been featured in the book's title. The book is on how to make computers more capable, especially with respect to language, not on how to live with today's less capable computers.

A fourth potential concern is the book's emphasis on language as the signature accomplishment of human cognition and the preeminent task for AI to accomplish. Other AI research sees things differently. For example, Marcus and Davis (2019) emphasizes common sense, which relates to language as well as to other domains such as motor skills; Russell (2019) emphasizes the importance of ethics. Collins argues for the importance of language on grounds that it enables culture in a way that distinguishes humans from other animals, but this distinction is less relevant for AI. Which cognitive skills are most important for AI is beyond my own expertise, but my intuition is that the book would have been better if it treated language as important but not necessarily as more important than other cognitive skills.

With these concerns in mind, let us now turn to the main focus of the book, which is its treatment of language proficiency in humans and AI.

Collins argues that deep learning is fundamentally incapable of mastering human language. Deep learning relies on statistical pattern recognition in large, complex datasets. Its recent success derives largely from the recent availability of large datasets ("big data", such as from the internet) and the computing power to process them. Deep learning struggles when data is scarce or when situations are novel, and it struggles to handle conceptual relationships like "a dog is a type of animal" or "pushing something off a table causes it to fall". Deep learning is therefore relatively good at handling language tasks for which there are many precedents to be found, but not so good for many other language tasks. My sense is that this argument is probably correct, and others have made it as well, including Marcus and Davis (2019).

Consider, for example, language translation. Deep learning is quite capable of translating phrases that are common in two languages because there is plenty of data to be found. However, some phrases are only common in one language. Deep learning struggles with this. Collins demonstrates this by using Google Translate to translate certain phrases into other languages then back to English. The initial phrase is returned for some languages but

not others—presumably the correct phrase is returned when the other language is one that commonly uses a comparable phrase. For example, “I field at short leg”, an expression from cricket, is successfully returned when translating into Hindi and Afrikaans, both being languages of places where cricket is popular. In contrast, French returns “I plant with short leg”; Chinese returns “I’m on short legs” (see p.62). Google Translate is constantly evolving, so I do not get quite the same results, but I do get something similar, e.g., using Chinese, “I am in short leg field”.)

These spurious translations occur because deep learning has no internal model, no “understanding” of what it means to field at short leg. I put “understanding” in scare quotes because whether the AI “understands” in the sense of being cognitively aware is beside the point; what matters here is its computational representation of the concept, not its sapience. Human minds form conceptual models and can use them to make sense of sentences, even when they are about unfamiliar concepts. For example, I am American and know very little about cricket—I do not know what a short leg is—but even to me “I am in short leg field” is obviously wrong. This translation has changed “field” from a verb to a noun. But deep learning translation only uses patterns in language data; it lacks a conception of grammar, let alone a conception of cricket.

The central argument of the book is that for AI to achieve human-level, human-like cognition, it must be embedded in human society. It is not enough for AI to observe and process data about human society, as in deep learning. Without active participation, the AI will inevitably fall short. Collins even entertains the idea that AI may need a human-like robotic body so it can participate in society in the same ways that humans do. For example, humans commonly bond over meals, so perhaps robots would need to be able to do something analogous to eating. On the other hand, Collins notes that humans born with physical (i.e. bodily) disabilities can still gain full linguistic fluency, so perhaps an AI could also get by with limited robotic capability. Regardless, the main argument is that by participating in human society, an AI will be able to learn the concepts needed to resolve problems like translation.

As an aside, Collins’s discussion of the body responds to an earlier outside critique of AI by Dreyfus (1972). Dreyfus turned out to be right about a lot, but Collins argues that Dreyfus overstated the importance of the body. The success of Dreyfus’s other arguments is evidence that outside critics like Collins should be taken seriously.

Collins’s argument for embeddedness is advanced by analogy to the field of sociology. In sociology, it is common for researchers to embed themselves within the societies they study. This yields a deeper understanding of the society than can be obtained by observing the society from the outside. In a sense, the sociologist is the piece of scientific equipment, the device through which the world is surveyed. The process culminates in the sociologist becoming able to make the same sorts of everyday judgments that members of the society make, almost (but not quite) as if the sociologist is a member of that society. Perhaps an AI could do the same. (This is the sort of distinctive contribution that Collins makes to the study of AI.)

Collins calls for success in AI to be measured via carefully constructed versions of the Turing test. The test is an imitation game in which a human judge must engage in text-based dialog with another human and an AI. The AI passes the test if the judge cannot reliably discern which is the human and which is the AI. Versions of the Turing test have already been passed, but Collins argues persuasively that these versions are inadequate. A more

demanding test would have highly skilled judges able to ask a wide range of questions over an extended period of time, in order to dive into the nuances of the AI's linguistic behavior. The book notes that Collins has passed the imitation game in gravitational wave physicists, with gravitational wave physicists not being able to distinguish gravitational wave sociologist from gravitational wave physicist. This experience is used to demonstrate the importance of being embedded within a society in order to master its linguistic quirks.

Would an AI need to be embedded in society in order to pass a rigorous Turing test or accomplish other advanced language tasks such as translation? About a decade ago, I asked a similar question in a survey of experts in the field of artificial general intelligence (AGI) (Baum et al. 2011). AGI is AI that can reason across a wide range of domains, an advanced task in its own right. My question was whether AI would require physical robotics, virtual robotics (as in a virtual reality setting), or only a more minimal text- or voice-based embodiment. Most of the experts said that the minimal embodiment would probably suffice. Collins would likely disagree with this—the book emphasizes the importance of face-to-face interaction. My suspicion is that the AGI experts are right, that advanced AI can be developed with relatively limited interaction, but it seems like a difficult matter to resolve at this time. It would be interesting to study the current views of experts in AGI and other subfields of AI, and to obtain their reactions to Collins's argument.

Ultimately, the most important question is what exactly should be the focus of the field of AI and the broader communities of scholars, policymakers, and others who are trying to shift the field in a better direction. Collins has provided an interesting perspective on the nuances of social interaction that AI may need to master, but little is said on the algorithmic techniques needed for this except that deep learning is not enough. There is also the essential ethical issue of which AI techniques and capabilities *should* be pursued, and the social and political matter of how to orient the field of AI in the right direction. Readers interested in these matters must look elsewhere. Collins has provided a distinctive perspective to the conversation on AI. This perspective is worth consulting as long as readers recognize how it fits into the broader study of AI.

Acknowledgments

Robert de Neufville provided helpful feedback on an earlier version of this review. Any remaining errors are the author's alone.

References

Baum, S.D. 2018. Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & Society* 33(4): 565-572.

Baum, S.D., Goertzel, B., Goertzel, T.G. 2011. How long until human-level AI? Results from an expert assessment. *Technological Forecasting & Social Change* 78(1): 185-195.

Dreyfus, H. 1972. *What computers can't do*. Cambridge, MA: MIT Press.

Lyell, D., Coiera, E. 2017. Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association* 24(2): 423-431.

Marcus, G.F., Davis, E. 2019. *Rebooting AI: Building artificial intelligence we can trust*. New York: Pantheon Books.

Russell, S. 2019. *Human compatible: Artificial intelligence and the problem of control*. New York: Viking.