2-2019

Artificial Intelligence Hits the Barrier of Meaning

Melanie Mitchell
Portland State University, mm@pdx.edu

Citation Details

Mitchell, M. (2019). Artificial Intelligence Hits the Barrier of Meaning. Information, 10(2), 51.

information

MDPI

Opinion

Artificial Intelligence Hits the Barrier of Meaning †

Melanie Mitchell [1,2]

[1] Department of Computer Science, Portland State University, Portland, OR 97207-0751, USA; mm@pdx.edu
[2] Santa Fe Institute, Santa Fe, NM 87501, USA

† This article is reprinted from Mitchell M. Artificial Intelligence Hits the Barrier of Meaning, *New York Times*, 5 November 2018.

Received: 20 January 2019; Accepted: 2 February 2019; Published: 5 February 2019

Abstract: Today's AI systems sorely lack the essence of human intelligence: Understanding the situations we experience, being able to grasp their meaning. The lack of humanlike understanding in machines is underscored by recent studies demonstrating lack of robustness of state-of-the-art deep-learning systems. Deeper networks and larger datasets alone are not likely to unlock AI's "barrier of meaning"; instead the field will need to embrace its original roots as an interdisciplinary science of intelligence.

Keywords: deep neural networks; meaning; understanding

You've probably heard that we're in the midst of an AI revolution. We're told that machine intelligence is progressing at an astounding rate, powered by "deep learning" algorithms that use vast amounts of data to train complicated programs known as "neural networks." Today's AI programs can recognize faces and transcribe spoken sentences. We have programs that can spot subtle financial fraud, find relevant web pages in response to ambiguous queries, map the best driving route to most any destination, beat human grandmasters at chess and Go, and translate between hundreds of languages. What's more, we've been promised that self-driving cars, automated cancer diagnoses, housecleaning robots, and even automated scientific discovery are on the verge of becoming mainstream. Facebook founder Mark Zuckerberg recently declared that, over the next five to ten years, the company will push its AI to "get better than human level at all of the primary human senses: vision, hearing, language, general cognition" [1]. Shane Legg, Chief Scientist of Google's DeepMind group, predicted that "human-level AI will be passed in the mid-2020s" [2].

As someone who has worked in AI for decades, I've witnessed the failure of similar predictions of imminent human-level AI, and I'm certain these latest forecasts will fall short as well. The challenge of creating humanlike intelligence in machines remains vastly underestimated. Today's AI systems sorely lack the essence of human intelligence: Understanding the situations we experience, being able to grasp their meaning. The mathematician and philosopher Gian-Carlo Rota famously asked, "I wonder whether or when AI will ever crash the barrier of meaning." To me, this is still the most important question concerning AI.

The lack of humanlike understanding in machines is underscored by recent cracks that have appeared in the foundations of modern AI. While today's programs are much more impressive than the systems we had twenty or thirty years ago, a series of research studies have shown that deep-learning systems can be unreliable in decidedly unhuman-like ways.

I'll give a few examples. "The bareheaded man needed a hat" is transcribed by my phone's speech-recognition program as "The bear headed man needed a hat." Google Translate renders "I put the pig in the pen" into French as "Je mets le cochon dans le stylo" (mistranslating "pen" in the sense of a writing instrument). Programs that "read" documents and answer questions about them can easily be fooled into giving wrong answers when short, irrelevant snippets of text are appended

to the document [3]. Similarly, programs that recognize faces and objects, lauded as a major triumph of deep learning, can fail dramatically when their input is modified even in modest ways by certain types of lighting, image filtering, and other alterations that do not affect humans' recognition abilities in the slightest [4]. One recent study showed that adding small amounts of noise to a face image can seriously harm the performance of state-of-the-art face-recognition programs [5]. Another study, humorously called "The Elephant in the Room", showed that inserting a small image of an out-of-place object, such as an elephant, in the corner of a living-room image strangely caused deep-learning vision programs to suddenly misclassify other objects in the image [6]. Furthermore, programs that have learned to play a particular video or board game at a "superhuman" level are completely lost when the game they have learned is slightly modified (e.g., the background color on a video-game screen is changed, or the virtual "paddle" for hitting "balls" changes position) [7,8].

These are only a few examples demonstrating that the best AI programs can be unreliable when faced with situations that differ—even in a small degree—from what they have been trained on. The errors made by such systems range from harmless and humorous to potentially disastrous: Imagine, for example, an airport security system that won't let you board your flight because your face is confused with that of a criminal, or a self-driving car that, due to unusual lighting conditions, fails to notice that you are about to cross the street.

Even more worrisome are recent demonstrations of the vulnerability of AI systems to so-called "adversarial examples", in which a malevolent hacker can make specific changes to images, sound waves, or text documents, which, while imperceptible or irrelevant to humans, will cause a program to make potentially catastrophic errors. The possibility of such attacks has been demonstrated in nearly every application domain of AI, including computer vision, medical image processing, speech recognition, and language processing. Numerous studies have demonstrated the ease with which hackers could, in principle, fool face- and object-recognition systems with specific minuscule changes to images [9], put inconspicuous stickers on a stop sign to make a self-driving car's vision system mistake it as a yield sign [10], or modify an audio signal so that it sounds like background music to a human but instructs a Siri or Alexa system to perform a hidden command [11].

These potential vulnerabilities illustrate the ways in which current progress in AI is stymied by the barrier of meaning. Anyone who works with AI systems knows that behind the facade of humanlike visual abilities, linguistic fluency, and game-playing prowess, these programs do not—in any humanlike way—understand the inputs they process or the outputs they produce. The lack of such understanding renders these programs susceptible to unexpected errors and undetectable attacks.

What would be required to surmount this barrier, to give machines the ability to more deeply understand the situations that they face, rather than have them rely on shallow features? To find the answer, we need to look to the study of human cognition. Our own understanding of the situations we encounter is grounded in broad intuitive "commonsense knowledge" about how the world works, and about the goals, motivations, and likely behavior of other living creatures, particularly other humans. Additionally, our understanding of the world relies on our core abilities to *generalize* what we know, to form abstract concepts, and to make analogies—in short, to flexibly adapt our concepts to new situations. Researchers have been experimenting for decades with methods for imbuing AI systems with intuitive common sense and robust humanlike generalization abilities, but as yet there has been little progress in this very difficult endeavor.

AI programs that lack common sense and other key aspects of human understanding are increasingly being deployed for real-world applications. While some people are worried about "superintelligent" AI, the most dangerous aspect of AI systems is that we will trust them too much and give them too much autonomy, while not being fully aware of their limitations. As AI researcher Pedro Domingos noted, "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world" [12].

The race to commercialize AI has put enormous pressure on researchers to produce systems that work "well enough" on narrow tasks. But ultimately, the goal of developing trustworthy AI will

require a deeper investigation into our own remarkable abilities, and new insights into the cognitive mechanisms we ourselves use to reliably and robustly understand the world. Unlocking AI's barrier of meaning will likely require a step backward for the field, away from ever bigger networks and more massive data collections, and back to the field's original roots as an interdisciplinary science studying the most challenging of scientific problems: The nature of intelligence.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McCracken, H. Inside Mark Zuckerberg's Bold Plan for the Future of Facebook (2015). *Zuckerberg Transcripts*. 16 November 2015. Available online: https://dc.uwm.edu/zuckerberg_files_transcripts/210 (accessed on 3 February 2019).
2. Despres, J. Scenario: Shane Legg. Available online: http://future.wikia.com/wiki/Scenario:_Shane_Legg (accessed on 4 December 2018).
3. Jia, R.; Liang, P. Adversarial Examples for Evaluating Reading Comprehension Systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 1–7 September 2017.
4. Hendrycks, D.; Dietterich, T.G. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv* **2018**, arXiv:1807:01697.
5. Goswami, G.; Ratha, N.; Agarwal, A.; Singh, R.; Vatsa, M. Unravelling robustness of deep learning based face recognition against adversarial attacks. In Proceedings of the American Association for Artificial Intelligence (AAAI 2018), New Orleans, LA, USA, 2–7 February 2018; pp. 6829–6836.
6. Rosenfeld, A.; Zemel, R.; Tsotsos, J.K. The elephant in the room. *arXiv* **2018**, arXiv:1808.03305.
7. Rusu, A.A.; Rabinowitz, N.C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsel, R. Progressive neural networks. *arXiv* **2016**, arXiv:1606.04671.
8. Kansky, K.; Silver, T.; Mély, D.A.; Eldawy, M.; Lázaro-Gredilla, M.; Lou, X.; Dorfman, N.; Sidor, S.; Phoenix, S.; George, D. Schema Networks: Zero-Shot Transfer With a Generative Causal Model of Intuitive Physics. In Proceedings of the International Conference on Machine Learning (2017), Stockholm, Sweden, 10–15 July 2018; pp. 1809–1818.
9. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. In Proceedings of the International Conference on Learning Representations (2014), Banff, MB, Canada, 14–16 April 2014.
10. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1625–1634.
11. Carlini, N.; Wagner, D. Audio Adversarial Examples: Targeted Attacks on Speech-To-Text. In Proceedings of the First Deep Learning and Security Workshop (2018), San Francisco, CA, USA, 24 May 2018.
12. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*; Basic Books: New York, NY, USA, 2015.