Charles Bolton
1/15/2019

They Can Pew, but Will Robots Ever Zing?

Melanie Mitchell's interesting new book, *Artificial Intelligence: A Guide for Thinking Humans,* is mixed-parts pop-science, history, ethics, autobiography and, naturally, computers. To the lattermost end, it's light on detailed algorithms, hideous textbook pseudocode and mathematical proofs; this is of course intentional, as the net cast toward its intended audience could be loosely labelled "whoever reads the New York Times or likes Brian Greene." And that's a *good* thing. I was able to put it down but only did so to sleep over the four days I was glued to its pages or whatever. It may not be "riveting," but it's readable in an eminent way.

This is largely due to Mitchell's excellently clear writing and occasional humor. For instance, concluding a sub-chunk called "Lost in Translation," a discussion on Natural Language Processing (NLP) in AI, Mitchell cautions us to "take [the results of machine translation] with a grain of salt." She notes that, given the idiom above, Google Translate will return "bring a salt bar" if you revolving-door translate it through Chinese. In perfect comedic timing, she quips "that might be a better idea" (p. 208). I laughed out loud on the treadmill at the gym; you can feel Mitchell beaming while a ghostly, inaudible "zing!" incants, as somewhere a neural structure that appreciates that meme shoots electricity around (is that the sound it makes?) in what is however a brain actually works[1]. Unfortunately, when I tried to recreate this experience on my own, Google Translate gave

[1]

me “sprinkle a grain of salt.” One imagines a sheepish engineer from Google’s DeepMind manually tweaking the encoder-decoder network at 4 am upon receipt of their advance copy. What’s more likely, though, is that continued iterative use of human “turk editors” in the form of you and me clicking the “suggest an edit” button changed the word vectors that created this newer translation. Because as we’ll see, either way, for now at least, AI needs our supervision.

But why should we help in the first place? I imagine people in the translation business wonder if they’re not “suggesting” their way toward unemployment, or worse. It turns out that this species of question goads us toward what could be the central ideas of *Artificial Intelligence*. Indeed, the book itself is at times explicitly *not* an ethics book (“an important discussion but beyond the scope of this book” (p. 99), but then there’s also a chapter “On Trustworthy and Ethical AI,” which covers in detail moral dilemmas and ethical issues involving AI. Despite its ambivalence toward ethics, if you think through the layers that Mitchell has quietly riveted together, *Artificial Intelligence* leaves uncamouflaged its persuasive objectives, namely: we need to update our tired (and wrong) evil machines memes, because we nurse the infancy of the robot race, guiding them with an intelligence that is necessarily human.

Don’t Be an Idiot

Needless to say, artificial intelligence *as a character*, as it has(n’t) evolved in the domains of science fiction and pop culture, remains a rampaging demon-droid, spurring promulgations of future-imagined dystopias with such frequency that fear of machines has by now been memetically embedded in our psyches enough to make this sentence unnecessary. Though there is certainly value in fear, Mitchell encourages us to separate real fear from phobia.

Artificial Intelligence is pointedly subtitled *A Guide for* ***Thinking*** *Humans*. Mitchell's subtle jab/joke here is that tackling questions about the consequences of machine intelligence requires *thinking*, and carefully at that. The other clue is that we are, by taxonomy, *sapient* humans; the adjective is superfluous; the subtitle might well have been *A Guide for Humans*; but the title, following the precedent set by its inspiration and predecessor, *Godel Escher Bach*, is intentionally caked with several interleaved layers of meaning. With respect to the *Thinking* part, throughout the text, Mitchell cajoles us into *re*thinking what we thought we knew about AI.

The famous Turing test serves as the canonical example for the misunderstanding that most people have about the advancement of AI. In Turing's formulation of the so-called "imitation game," he sets the benchmark for passing the test when "the average judge [is] fooled 30 percent of the time" (p. 50). The key word here is *fool*, and Mitchell goes to lengths to show that all heretofore announcements of machines passing the test have involved mainly "superficial trickery" which have been "unanimously scoffed at" by AI experts (p. 51). In one case, a chatbot called Eugene was programmed to change the subject every time it didn't know how to respond. Given enough time, any human would begin to recognize the switch. It turns out that in multiple other moments in AI history, what people thought was a big deal turned out, upon consideration, to be a trick.

For example, in her fascinating discussion of convolutional networks, Mitchell describes how machines, using layered activation maps which output a weighted confidence measure based on some object in question from an image input. The networks are trained to recognize objects such as animals based on shapes, edges, colors, and other "activations." The rise of deep learning spawned numerous competitions that saw startling increases in programs with object-recognition capabilities. One of the programs discussed at length, AlexNet, proved to have "intriguing properties" that showed "AlexNet could easily be fooled" by an adversary if pixels were altered or other imperceptible (to humans) changes were made. It was later shown that several different CNNs with varying designs could also easily be broken with similar techniques (p.110). Despite the clamor in the media about AI that could recognize objects better than humans, these amazing, complex networks are still what Mitchell calls "brittle," easily exposed as one-trick ponies.

Facial recognition technology, again using deep neural networks, is similarly advanced and impressive. Mitchell tells the story of the ACLU testing Amazon's Rekognition system on members of Congress, finding that "the system incorrectly matched 28 out of the 535 members of Congress with people in [a] criminal database." While this number is clearly too low, the point here is actually that the system was not only incorrect, it was racist, a double dose of foolishness. This "unreliability and biases of face recognition" (p. 123) is coupled with the fact that these systems can be easily fooled by humans to recognize a face as belonging to another person.

Games have also played a starring role in the history of AI, and the treatment here is mostly positive with a detailed chronology of early chess and checkers-playing programs, to more sophisticated investments like DeepBlue and finally the recent AlphaGo tournament. These games, using a combination of CNNs, reinforcement learning and searching algorithms like Monte Carlo search, have outwitted humans in game after game. But though these programs are perennially proclaimed in headlines as harbingers of the coming robot smackdown, Mitchell again tempers our fears by explaining that claims of game-playing machines and their real capabilities are largely "wrong and misleading" because "unlike humans, none of these programs can 'transfer' anything it has learned about one game to help it learn a different game" (p.166). She emphasizes that transfer learning "is, well, *learning*" and that this very "ability to generalize what we learn is a core part of what it means for us to *think.*" In other words, these programs, given a task other than playing *Pong* better than anyone on the planet, might as well be a stone.

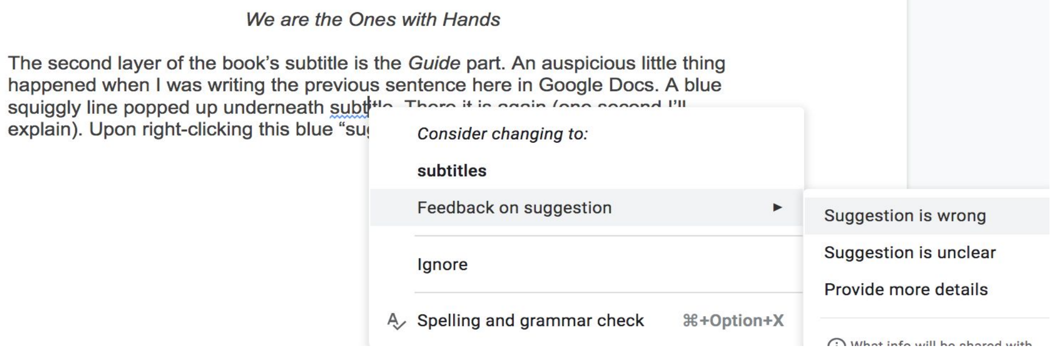
Related examples of the *single mindedness* involved in many AI programs abound throughout the book, and Mitchell remarks about the ease with which an adversary can fool image caption bots, self-driving vehicles, reading comprehension programs and other machines. She also allows us to witness some cringey robot gaffes in the domain of NLP; in a language analogy exercise, given the prompt "*Woman* is to *genius* as *man* is to ___?," the robot answered "geniuses" (p.195). Rather than delve deeper into each example of *brittle bots*, I'll conclude with one last example in the tale of Watson, which Mitchell recounts in detail. The robot was designed by IBM to play *Jeopardy!* by leveraging NLP methods with huge amounts of data, parallel computing, and neural networks; it obviously won the televised tournament in front of millions of spectators. What followed was the usual craze and speculation about AI and whether Watson would soon be able to work in fields such as medicine or law. The conclusion: "IBM have far overpromised what the technology can offer” (p.221) and no amount of hype or marketing could save Watson from being a (albeit an incredibly impressive one) “parlor trick.”

What do AI researchers really want? According to Mitchell, to build “an AI system that truly learns on its own—one that is more trustworthy” (p.129). For now and in the foreseeable future, though, any claim that a given artificial intelligence system has some superhuman ability should be taken with a sprinkle of salt. Indeed, in one of the more memorable passages of the book, Mitchell suggests we would be smarter to be fearful not of intelligent machines, but of “machine stupidity” (p. 279). If we allow ourselves to believe that machines are smarter than they actually are, we open ourselves up to risking our daily lives on dangerous technology like crashing self-driving cars or planes that fall out of the sky.

Anyone who has read Plato can recall how cave-shackled folk eventually come to believe they exist in a reality whose dimensionality is at least one level removed. *The Matrix* adds robots to the myth but the idea is the same. Imagining all those people watching Watson win in *Jeopardy!* on their couches in their living rooms, it’s hard not to think of the entire spectacle as a kind of shadow mirage. It’s easy to see why non-experts come to believe AI is far more advanced than it truly is. If Mitchell wants you to learn anything from this book, it’s that underneath every chess-playing automaton, there’s probably a mechanical turk. Our job is to not be fooled.

We Are the Ones With Hands

The second layer of the book’s subtitle is the *Guide* part. But first: a meta-digression: an auspicious little thing happened when I was writing the previous sentence “here” in Google Docs. A blue squiggly line popped up underneath ‘subtitle.’ There it is again (one second I’ll explain). Upon right-clicking this blue “suggestion” I’m given the option to offer “Feedback” and so I tell Google this “suggestion is wrong.” This is a new thing that I have never done before but it gets at the liver of the second important beam of *Artificial Intelligence*: we are the ones with hands held downward. Except that robots don’t have hands (usually) and if they did, they are infants so we must coddle them.



We are the Ones with Hands

The second layer of the book’s subtitle is the *Guide* part. An auspicious little thing happened when I was writing the previous sentence here in Google Docs. A blue squiggly line popped up underneath subtitle. There it is again (one second I’ll explain). Upon right-clicking this blue “su

The reason I gave Google my suggestion, something I normally can’t be bothered to do, is because I know, after reading this book, that I’m helping Google’s natural language processing robot here in Google Docs. Behind the scenes, my suggestion, along with (ostensibly) the many thousands of others, will call some function somewhere in some network which will adjust a weight or value on some wordvec. For all that we make of artificial intelligence, given a world without humans, all robots would simply not exist (obviously).

When I was in my early 20s, I briefly worked for Amazon's Mechanical Turk service, performing seemingly menial tasks for pennies. At the time, the tasks seemed meaningless, but I may have been helping contribute to a datasets for AI research groups. The need for 'turkers' comes from the fact that training an AI model requires huge amounts of data. This data often needs to be categorized, labeled, edited, or otherwise given some information that currently only humans can provide. People are needed not only to process menial labeling tasks, but researchers are needed to manually adjust code, hardware, datasets, and write newer, better networks. For the ImageNet Visual Recognition Challenge, "competitors were given labeled training images—1.2 million of them" (p. 85). Similar examples abound throughout the book.

Perhaps the most interesting part of the book is Mitchell's discussion of training a Robo-dog to play soccer. The main technique used in this domain is called reinforcement learning, whereby the robot is allowed to choose randomly from some set of actions in some state until it eventually does something desirable. Once this goal is met, usually after many "episodes" a value in the robot's Q-table is updated to reflect a virtual reward, and the process repeats. After much more elaboration, Mitchell points out that "designing successful reinforcement-learning systems" remains a kind of art, "mastered by a relatively small group of experts who, like their deep-learning counterparts, spend a lot of time tuning hyperparameters" (p. 143).

Reinforcement learning and searching algorithms were apotheosized one echelon higher with the recent success of AlphaGo after it defeated the world champion Go player, Lee Sedol. Following this achievement, AlphaGo's design team DeepMind stated

> Our results comprehensively demonstrate that a pure reinforcement learning approach is fully feasible, even in the most challenging of domains: it is possible to train to superhuman level, without human examples or guidance, given no knowledge of the domain beyond basic rules.

Mitchell refutes this claim by arguing that human guidance was "critical to [AlphaGo's] success" and that "the specific architecture of its convolutional neural network, the use of Monte Carlo tree search, and the setting of the many hyperparameters" were built in by the programmers (p. 167). Even though reinforcement itself implies that the robots learn by trial and error, it's clear that without the world's best researchers guiding robots or using enormous troves of human-generated data, they will never succeed on their own.

Intelligence is Human

The last layer of the book's title is the most felt, and it is of course the *Thinking* ***Humans*** layer. In the introduction, titled "Terrified," Mitchell shares a conversation she had with a "terrified" Douglas Hofstadter who lamented not the imminent robot apocalypse, but rather, the fear of human "death," darkly remarking "if such minds of infinite subtlety and complexity and emotional depth could be trivialized by a small chip, it would destroy my sense of what humanity is about" (p. 12). This is preceded by a section about EMI, a robot which created classical piano pieces in the style of Chopin so credibly that people believed they were written by Chopin. Hofstadter's fear is about existential obsolescence and the resulting abyss that humans will forever stare into when/if machines supersede our creative powers. Mitchell’s take on this is heavy on the *if* side, and ultimately tends toward something completely different.

Earlier I made the dumb-sounding comment that Amazon’s racist facial recognition software was “a double dose of foolishness,” but the real story is that it’s just a single dose, carried through to machines from humans. This incident, along with other examples of racist AI like the image-caption program that tagged a group of African-Americans as “gorillas,” or the digital camera that thought that one Asian’s eyes were closed (p. 107), highlight the human intelligence (or lack thereof) in our artificial programs. The study of artificial intelligence serves as a cultural mirror, showing that our mental biases are reflected in the machines we make.

“The true challenge is to create machines that can actually *understand* the situations that they confront” (p. 123). But what is more revelatory about artificial intelligence is what is *not* carried over from our human minds, and it turns out this is practically everything. While robots can be trained to do certain tricks, Mitchell repeatedly reminds us that they still don’t understand that they’ve performed a trick. Not only that, but giving them the ability to understand might be tantamount to understanding everything about how a brain actually works.

My favorite instance of this problem of “the barrier of meaning” in the book occurs when Mitchell is talking about language and metaphor. For example, a warm beverage appears to be related to our feeling of emotional warmth, and so on. In summary, she suggests that merely *being* a human and living a physical human existence informs our mental abstractions and the connections we make intelligently. The idea is that, since robots are not human, they will fundamentally not have the same intuitions we have, and therefore will not be able to think as a human. However they reason, they may never be "reasonable."

Efforts to instill human intelligence have been embarked upon, among them the project Cyc, a symbolic AI system intended to be an encyclopedia about human common sense. According to Mitchell, most of these efforts to provide a human ability to make analogies and abstractions have failed or are still floundering in stages of infancy (p. 250).

Take "zing!"—I don't remember where it came from. I don't know if anybody ever told me that zing *meant* something. Nor does the dictionary definition really capture what zing means. Aside from the main definitions, there's a note which says:

> *[with object] North American*; attack or criticize sharply: 'he zinged the budget deal in interviews with journalists.'

This, too, is not quite the zing we all know (somehow) and love. The zing I speak of is the zing that someone shouts in a strange alien tone after they've said something that even they didn't expect to say, but which combines some kind of weird cocktail of joke, wordplay, cultural and social competence and situational intuition. I was watching an MIT lecture the other day where the lecturer asked the audience a question, then commented, "I see one person nodding, but then again they are also yawning. So, maybe they are just—nodding off..? Zing!" Is this the appropriate way to use this word? Nobody knows, but Urban Dictionary offers: "a self-proclaimed exclamation of superior bantering abilities," and "an interjection commonly used after making a witty joke at someone else's expense while they are present" among other definitions, so, maybe? This cosmic and novel ability to zing is bravely human.

To close: in the opening I remarked that "I was able to put *it* down but only did so to sleep over the four days I was glued to its pages or whatever. While not quite 'riveting', it is readable in an eminent way." This statement combines a bunch of semantic language creation skills that robots are nowhere near possessing. First, I'm rearranging the idiom of "not being able to put a book down." Then, I'm fragmenting another idiom about reading ("eyes glued to the screen/page") that I've evidently become bored with and am allowing the reader to finish, relying on her intelligence to "get" the reference. Finally, I'm making fun of the oft-blurbed "riveting" (a meme by now) and inverting the cliche of a book being "eminently" readable by denying this descriptive coupling while simultaneously making light of what I apparently perceive to be its ultimately boring-sounding persuasion by dint[2] of its alleged overuse. All of this transpires across a few sentences (and I'm one of the stupid ones!). Mitchell provokes us to ponder: how do we teach a robot to understand what the word "one" means in the sentence above (or "it" in the previous sentence)? Wrangling with these kinds of questions can actually teach *us* what intelligence means, at least in the domain of language generation.

The darkest page in the book for me? A survey found that:

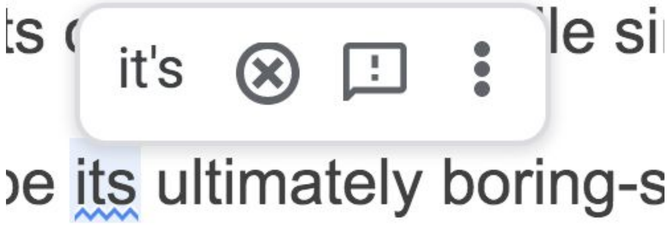
> 76 percent of participants answered that it would be morally preferable for a self-driving car to sacrifice one passenger rather than killing ten pedestrians. But when asked if they would buy a self-driving car programmed to sacrifice its

[2] A word I may not be using correctly but is likely nonetheless contextually admissible.

> passengers in order to save a much larger number of pedestrians, the overwhelming majority [...] responded that they themselves would not buy such a car. (p.128)

Mitchell quotes psychologist Joshua Greene, who says “Before we can put our values into machines, we have to figure out how to make our values clear and consistent.” Even if “the aspects of our humanity that we most cherish are not going to be matched by a ‘bag of tricks’,” if AI can help us to better recognize our biases, our inconsistencies, and, year by year, unfurl whatever makes us and our thinking minds so irrevocably sapient, it is, if nothing else, worth thinking about.

And for the someday robot-appreciators of high-culture, here you go:



ts (le si

e its ultimately boring-s

Zing!