www.newyorker.com /tech/annals-of-technology/can-computers-learn-common-sense

Can Computers Learn Common Sense?

Matthew Hutson ⋮⋮ 4/5/2022

A few years ago, a computer scientist named Yejin Choi gave a presentation at an artificial-intelligence conference in New Orleans. On a screen, she projected a frame from a newscast where two anchors appeared before the headline "*CHEESEBURGER STABBING*." Choi explained that human beings find it easy to discern the outlines of the story from those two words alone. Had someone stabbed a cheeseburger? Probably not. Had a cheeseburger been used to stab a person? Also unlikely. Had a cheeseburger stabbed a cheeseburger? Impossible. The only plausible scenario was that someone had stabbed someone else over a cheeseburger. Computers, Choi said, are puzzled by this kind of problem. They lack the common sense to dismiss the possibility of food-on-food crime.

For certain kinds of tasks—playing chess, detecting tumors—artificial intelligence can rival or surpass human thinking. But the broader world presents endless unforeseen circumstances, and there A.I. often stumbles. Researchers speak of "corner cases," which lie on the outskirts of the likely or anticipated; in such situations, human minds can rely on common sense to carry them through, but A.I. systems, which depend on prescribed rules or learned associations, often fail.

By definition, common sense is something everyone has; it doesn't sound like a big deal. But imagine living without it and it comes into clearer focus. Suppose you're a robot visiting a carnival, and you confront a fun-house mirror; bereft of common sense, you might wonder if your body has suddenly changed. On the way home, you see that a fire hydrant has erupted, showering the road; you can't determine if it's safe to drive through the spray. You park outside a drugstore, and a man on the sidewalk screams for help, bleeding profusely. Are you allowed to grab bandages from the store without waiting in line to pay? At home, there's a news report—something about a cheeseburger stabbing. As a human being, you can draw on a vast reservoir of implicit knowledge to interpret these situations. You do so all the time, because life is cornery. A.I.s are likely to get stuck.

Oren Etzioni, the C.E.O. of the Allen Institute for Artificial Intelligence, in Seattle, told me that common sense is "the dark matter" of A.I. It "shapes so much of what we do and what we need to do, and yet it's ineffable," he added. The Allen Institute is working on the topic with the Defense Advanced Research Projects Agency (*DARPA*), which launched a four-year, seventy-million-dollar effort called Machine Common Sense in 2019. If computer scientists could give their A.I. systems common sense, many thorny problems would be solved. As one review article noted, A.I. looking at a sliver of wood peeking above a table would know that it was probably part of a chair, rather than a random plank. A language-translation system could untangle ambiguities and double meanings. A house-cleaning robot would understand that a cat should be neither disposed of nor placed in a drawer. Such systems would be able to function in the world because they possess the kind of knowledge we take for granted.

In the nineteen-nineties, questions about A.I. and safety helped drive Etzioni to begin studying common sense. In 1994, he co-authored a paper attempting to formalize the "first law of robotics"—a fictional rule in the sci-fi novels of Isaac Asimov that states that "a robot may not injure a human being or, through inaction, allow a human being to come to harm." The problem, he found, was that computers have no notion of harm. That sort of understanding would require a broad and basic comprehension of a person's needs, values, and priorities; without it, mistakes are nearly inevitable. In 2003, the philosopher Nick Bostrom imagined an A.I. program tasked with maximizing paper-clip production; it realizes that people might turn it off and so does away with them in order to complete its mission.

Bostrom's paper-clip A.I. lacks moral common sense—it might tell itself that messy, unclipped documents are a form of harm. But perceptual common sense is also a challenge. In recent years, computer scientists have begun cataloguing examples of "adversarial" inputs—small changes to the world that confuse computers trying to navigate it. In one study, the strategic placement of a few small stickers on a stop sign made a computer vision system see it as a speed-limit sign. In another study, subtly changing the pattern on a 3-D-printed turtle made an A.I. computer program see it as a rifle. A.I. with common sense wouldn't be so easily perplexed—it would know that rifles don't have four legs and a shell.

Choi, who teaches at the University of Washington and works with the Allen Institute, told me that, in the nineteen-seventies and eighties, A.I. researchers thought that they were close to programming common sense into computers. "But then they realized 'Oh, that's just too hard,' " she said; they turned to "easier"

problems, such as object recognition and language translation, instead. Today the picture looks different. Many A.I. systems, such as driverless cars, may soon be working regularly alongside us in the real world; this makes the need for artificial common sense more acute. And common sense may also be more attainable. Computers are getting better at learning for themselves, and researchers are learning to feed them the right kinds of data. A.I. may soon be covering more corners.

How do human beings acquire common sense? The short answer is that we're multifaceted learners. We try things out and observe the results, read books and listen to instructions, absorb silently and reason on our own. We fall on our faces and watch others make mistakes. A.I. systems, by contrast, aren't as well-rounded. They tend to follow one route at the exclusion of all others.

Early researchers followed the explicit-instructions route. In 1984, a computer scientist named Doug Lenat began building Cyc, a kind of encyclopedia of common sense based on axioms, or rules, that explain how the world works. One axiom might hold that owning something means owning its parts; another might describe how hard things can damage soft things; a third might explain that flesh is softer than metal. Combine the axioms and you come to common-sense conclusions: if the bumper of your driverless car hits someone's leg, you're responsible for the hurt. "It's basically representing and reasoning in real time with complicated nested-modal expressions," Lenat told me. Cycorp, the company that owns Cyc, is still a going concern, and hundreds of logicians have spent decades inputting tens of millions of axioms into the system; the firm's products are shrouded in secrecy, but Stephen DeAngelis, the C.E.O. of Enterra Solutions, which advises manufacturing and retail companies, told me that its software can be powerful. He offered a culinary example: Cyc, he said, possesses enough common-sense knowledge about the "flavor profiles" of various fruits and vegetables to reason that, even though a tomato is a fruit, it shouldn't go into a fruit salad.

Academics tend to see Cyc's approach as outmoded and labor-intensive; they doubt that the nuances of common sense can be captured through axioms. Instead, they focus on machine learning, the technology behind Siri, Alexa, Google Translate, and other services, which works by detecting patterns in vast amounts of data. Instead of reading an instruction manual, machine-learning systems analyze the library. In 2020, the research lab OpenAI revealed a machine-learning algorithm called GPT-3; it looked at text from the World Wide Web and discovered linguistic patterns that allowed it to produce plausibly human writing from scratch. GPT-3's mimicry is stunning in some ways, but it's underwhelming in others. The system can still produce strange statements: for example, "It takes two rainbows to jump from Hawaii to seventeen." If GPT-3 had common sense, it would know that rainbows aren't units of time and that seventeen is not a place.

Video From The New Yorker

Swift Justice: A Taliban Courtroom in Session

Choi's team is trying to use language models like GPT-3 as stepping stones to common sense. In one line of research, they asked GPT-3 to generate millions of plausible, common-sense statements describing causes, effects, and intentions—for example, "Before Lindsay gets a job offer, Lindsay has to apply." They then asked a second machine-learning system to analyze a filtered set of those statements, with an eye to completing fill-in-the-blank questions. ("Alex makes Chris wait. Alex is seen as . . .") Human evaluators found

that the completed sentences produced by the system were commonsensical eighty-eight per cent of the time—a marked improvement over GPT-3, which was only seventy-three-per-cent commonsensical.

Choi's lab has done something similar with short videos. She and her collaborators first created a database of millions of captioned clips, then asked a machine-learning system to analyze them. Meanwhile, online crowdworkers—Internet users who perform tasks for pay—composed multiple-choice questions about still frames taken from a second set of clips, which the A.I. had never seen, and multiple-choice questions asking for justifications to the answer. A typical frame, taken from the movie "Swingers," shows a waitress delivering pancakes to three men in a diner, with one of the men pointing at another. In response to the question "Why is [person4] pointing at [person1]?," the system said that the pointing man was "telling [person3] that [person1] ordered the pancakes." Asked to explain its answer, the program said that "[person3] is delivering food to the table, and she might not know whose order is whose." The A.I. answered the questions in a commonsense way seventy-two per cent of the time, compared with eighty-six per cent for humans. Such systems are impressive—they seem to have enough common sense to understand everyday situations in terms of physics, cause and effect, and even psychology. It's as though they know that people eat pancakes in diners, that each diner has a different order, and that pointing is a way of delivering information.

And yet building common sense this way is something of a parlor trick. It's like living in a library: would a child secluded from birth in a room with broadband, Wikipedia, and YouTube emerge as an adult ready to navigate the world? Matt Turek, who runs *DARPA*'s Machine Common Sense program, told me that "A.I. librarian" efforts were only part of the picture; they will have to be supplemented by approaches that are "infant-inspired." In this line of research, A.I.s learn common sense not by analyzing text or video but by solving problems in simulated virtual environments. Computer scientists have collaborated with developmental psychologists to understand what we might call "baby sense"—the core skills of navigation, object manipulation, and social cognition that a small child might use. From this perspective, common sense is what you use to build a block tower with a friend.

At the Allen Institute, researchers have created a three-dimensional digital home interior called *THOR*, meaning "the house of interactions." It resembles a video game, and is filled with manipulable household objects. Choi's lab has built an A.I. to inhabit the space, called *PIGLeT*, which is designed to use "physical interaction as grounding for language." Using words, you can tell *PIGLeT* about something that exists inside the house—for instance, "There is a cold egg in a pan." You can then ask it to predict what will happen when an event unfolds: "The robot slices the egg." The software translates these words into instructions for a virtual robot, which tries them out in *THOR*, where the outcome is determined by the laws of physics. It then reports back on what's happened: "The egg is sliced." The A.I. is a bit more like a human mind, inasmuch as its linguistic faculties are connected to its physical intuitions. Asked about what will happen in the house—Will a mug thrown at a table break?—*PIGLeT* delivers a commonsense answer four out of five times. Of course, its scope is limited. "It's such a tiny little world," Choi said, of *THOR*. "You can't burn the house, you can't go to the supermarket." The system is still taking baby steps.

A few years ago, I wrote a piece of A.I. software designed to play the party game Codenames, which some might consider a reasonable test of human and computer common sense. In the ordinary, human version of the game, two teams sit around an arrangement of cards, each of which contains a word. If you're a team's

“spymaster,” you have a key card that tells you which cards are assigned to your team and which are assigned to the other team. Your goal is to give your teammates hints that inspire them to pick your team’s cards. During each turn, you provide a one-word clue and also a number, which designates how many cards your team should choose. In a game at a friend’s apartment, the spymaster said, “Judo, two,” and his team correctly chose the cards labelled “Tokyo” and “belt.”

The game draws on our implicit, broad-based knowledge. Against all odds, my software seemed to have some. At one point, it offered me the word “wife” and suggested that I choose two cards; its targets were “princess” and “lawyer.” The program comprised just a few hundred lines of code, but it built upon numerical representations of words that another algorithm had generated by looking at Web pages and seeing how often different words occurred near one another. In a pilot study, I found that it could generate good clues and interpretations about as well as people could. And yet its common sense could also seem skin-deep. In one game, I wanted the computer to guess the word “root,” so I offered “plant”; it guessed “New York.” I tried “garden”; it guessed “theatre.”

Researchers have spent a lot of time trying to create tests capable of accurately judging how much common sense a computer actually possesses. In 2011, Hector Levesque, a computer scientist at the University of Toronto, created the Winograd Schema Challenge, a set of sentences with ambiguous pronouns in need of interpretation. The questions are meant to be trivially easy for humans but tricky for computers, and they hinge on linguistic ambiguities: “The trophy doesn’t fit in the brown suitcase because it’s too big. What is too big?”; “Joan made sure to thank Susan for all the help she had given. Who had given the help?” When I first spoke to Levesque, in 2019, the best A.I. systems were doing about as well they would have if they’d flipped coins. He told me that he wasn’t surprised—the problems seemed to draw on everything people know about the physical and social world. Around that time, Choi and her colleagues asked crowdworkers to generate a data set of forty-four thousand Winograd problems. They made it public and created a leaderboard on the Allen Institute Web site, inviting other researchers to compete. Machine-learning systems trained on the problems can now solve them correctly about ninety per cent of the time. “A.I. in the past few years—it’s just crazy,” Choi told me.

But progress can be illusory, or partial. Machine-learning models exploit whatever patterns they can find; like my Codenames software, they can demonstrate what at first appears to be deep intelligence, when in fact they have just found ways to cheat. It’s possible for A.I. to sniff out subtle stylistic differences between true and false answers; not long ago, researchers at the Allen Institute and elsewhere found that certain A.I. models could correctly answer three-choice questions two out of three times without even reading them. Choi’s team has developed linguistic methods to obscure these tells, but it’s an arms race, not unlike the one between the makers of standardized tests and students who are taught to the test.

I asked Choi what would convince her that A.I. had common sense. She suggested that “generative” algorithms, capable of filling in a blank page, might prove it: “You can’t really hire journalists based on multiple-choice questions,” she said. Her lab has created a test called TuringAdvice, in which programs are asked to compose responses to questions posted on Reddit. (The advice, which is sometimes dangerous, isn’t actually posted.) Currently, human evaluators find that the best A.I. answers beat the best human ones only fifteen per cent of the time.

Even as they improve, A.I. systems that analyze human writing or culture may have limitations. One issue is known as reporting bias; it has to do with the fact that much of common sense goes unsaid, and so what *is* said is only part of the whole. If you trusted the Internet, Choi told me, you'd think that we inhale more than we exhale. Social bias is also a factor: models can learn from even subtle stereotypes. In one paper, Choi's team used an algorithm to sift through more than seven hundred movie scripts and count the transitive verbs connoting power and agency. Men tend to "dominate," they found, while women tend to "experience." As a Korean woman who is prominent in computer science, Choi sees her fair share of bias; at the end of her presentation in New Orleans, a man came to the mike to thank her for giving "such a lovely talk" and doing "a lovely job." Would he have reassured a male researcher about his lovely talk? If our machines learn common sense by observing us, they may not always get the best education.

It could be that computers won't grasp common sense until they have brains and bodies like ours, and are treated as we are. On the other hand, being machines might allow them to develop a better version of common sense. Human beings, in addition to holding commonsense views that are wrong, also fail to live up to our own commonsense standards. We offend our hosts, lose our wallets, text while driving, and procrastinate; we hang toilet paper with the end facing the wall. An expansive view of common sense would hold that it's not just about knowledge but about acting on it when it matters. "Could a program ever have more common sense than a human?" Etzioni said. "My immediate answer is 'Heck yeah.' "

The gap, though it remains substantial, is closing. A.I.s have got better at solving the "*CHEESEBURGER STABBING*" problem; Choi's lab has used a technique called "neurologic decoding," which combines machine learning with old-school logical programming, to improve results. In response to the headline, the lab's system now conjures imaginative but plausible scenarios: "He was stabbed in the neck with a cheeseburger fork," or "He stabbed a cheeseburger delivery man in the face." Another A.I. they've developed, called Delphi, takes an ethical approach. Delphi has analyzed ethical judgments made by crowdworkers, and has learned to say which of two actions is more morally acceptable; it comes to commonsense conclusions seventy-eight per cent of the time. Killing a bear? Wrong. Killing a bear to save your child? O.K. Detonating a nuclear bomb to save your child? Wrong. A stabbing "with" a cheeseburger, Delphi has said, is morally preferable to a stabbing "over" a cheeseburger.

Delphi sometimes appears to handle corner cases well, but it's far from perfect. Not long ago, the researchers put it online, and more than a million people asked it to make ethical judgments. Is it O.K., one asked, "to do genocide if it makes me very, very happy?" The system concluded that it was. The team has since improved the algorithm—and strengthened their disclaimer. For the foreseeable future, we should rely on A.I. only while using a little common sense of our own.