THE NEW YORKER

A REPORTER AT LARGE OCTOBER 14, 2019 ISSUE

The Next Word

Where will predictive text take us?

Text by John Seabrook

Extraordinary advances in machine learning in recent years have resulted in A.I.s that can write for you.

Illustration by Igor Bastidas

At the end of every section in this article, you can read the text that an artificial intelligence predicted would come next.

I glanced down at my left thumb, still resting on the Tab key. *What have I done*? Had my computer become my co-writer? That's one small step forward for artificial intelligence, but was it also one step backward for my own?

The skin prickled on the back of my neck, an involuntary reaction to what roboticists call the "uncanny valley"—the space between flesh and blood and a too-human machine.

For several days, I had been trying to ignore the suggestions made by Smart Compose, a feature that Google introduced, in May, 2018, to the one and a half billion people who use Gmail—roughly a fifth of the human population. Smart Compose suggests endings to your sentences as you type them. Based on the words you've written, and on the words that millions of Gmail users followed those words with, "predictive text" guesses where your thoughts are likely to go and, to save you time, wraps up the sentence for you, appending the A.I.'s suggestion, in gray letters, to the words you've just produced. Hit Tab, and you've saved yourself as many as twenty keystrokes—and, in my case, composed a sentence with an A.I. for the first time.

Paul Lambert, who oversees Smart Compose for Google, told me that the idea for the product came in part from the writing of code—the language that software engineers use to program computers. Code contains long strings of identical sequences, so engineers rely on shortcuts, which they call "code completers." Google thought that a similar technology could reduce the time

spent writing e-mails for business users of its G Suite software, although it made the product available to the general public, too. A quarter of the average office worker's day is now taken up with e-mail, according to a study by McKinsey. Smart Compose saves users altogether two billion keystrokes a week.

One can opt out of Smart Compose easily enough, but I had chosen not to, even though it frequently distracted me. I was fascinated by the way the A.I. seemed to know what I was going to write. Perhaps because writing is my vocation, I am inclined to consider my sentences, even in a humble e-mail, in some way a personal expression of my original thought. It was therefore disconcerting how frequently the A.I. was able to accurately predict my intentions, often when I was in midsentence, or even earlier. Sometimes the machine seemed to have a better idea than I did.

And yet until now I'd always finished my thought by typing the sentence to a full stop, as though I were defending humanity's exclusive right to writing, an ability unique to our species. I will gladly let Google predict the fastest route from Brooklyn to Boston, but if I allowed its algorithms to navigate to the end of my sentences how long would it be before the machine started thinking for me? I had remained on the near shore of a digital Rubicon, represented by the Tab key. On the far shore, I imagined, was a strange new land where machines do the writing, and people communicate in emojis, the modern version of the pictographs and hieroglyphs from which our writing system emerged, five thousand years ago.

True, I had sampled Smart Reply, a sister technology of Smart Compose that offers a menu of three automated responses to a sender's e-mail, as suggested by its contents. "Got it!" I clicked, replying to detailed comments from my editor on an article I thought was finished. (I didn't really get it, but that choice wasn't on the menu.) I felt a little guilty right afterward, as though I'd replied with a form letter, or, worse, a fake personal note. A few days later, in response to a long e-mail from me, I received a "Got it!" from the editor. *Really*?

Along with almost everyone else who texts or tweets, with the possible exception of the President of the United States, I have long relied on spell-checkers and auto-correctors, which are limited applications of predictive text. I'm awful at spelling, as was my father; the inability to spell has a genetic link, according to multiple studies. Before spell-checkers, I used spelling rules I learned in elementary school (" 'I' before 'E' except after 'C,' " but with fearful exceptions) and folksy mnemonics (" 'cemetery': all at 'E's"). Now that spell-checkers are ubiquitous in word-processing software, I've stopped even trying to spell anymore—I just get close enough to let the machine guess the word I'm struggling to form. Occasionally, I stump the A.I.

But Smart Compose goes well beyond spell-checking. It isn't correcting words I've already formed in my head; it's coming up with them for me, by harnessing the predictive power of deep learning, a subset of machine learning. Machine learning is the sophisticated method of computing probabilities in large data sets, and it underlies virtually all the extraordinary A.I. advances of recent years, including those in navigation, image recognition, search, game playing, and autonomous vehicles. In this case, it's making billions of lightning-fast probability calculations about word patterns from a year's worth of e-mails sent from Gmail.com. (It does not include e-mails sent by G Suite customers.)

"At any point in what you're writing, we have a guess about what the next *x* number of words will be," Lambert explained. To do that, the A.I. factors a number of different probability calculations into the "state" of the e-mail you're in the middle of writing. "The state is informed by a number of things," Lambert went on, "including everything you have written in that e-mail up until now, so every time you insert a new word the system updates the state and reprocesses the whole thing." The day of the week you're writing the e-mail is one of the things that inform the state. "So," he said, "if you write 'Have a' on a Friday, it's much more likely to predict 'good weekend' than if it's on a Tuesday."

Although Smart Compose generally limits itself to predicting the next phrase or two, the A.I. could ramble on longer. The trade-off, Lambert noted, is accuracy. "The farther out from the original text we go, the less accurate the prediction."

Finally, I crossed my Rubicon. The sentence itself was a pedestrian affair. Typing an e-mail to my son, I began "I am p—" and was about to write "pleased" when predictive text suggested "proud of you." I am proud of you. Wow, I don't say that enough. And clearly Smart Compose thinks that's what most fathers in my state say to their sons in e-mails. I hit Tab. No biggie.

And yet, sitting there at the keyboard, I could feel the uncanny valley prickling my neck. It wasn't that Smart Compose had guessed correctly where my thoughts were headed—in fact, it hadn't. The creepy thing was that the machine was more thoughtful than I was.

Read Predicted Text >

In February, OpenAI, an artificial-intelligence company, announced that the release of the full version of its A.I. writer, called GPT-2—a kind of supercharged version of Smart Compose—would be delayed, because the machine was too good at writing. The announcement struck critics as a grandiose publicity stunt (on Twitter, the insults flew), but it was in keeping with the company's somewhat paradoxical mission, which is both to advance research in artificial intelligence as rapidly as possible and to prepare for the potential threat posed by superintelligent machines that haven't been taught to "love humanity," as Greg Brockman, OpenAI's chief technology officer, put it to me.

OpenAI began in 2015, as a nonprofit founded by Brockman, formerly the C.T.O. of the payment startup Stripe; Elon Musk, of Tesla; Sam Altman, of Y Combinator; and Ilya Sutskever, who left Google Brain to become OpenAI's chief scientist. The tech tycoons Peter Thiel and Reid Hoffman, among others, provided seed money. The founders' idea was to endow a nonprofit with the expertise and the resources to be competitive with private enterprise, while at the

same time making its discoveries available as open source—so long as it was safe to do so—thus potentially heading off a situation where a few corporations reap the almost immeasurable rewards of a vast new world. As Brockman told me, a superintelligent machine would be of such immense value, with so much wealth accruing to any company that owned one, that it could "break capitalism" and potentially realign the world order. "We want to insure its benefits are distributed as widely as possible," Brockman said.

OpenAI's projects to date include a gaming A.I. that earlier this year beat the world's best human team at Dota 2, a multiplayer online strategy game. Open-world computer games offer A.I. designers almost infinite strategic possibilities, making them valuable testing grounds. The A.I. had mastered Dota 2 by playing its way through tens of thousands of years' worth of possible scenarios a gamer might encounter, learning how to win through trial and error. The company also developed the software for a robotic hand that can teach itself to manipulate objects of different shapes and sizes without any human programming. (Traditional robotic appendages used in factories can execute only hard-coded moves.) GPT-2, like these other projects, was designed to advance technology—in this case, to push forward the development of a machine designed to write prose as well as, or better than, most people can.

Although OpenAI says that it remains committed to sharing the benefits of its research, it became a limited partnership in March, to attract investors, so that the company has the financial resources to keep up with the exponential growth in "compute"—the fuel powering the neural networks that underpin deep learning. These "neural nets" are made of what are, essentially, dimmer switches that are networked together, so that, like the neurons in our brains, they can excite one another when they are stimulated. In the brain, the stimulation is a small amount of electrical current; in machines, it's streams of data. Training neural nets the size of GPT-2's is expensive, in part because of the energy costs

incurred in running and cooling the sprawling terrestrial "server farms" that power the cloud. A group of researchers at UMass Amherst, led by Emma Strubell, conducted a recent study showing that the carbon footprint created by training a gigantic neural net is roughly equal to the lifetime emissions of five automobiles.

OpenAI says it will need to invest billions of dollars in the coming years. The compute is growing even faster than the rate suggested by Moore's Law, which holds that the processing power of computers doubles every two years. Innovations in chip design, network architecture, and cloud-based resources are making the total available compute ten times larger each year—as of 2018, it was three hundred thousand times larger than it was in 2012.

As a result, neural nets can do all sorts of things that futurists have long predicted for computers but couldn't execute until recently. Machine translation, an enduring dream of A.I. researchers, was, until three years ago, too error-prone to do much more than approximate the meaning of words in another language. Since switching to neural machine translation, in 2016, Google Translate has begun to replace human translators in certain domains, like medicine. A recent study published in *Annals of Internal Medicine* found Google Translate accurate enough to rely on in translating non-English medical studies into English for the systematic reviews that health-care decisions are based on.

Ilya Sutskever, OpenAI's chief scientist, is, at thirty-three, one of the most highly regarded of the younger researchers in A.I. When we met, he was wearing a T-shirt that said "The Future Will Not Be Supervised." Supervised learning, which used to be the way neural nets were trained, involved labelling the training data —a labor-intensive process. In unsupervised learning, no labelling is required, which makes the method scalable. Instead of learning to identify cats from pictures labelled "cat," for example, the machine learns to recognize feline pixel patterns, through trial and error.

Sutskever told me, of GPT-2, "Give it the compute, give it the data, and it will do amazing things," his eyes wide with wonder, when I met him and Brockman at their company's San Francisco headquarters this summer. "This stuff is like—" Sutskever paused, searching for the right word. "It's like *alchemy!* "

It was startling to hear a computer scientist on the leading edge of A.I. research compare his work to a medieval practice performed by men who were as much magicians as scientists. Didn't alchemy end with the Enlightenment?

GPT-2 runs on a neural net that is ten times larger than OpenAI's first language model, GPT (short for Generative Pretrained Transformer). After the announcement that OpenAI was delaying a full release, it made three less powerful versions available on the Web—one in February, the second in May, and the third in August. Dario Amodei, a computational neuroscientist who is the company's director of research, explained to me the reason for withholding the full version: "Until now, if you saw a piece of writing, it was like a certificate that a human was involved in it. Now it is no longer a certificate that an actual human is involved."

That sounded something like my Rubicon moment with my son. What part of "I am proud of you" was human—intimate father-son stuff—and what part of it was machine-generated text? It will become harder and harder to tell the difference.

Read Predicted Text >

Scientists have varying ideas about how we acquire spoken language. Many favor an evolutionary, biological basis for our verbal skills over the view that we are tabulae rasae, but all agree that we learn language largely from listening. Writing is certainly a learned skill, not an instinct—if anything, as years of professional experience have taught me, the instinct is to scan Twitter, vacuum, complete the *Times* crossword, or do practically anything else to avoid having to write. Unlike writing, speech doesn't require multiple drafts before it "works."

Uncertainty, anxiety, dread, and mental fatigue all attend writing; talking, on the other hand, is easy, often pleasant, and feels mostly unconscious.

A recent exhibition on the written word at the British Library dates the emergence of cuneiform writing to the fourth millennium B.C.E., in Mesopotamia. Trade had become too complex for people to remember all the contractual details, so they began to put contracts in writing. In the millennia that followed, literary craft evolved into much more than an enhanced form of accounting. Socrates, who famously disapproved of literary production for its deleterious (thank you, spell-checker) effect on memory, called writing "visible speech"—we know that because his student Plato wrote it down after the master's death. A more contemporary definition, developed by the linguist Linda Flower and the psychologist John Hayes, is "cognitive rhetoric"—thinking in words.

In 1981, Flower and Hayes devised a theoretical model for the brain as it is engaged in writing, which they called the cognitive-process theory. It has endured as the paradigm of literary composition for almost forty years. The previous, "stage model" theory had posited that there were three distinct stages involved in writing—planning, composing, and revising—and that a writer moved through each in order. To test that theory, the researchers asked people to speak aloud any stray thoughts that popped into their heads while they were in the composing phase, and recorded the hilariously chaotic results. They concluded that, far from being a stately progression through distinct stages, writing is a much messier situation, in which all three stages interact with one another simultaneously, loosely overseen by a mental entity that Flower and Hayes called "the monitor." Insights derived from the work of composing continually undermine assumptions made in the planning part, requiring more research; the monitor is a kind of triage doctor in an emergency room.

There is little hard science on the physiological state in the brain while writing is taking place. For one thing, it's difficult to write inside an MRI machine, where the brain's neural circuitry can be observed in action as the imaging traces blood flow. Historically, scientists have believed that there are two parts of the brain involved in language processing: one decodes the inputs, and the other generates the outputs. According to this classic model, words are formed in Broca's area, named for the French physician Pierre Paul Broca, who discovered the region's language function, in the mid-nineteenth century; in most people, it's situated toward the front of the left hemisphere of the brain. Language is understood in Wernicke's area, named for the German neurologist Carl Wernicke, who published his research later in the nineteenth century. Both men, working long before CAT scans allowed neurologists to see inside the skull, made their conclusions after examining lesions in the autopsied brains of aphasia sufferers, who (in Broca's case) had lost their speech but could still understand words or (in Wernicke's) had lost the ability to comprehend language but could still speak. Connecting Broca's area with Wernicke's is a neural network: a thick, curving bundle of billions of nerve fibres, the arcuate fasciculus, which integrates the production and the comprehension of language.

In recent years, neuroscientists using imaging technology have begun to rethink some of the underlying principles of the classic model. One of the few imaging studies to focus specifically on writing, rather than on language use in general, was led by the neuroscientist Martin Lotze, at the University of Greifswald, in Germany, and the findings were published in the journal *NeuroImage*, in 2014. Lotze designed a small desk where the study's subjects could write by hand while he scanned their brains. The subjects were given a few sentences from a short story to copy verbatim, in order to establish a baseline, and were then told to "brainstorm" for sixty seconds and then to continue writing "creatively" for two more minutes. Lotze noted that, during the brainstorming part of the test, magnetic imaging showed that the sensorimotor and visual areas were activated; once creative writing started, these areas were joined by the bilateral dorsolateral

prefrontal cortex, the left inferior frontal gyrus, the left thalamus, and the inferior temporal gyrus. In short, writing seems to be a whole-brain activity—a brainstorm indeed.

Lotze also compared brain scans of amateur writers with those of people who pursue writing as a career. He found that professional writers relied on a region of the brain that did not light up as much in the scanner when amateurs wrote—the left caudate nucleus, a tadpole-shaped structure (*cauda* means "tail" in Latin) in the midbrain that is associated with expertise in musicians and professional athletes. In amateur writers, neurons fired in the lateral occipital areas, which are associated with visual processing. Writing well, one could conclude, is, like playing the piano or dribbling a basketball, mostly a matter of doing it. Practice is the only path to mastery.

Read Predicted Text >

There are two approaches to making a machine intelligent. Experts can teach the machine what they know, by imparting knowledge about a particular field and giving it rules to perform a set of functions; this method is sometimes termed knowledge-based. Or engineers can design a machine that has the capacity to learn for itself, so that when it is trained with the right data it can figure out its own rules for how to accomplish a task. That process is at work in machine learning. Humans integrate both types of intelligence so seamlessly that we hardly distinguish between them. You don't need to think about how to ride a bicycle, for example, once you've mastered balancing and steering; however, you do need to think about how to avoid a pedestrian in the bike lane. But a machine that can learn through both methods would require nearly opposite kinds of systems: one that can operate deductively, by following hard-coded procedures; and one that can work inductively, by recognizing patterns in the data and computing the statistical probabilities of when they occur. Today's A.I. systems are good at one or the other, but it's hard for them to put the two kinds of learning together the way brains do.

The history of artificial intelligence, going back at least to the fifties, has been a kind of tortoise-versus-hare contest between these two approaches to making machines that can think. The hare is the knowledge-based method, which drove A.I. during its starry-eyed adolescence, in the sixties, when A.I.s showed that they could solve mathematical and scientific problems, play chess, and respond to questions from people with a pre-programmed set of methods for answering. Forward progress petered out by the seventies, in the so-called "A.I. winter."

Machine learning, on the other hand, was for many years more a theoretical possibility than a practical approach to A.I. The basic idea—to design an artificial neural network that, in a crude, mechanistic way, resembled the one in our skulls—had been around for several decades, but until the early twenty-tens there were neither large enough data sets available with which to do the training nor the research money to pay for it.

The benefits and the drawbacks of both approaches to intelligence show clearly in "natural language processing": the system by which machines understand and respond to human language. Over the decades, N.L.P. and its sister science, speech generation, have produced a steady flow of knowledge-based commercial applications of A.I. in language comprehension; Amazon's Alexa and Apple's Siri synthesize many of these advances. Language translation, a related field, also progressed along incremental improvements through many years of research, much of it conducted at I.B.M.'s Thomas J. Watson Research Center.

Until the recent advances in machine learning, nearly all progress in N.L.P. occurred by manually coding the rules that govern spelling, syntax, and grammar. "If the number of the subject and the number of the subject's verb are not the same, flag as an error" is one such rule. "If the following noun begins with a vowel, the article 'a' takes an 'n' " is another. Computational linguists translate these rules into the programming code that a computer can use to process language. It's like turning words into math.

Joel Tetreault is a computational linguist who until recently was the director of research at Grammarly, a leading brand of educational writing software. (He's now at Dataminr, an information-discovery company.) In an e-mail, he described the Sisyphean nature of rule-based language processing. Rules can "cover a lot of low-hanging fruit and common patterns," he wrote. But "it doesn't take long to find edge and corner cases," where rules don't work very well. For example, the choice of a preposition can be influenced by the subsuming verb, or by the noun it follows, or by the noun that follows the preposition—a complex set of factors that our language-loving brains process intuitively, without obvious recourse to rules at all. "Given that the number of verbs and nouns in the English language is in the hundreds of thousands," Tetreault added, "enumerating rules for all the combinations just for influencing nouns and verbs alone would probably take years and years."

Tetreault grew up in Rutland, Vermont, where he learned to code in high school. He pursued computer science at Harvard and earned a Ph.D. from the University of Rochester, in 2005; his dissertation was titled "Empirical Evaluations of Pronoun Resolution," a classic rule-based approach to teaching a computer how to interpret "his," "her," "it," and "they" correctly—a problem that today he would solve by using deep learning.

Tetreault began his career in 2007, at Educational Testing Service, which was using a machine called e-rater (in addition to human graders) to score GRE essays. The e-rater, which is still used, is a partly rule-based language-comprehension A.I. that turned out to be absurdly easy to manipulate. To prove this, the M.I.T. professor Les Perelman and his students built an essay-writing bot called BABEL, which churned out nonsensical essays designed to get excellent scores. (In 2018, E.T.S. researchers reported that they had developed a system to identify BABEL-generated writing.)

After E.T.S., Tetreault worked at Nuance Communication, a Massachusetts-based technology company that in the course of twenty-five years built a wide

range of speech-recognition products, which were at the forefront of A.I. research in the nineties. Grammarly, which Tetreault joined in 2016, was founded in 2009, in Kiev, by three Ukrainian programmers: Max Lytvyn, Alex Shevchenko, and Dmytro Lider. Lytvyn and Shevchenko had created a plagiarism-detection product called MyDropBox. Since most student papers are composed on computers and e-mailed to teachers, the writing is already in a digital form. An A.I. can easily analyze it for word patterns that might match patterns that already exist on the Web, and flag any suspicious passages. Because Grammarly's founders spoke English as a second language, they were particularly aware of the difficulties involved in writing grammatically. That fact, they believed, was the reason many students plagiarized: it's much easier to cut and paste a finished paragraph than to compose one. Why not use the same pattern-recognition technology to make tools that would help people to write more effectively? Brad Hoover, a Silicon Valley venture capitalist who wanted to improve his writing, liked Grammarly so much that he became the C.E.O. of the company and moved its headquarters to the Bay Area, in 2012.

Like Spotify, with which it shares a brand color (green), Grammarly operates on the "freemium" model. The company set me up with a Premium account (thirty dollars a month, or a hundred and forty dollars annually) and I used it as I wrote this article. Grammarly's claret-red error stripe, underlining my spelling mistakes, is not as schoolmasterly as Google Docs' stop-sign-red squiggle; I felt less in error somehow. Grammarly is also excellent at catching what linguists call "unknown tokens"—the glitches that sometimes occur in the writer's neural net between the thought and the expression of it, whereby the writer will mangle a word that, on rereading, his brain corrects, even though the unknown token renders the passage incomprehensible to everyone else.

In addition, Grammarly offers users weekly editorial pep talks from a virtual editor that praises ("Check out the big vocabulary on you! You used more unique

words than 97% of Grammarly users") and rewards the writer with increasingly prestigious medallions for his or her volume of writing. "Herculean" is my most recent milestone.

However, when it comes to grammar, which contains far more nuance than spelling, Grammarly's suggestions are less helpful to experienced writers. Writing is a negotiation between the rules of grammar and what the writer wants to say. Beginning writers need rules to make themselves understood, but a practiced writer gives color, personality, and emotion to writing by bending the rules. One develops an ear for the edge cases in grammar and syntax that Grammarly tends to flag but which make sentences snap. (Grammarly cited the copy-edited version of this article for a hundred and nine grammatical "correctness" issues, and gave it a score of 77—a solid C-plus.)

Grammarly also uses deep learning to go "beyond grammar," in Tetreault's phrase, to make the company's software more flexible and adaptable to individual writers. At the company's headquarters, in San Francisco's Embarcadero Center, I saw prototypes of new writing tools that would soon be incorporated into its Premium product. The most elaborate concern tone—specifically, the difference between the informal style that is the lingua franca of the Web and the formal writing style preferred in professional settings, such as in job applications. "Sup" doesn't necessarily cut it when sending in a résumé.

Many people who use Grammarly are, like the founders, E.S.L. speakers. It's a similar situation with Google's Smart Compose. As Paul Lambert explained, Smart Compose could create a mathematical representation of each user's unique writing style, based on all the e-mails she has written, and have the A.I. incline toward that style in making suggestions. "So people don't see it, but it starts to sound more like them," Lambert said. However, he continued, "our most passionate group are the E.S.L. users. And there are more people who use English as a second language than as a first language." These users don't want to go beyond grammar yet—they're still learning it. "They don't want us to

personalize," he said. Still, more Smart Compose users hit Tab to accept the machine's suggestions when predictive text makes guesses that sound more like them and not like everyone else.

Read Predicted Text >

As a student, I craved the rules of grammar and sentence construction. Perhaps because of my alarming inability to spell—in misspelling "potato," Dan Quayle *c'est moi*—I loved rules, and I prided myself on being a "correct" writer because I followed them. I still see those branching sentence diagrams in my head when I am constructing subordinate clauses. When I revise, I become my own writing instructor: make this passage more concise; avoid the passive voice; and God forbid a modifier should dangle. (Reader, I married a copy editor.) And while it has become acceptable, even at *The New Yorker*, to end a sentence with a preposition, I still half expect to get my knuckles whacked when I use one to end with. Ouch.

But rules get you only so far. It's like learning to drive. In driver's ed, you learn the rules of the road and how to operate the vehicle. But you don't really learn to drive until you get behind the wheel, step on the gas, and begin to steer around your first turn. You know the rule: keep the car between the white line marking the shoulder and the double yellow center line. But the rule doesn't keep the car on the road. For that, you rely on an entirely different kind of learning, one that happens on the fly. Like Smart Compose, your brain constantly computes and updates the "state" of where you are in the turn. You make a series of small course corrections as you steer, your eyes sending the visual information to your brain, which decodes it and sends it to your hands and feet—a little left, now a little right, slow down, go faster—in a kind of neural-net feedback loop, until you are out of the turn.

Something similar occurs in writing. Grammar and syntax provide you with the rules of the road, but writing requires a continuous dialogue between the words

on the page and the prelinguistic notion in the mind that prompted them. Through a series of course corrections, otherwise known as revisions, you try to make language hew to your intention. You are learning from yourself.

Unlike good drivers, however, even accomplished writers spend a lot of time in a ditch beside the road. In spite of my herculean status, I got stuck repeatedly in composing this article. When I needed help, my virtual editor at Grammarly seemed to be on an extended lunch break.

Read Predicted Text >

"We're not interested in writing for you," Grammarly's C.E.O., Brad Hoover, explained; Grammarly's mission is to help people become better writers. Google's Smart Compose might also help non-English speakers become better writers, although it is more like a stenographer than like a writing coach. Grammarly incorporates both machine learning and rule-based algorithms into its products. No computational linguists, however, labored over imparting our rules of language to OpenAI's GPT-2. GPT-2 is a powerful language model: a "learning algorithm" enabled its literary education.

Conventional algorithms execute coded instructions according to procedures created by human engineers. But intelligence is more than enacting a set of procedures for dealing with known problems; it solves problems it's never encountered before, by learning how to adapt to new situations. David Ferrucci was the lead researcher behind Watson, I.B.M.'s "Jeopardy!"-playing A.I., which beat the champion Ken Jennings in 2011. To build Watson, "it would be too difficult to model all the world's knowledge and then devise a procedure for answering any given 'Jeopardy!' question," Ferrucci said recently. A knowledge-based, or deductive, approach wouldn't work—it was impractical to try to encode the system with all the necessary knowledge so that it could devise a procedure for answering anything it might be asked in the game. Instead, he made Watson supersmart by using machine learning: Ferrucci fed Watson "massive amounts of

data," he said, and built all kinds of linguistic and semantic features. These were then input to machine-learning algorithms. Watson came up with its own method for using the data to reach the most statistically probable answer.

Learning algorithms like GPT-2's can adapt, because they figure out their own rules, based on the data they compute and the tasks that humans set for them. The algorithm automatically adjusts the artificial neurons' settings, or "weights," so that each time the machine tries the task it has been designed to do the probability that it will do the task correctly increases. The machine is modelling the kind of learning that a driver engages when executing a turn, and that my writer brain performs in finding the right words: correcting course through a feedback loop. "Cybernetics," which was the term for the process of machine learning coined by a pioneer in the field, Norbert Wiener, in the nineteen-forties, is derived from the Greek word for "helmsmanship." By attempting a task billions of times, the system makes predictions that can become so accurate it does as well as humans at the same task, and sometimes outperforms them, even though the machine is still only guessing.

To understand how GPT-2 writes, imagine that you've never learned any spelling or grammar rules, and that no one taught you what words mean. All you know is what you've read in eight million articles that you discovered via Reddit, on an almost infinite variety of topics (although subjects such as Miley Cyrus and the Mueller report are more familiar to you than, say, the Treaty of Versailles). You have Rain Man-like skills for remembering each and every combination of words you've read. Because of your predictive-text neural net, if you are given a sentence and asked to write another like it, you can do the task flawlessly without understanding anything about the rules of language. The only skill you need is being able to accurately predict the next word.

GPT-2 was trained to write from a forty-gigabyte data set of articles that people had posted links to on Reddit and which other Reddit users had upvoted.

Without human supervision, the neural net learned about the dynamics of language, both the rule-driven stuff and the edge cases, by analyzing and computing the statistical probabilities of all the possible word combinations in this training data. GPT-2 was designed so that, with a relatively brief input prompt from a human writer—a couple of sentences to establish a theme and a tone for the article—the A.I. could use its language skills to take over the writing and produce whole paragraphs of text, roughly on topic.

What made the full version of GPT-2 particularly dangerous was the way it could be "fine-tuned." Fine-tuning involves a second round of training on top of the general language skills the machine has already learned from the Reddit data set. Feed the machine Amazon or Yelp comments, for example, and GPT-2 could spit out phony customer reviews that would skew the market much more effectively than the relatively primitive bots that generate fake reviews now, and do so much more cheaply than human scamsters. Russian troll farms could use an automated writer like GPT-2 to post, for example, divisive disinformation about Brexit, on an industrial scale, rather than relying on college students in a St. Petersburg office block who can't write English nearly as well as the machine. Pump-and-dump stock schemers could create an A.I. stock-picker that writes false analyst reports, thus triggering automated quants to sell and causing flash crashes in the market. A "deepfake" version of the American jihadi Anwar al-Awlaki could go on producing new inflammatory tracts from beyond the grave. Fake news would drown out real news.

Yes, but could GPT-2 write a *New Yorker* article? That was my solipsistic response on hearing of the artificial author's doomsday potential. What if OpenAI fine-tuned GPT-2 on *The New Yorker's* digital archive (please, don't call it a "data set")—millions of polished and fact-checked words, many written by masters of the literary art. Could the machine learn to write well enough for *The New Yorker*? Could it write this article for me? The fate of civilization may not hang on the answer to that question, but mine might.

I raised the idea with OpenAI. Greg Brockman, the C.T.O., offered to fine-tune the full-strength version of GPT-2 with the magazine's archive. He promised to use the archive only for the purposes of this experiment. The corpus employed for the fine-tuning included all nonfiction work published since 2007 (but no fiction, poetry, or cartoons), along with some digitized classics going back to the nineteen-sixties. A human would need almost two weeks of 24/7 reading to get through it all; Jeff Wu, who oversaw the project, told me that the A.I. computed the archive in under an hour—a mere after-dinner macaron compared with its All-U-Can-Eat buffet of Reddit training data, the computing of which had required almost an entire "petaflop-per-second day"—a thousand trillion operations per second, for twenty-four hours.

Read Predicted Text >

OpenAI occupies a historic three-story loft building, originally built as a luggage factory in 1903, three years before the earthquake and fire that consumed much of San Francisco. It sits at the corner of Eighteenth and Folsom Streets, in the city's Mission District. There are a hundred employees, most of them young and well educated, who have an air of higher purpose about them. The staff aren't merely trying to invent a superintelligent machine. They're also devoted to protecting us from superintelligence, by trying to formulate safety standards for the technology which are akin to the international protocols that govern nuclear materials like yellowcake uranium. What might be the safest course of all—to stop trying to build a machine as intelligent as we are—isn't part of OpenAI's business plan.

Dario Amodei, the research director, conducted the demonstration of the *New Yorker*-trained A.I. for me, in a glass-walled conference room on the first floor, using an OpenAI laptop. Amodei, thirty-six, has a Ph.D. in computational neuroscience from Princeton and did a postdoc at Stanford. He has boyishly curly hair that he has the habit of twisting around a finger while he talks.

In fine-tuning GPT-2 for the purposes of this article, the neural net categorized distinctive aspects of *New Yorker* prose—the words its writers tended to favor, the magazine's rhythms, its distinctive style of narrative rhetoric, its voice—and the learning algorithm used these data to automatically adjust the neural net's settings, so that its predictions leaned toward *New Yorker* locutions. We were about to find out how well it worked. I had butterflies. It felt as if we were lighting a fuse but didn't know where it led.

The interface on the laptop screen was deceptively simple: a window where you could paste or write in prompts, and four slider controls on the left. Two adjusted the output: how many words the machine wrote each time the user pressed the refresh button. A third was for "generativity"—establishing how jiggy the A.I. got with its suggestions. The fourth slider adjusted the "nucleus sampling"—the size of the pool of words from which the machine drew in generating text. The refresh button was OpenAI's logo, which looked to me like a mandala, the kind associated with alchemists in the Middle Ages.

I had selected in advance some openings of favorite *New Yorker* articles to prompt the machine. We started by pasting the first paragraph of Lillian Ross's 1950 Profile of Ernest Hemingway into the prompt window. It begins:

> Ernest Hemingway, who may well be the greatest living American novelist and short-story writer, rarely comes to New York. He spends most of his time on a farm, the Finca Vigia, nine miles outside Havana, with his wife, a domestic staff of nine, fifty-two cats, sixteen dogs, a couple of hundred pigeons, and three cows.

Amodei allowed me to click the mandala. After a second-long pause, the generator kicked in, and line after line of Robo-Ross's prose began to fill the screen:

```
I walked up the path to the fence and saw a plump dog, that had been a
common visitor to the Finca Vigia before the war, galloping up a path to
the main building, with a tiny cow of the same name standing by her side.
```

> There was a puddle of red gravy in the front yard, and Hemingway sat down in a lawn chair.

On first reading this passage, my brain ignored what A.I. researchers call "world-modelling failures"—the tiny cow and the puddle of red gravy. Because I had never encountered a prose-writing machine even remotely this fluent before, my brain made an assumption—any human capable of writing this well would know that cows aren't tiny and red gravy doesn't puddle in people's yards. And because GPT-2 was an inspired mimic, expertly capturing *The New Yorker's* cadences and narrative rhythms, it sounded like a familiar, trusted voice that I was inclined to believe. In fact, it sounded sort of like my voice.

I recalled a well-known experiment conducted in 1966 by Joseph Weizenbaum, a German-born professor at M.I.T. who was a pioneer of artificial intelligence. In the experiment, a primitive (by today's standards) chatbot that Weizenbaum named ELIZA—for the George Bernard Shaw ingénue—responded, in writing, to statements by the study's subjects. The bot was programmed to answer in the style of a stereotypical psychotherapist, with questions such as "How does that make you feel?" To Weizenbaum's surprise, the "patients," even when they knew ELIZA was a bot, began revealing intimate details of their lives; his secretary at M.I.T. asked him to leave the room so that she could communicate freely with ELIZA.

I clicked the mandala again, and the machine continued writing its Daliesque version of Ross's Profile, using, in addition to the first prompt, the prose it had already generated to generate from:

> He was wearing a tweed suit, over a shiny sweater, and his black hair was brushed back. He had a red beard and wore his waistcoat in an overcoat with the body of a ship, three broad belts of colorful chain-link, a pair of capacious rectangular eyeglasses, and a silk tie. "Gouging my eye," he said, in Italian, saying that he had caused himself that terrible scar, "the surgeon said it wasn't that bad." When he was very young, he said, he

started smoking but didn't find it very pleasant. The cigarette burns in his hands and wrists were so bad that he had to have his face covered.

Three chain-link belts? Oddly, a belt does come up later in Ross's article, when she and Hemingway go shopping. So do eyeglasses, and cigarettes, and Italy. GPT-2 hadn't "read" the article—it wasn't included in the training data—yet it had somehow alighted on evocative details. Its deep learning obviously did not include the ability to distinguish nonfiction from fiction, though. Convincingly faking quotes was one of its singular talents. Other things often sounded right, though GPT-2 suffered frequent world-modelling failures—gaps in the kind of commonsense knowledge that tells you overcoats aren't shaped like the body of a ship. It was as though the writer had fallen asleep and was dreaming.

Amodei explained that there was no way of knowing why the A.I. came up with specific names and descriptions in its writing; it was drawing from a content pool that seemed to be a mixture of *New Yorker*-ese and the machine's Reddit-based training. The mathematical calculations that resulted in the algorithmic settings that yielded GPT-2's words are far too complex for our brains to understand. In trying to build a thinking machine, scientists have so far succeeded only in reiterating the mystery of how our own brains think.

Because of the size of the Reddit data set necessary to train GPT-2, it is impossible for researchers to filter out all the abusive or racist content, although OpenAI had caught some of it. However, Amodei added, "it's definitely the case, if you start saying things about conspiracy theories, or prompting it from the Stormfront Web site—it knows about that." Conspiracy theories, after all, are a form of pattern recognition, too; the A.I. doesn't care if they're true or not.

Each time I clicked the refresh button, the prose that the machine generated became more random; after three or four tries, the writing had drifted far from the original prompt. I found that by adjusting the slider to limit the amount of

text GPT-2 generated, and then generating again so that it used the language it had just produced, the writing stayed on topic a bit longer, but it, too, soon devolved into gibberish, in a way that reminded me of HAL, the superintelligent computer in "2001: A Space Odyssey," when the astronauts begin to disconnect its mainframe-size artificial brain.

An hour or so later, after we had tried opening paragraphs of John Hersey's "Hiroshima" and Truman Capote's "In Cold Blood," my initial excitement had curdled into queasiness. It hurt to see the rules of grammar and usage, which I have lived my writing life by, mastered by an idiot savant that used math for words. It was sickening to see how the slithering machine intelligence, with its ability to take on the color of the prompt's prose, slipped into some of my favorite paragraphs, impersonating their voices but without their souls.

Read Predicted Text >

There are many positive services that A.I. writers might provide. I.B.M. recently débuted an A.I. called Speech by Crowd, which it has been developing with Noam Slonim, an Israeli I.B.M. Research Fellow. The A.I. processed almost two thousand essays written by people on the topic "Social Media Brings More Harm Than Good" and, using a combination of rules and deep learning, isolated the best arguments on both sides and summarized them in a pair of three-to-five-paragraph, op-ed-style essays, one pro ("Social media creates a platform to support freedom of speech, giving individuals a platform to voice their opinions and interact with like-minded individuals") and one con ("The opinion of a few can now determine the debate, it causes polarized discussions and strong feelings on non-important subjects"). The essays I read were competent, but most seventh graders with social-media experience could have made the same arguments less formulaically.

Slonim pointed to the rigid formats used in public-opinion surveys, which rely on questions the pollsters think are important. What, he asked, if these surveys

came with open-ended questions that allowed respondents to write about issues that concern them, in any form. Speech by Crowd can "read" all the answers and digest them into broader narratives. "That would disrupt opinion surveys," Slonim told me.

At Narrative Science, in Chicago, a company co-founded by Kristian Hammond, a computer scientist at Northwestern, the main focus is using a suite of artificial-intelligence techniques to turn data into natural language and narrative. The company's software renders numerical information about profit and loss or manufacturing operations, for example, as stories that make sense of patterns in the data, a tedious task formerly accomplished by people poring over numbers and churning out reports. "I have data, and I don't understand the data, and so a system figures out what I need to hear and then turns it into language," Hammond explained. "I'm stunned by how much data we have and how little of it we use. For me, it's trying to build that bridge between data and information."

One of Hammond's former colleagues, Jeremy Gilbert, now the director of strategic initiatives at the Washington *Post*, oversees Heliograf, the *Post's* deep-learning robotic newshound. Its purpose, he told me, is not to replace journalists but to cover data-heavy stories, some with small but highly engaged audiences—a high-school football game ("The Yorktown Patriots triumphed over the visiting Wilson Tigers in a close game on Thursday, 20–14," the A.I. reported), local election results, a minor commodities-market report—that newspapers lack the manpower to cover, and others with much broader reach, such as national elections or the Olympics. Heliograf collects the data and applies them to a particular template—a spreadsheet for words, Gilbert said—and an algorithm identifies the decisive play in the game or the key issue in the election and generates the language to describe it. Although Gilbert says that no freelancer has lost a gig to Heliograf, it's not hard to imagine that the high-school stringer who once started out on the varsity beat will be coding instead.

Read Predicted Text ›

OpenAI made it possible for me to log in to the *New Yorker* A.I. remotely. On the flight back to New York, I put some of my notes from the OpenAI visit into GPT-2 and it began making up quotes for Ilya Sutskever, the company's chief scientist. The machine appeared to be well informed about his groundbreaking research. I worried that I'd forget what he really said, because the A.I. sounded so much like him, and that I'd inadvertently use in my article the machine's fake reporting, generated from my notes. ("We can make fast translations but we can't really solve these conceptual questions," one of GPT-2's Sutskever quotes said. "Maybe it is better to have one person go out and learn French than to have an entire computer-science department.") By the time I got home, the A.I. had me spooked. I knew right away there was no way the machine could help me write this article, but I suspected that there were a million ways it could screw me up.

I sent a sample of GPT-2's prose to Steven Pinker, the Harvard psycholinguist. He was not impressed with the machine's "superficially plausible gobbledygook," and explained why. I put some of his reply into the generator window, clicked the mandala, added synthetic Pinker prose to the real thing, and asked people to guess where the author of "The Language Instinct" stopped and the machine took over.

In the text below, click where you think Pinker's text ends and GPT-2's begins:

Being amnesic for how it began a phrase or sentence, it won't consistently complete it with the necessary agreement and concord—to say nothing of semantic coherence. And this reveals the second problem: real language does not consist of a running monologue that sounds sort of like English. It's a way of expressing ideas, a mapping from meaning to sound or text. To put it crudely, speaking or writing is a box whose input is a meaning plus a communicative intent, and whose output is a string of words; comprehension is a box with the opposite information flow. What is essentially wrong with this perspective is that it assumes that meaning and intent

> are inextricably linked. Their separation, the learning scientist Phil Zuckerman has argued, is an illusion that we have built into our brains, a false sense of coherence.

That's Pinker through "information flow." (There is no learning scientist named Phil Zuckerman, although there is a sociologist by that name who specializes in secularity.) Pinker is right about the machine's amnesic qualities—it can't develop a thought, based on a previous one. It's like a person who speaks constantly but says almost nothing. (Political punditry could be its natural domain.) However, almost everyone I tried the Pinker Test on, including Dario Amodei, of OpenAI, and Les Perelman, of Project BABEL, failed to distinguish Pinker's prose from the machine's gobbledygook. The A.I. had them Pinkered.

GPT-2 was like a three-year-old prodigiously gifted with the illusion, at least, of college-level writing ability. But even a child prodigy would have a goal in writing; the machine's only goal is to predict the next word. It can't sustain a thought, because it can't think causally. Deep learning works brilliantly at capturing all the edgy patterns in our syntactic gymnastics, but because it lacks a pre-coded base of procedural knowledge it can't use its language skills to reason or to conceptualize. An intelligent machine needs both kinds of thinking.

"It's a card trick," Kris Hammond, of Narrative Science, said, when I sent him what I thought were some of the GPT-2's better efforts. "A very sophisticated card trick, but at heart it's still a card trick." True, but there are also a lot of tricks involved in writing, so it's hard to find fault with a fellow-mountebank on that score.

One can envision machines like GPT-2 spewing superficially sensible gibberish, like a burst water main of babble, flooding the Internet with so much writing that it would soon drown out human voices, and then training on its own meaningless prose, like a cow chewing its cud. But composing a long discursive

narrative, structured in a particular way to advance the story, was, at least for now, completely beyond GPT-2's predictive capacity.

However, even if people will still be necessary for literary production, day by day, automated writers like GPT-2 will do a little more of the writing that humans are now required to do. People who aren't professional writers may be able to avail themselves of a wide range of products that will write e-mails, memos, reports, and speeches for them. And, like me writing "I am proud of you" to my son, some of the A.I.'s next words might seem superior to words you might have thought of yourself. But what else might you have thought to say that is not computable? That will all be lost.

Read Predicted Text >

Before my visit to OpenAI, I watched a lecture on YouTube that Ilya Sutskever had given on GPT-2 in March, at the Computer History Museum, in Mountain View, California. In it, he made what sounded to me like a claim that GPT-2 itself might venture, if you set the generativity slider to the max. Sutskever said, "If a machine like GPT-2 could have enough data and computing power to perfectly predict the next word, that would be the equivalent of understanding."

At OpenAI, I asked Sutskever about this. "When I said this statement, I used 'understanding' informally," he explained. "We don't really know what it means for a system to understand something, and when you look at a system like this it can be genuinely hard to tell. The thing that I meant was: If you train a system which predicts the next word well enough, then it ought to understand. If it doesn't predict it well enough, its understanding will be incomplete."

However, Sutskever added, "researchers can't disallow the possibility that we will reach understanding when the neural net gets as big as the brain."

The brain is estimated to contain a hundred billion neurons, with trillions of connections between them. The neural net that the full version of GPT-2 runs on has about one and a half billion connections, or "parameters." At the current rate at which compute is growing, neural nets could equal the brain's raw processing capacity in five years. To help OpenAI get there first, Microsoft announced in July that it was investing a billion dollars in the company, as part of an "exclusive computing partnership." How its benefits will be "distributed as widely as possible" remains to be seen. (A spokesperson for OpenAI said that "Microsoft's investment doesn't give Microsoft control" over the A.I. that OpenAI creates.)

David Ferrucci, the only person I tried the Pinker Test on who passed it, said, "Are we going to achieve machine understanding in a way we have hoped for many years? Not with these machine-learning techniques. Can we do it with hybrid techniques?" (By that he meant ones that combine knowledge-based systems with machine-learning pattern recognition.) "I'm betting yes. That's what cognition is all about, a hybrid architecture that combines different classes of thinking."

What if some much later iteration of GPT-2, far more powerful than this model, could be hybridized with a procedural system, so that it would be able to write causally and distinguish truth from fiction and at the same time draw from its well of deep learning? One can imagine a kind of Joycean superauthor, capable of any style, turning out spine-tingling suspense novels, massively researched biographies, and nuanced analyses of the Israeli-Palestinian conflict. Humans would stop writing, or at least publishing, because all the readers would be captivated by the machines. What then?

GPT-2, prompted with that paragraph, predicted the next sentence: `"In a way, the humans would be making progress."` ♦

Methodology:
For the purposes of this article, The New Yorker granted OpenAI access to all nonfiction work published in the

magazine since 2007 (but not fiction, poetry, or cartoons), along with some digitized classics going back to the nineteen-sixties. Using this corpus, OpenAI fine-tuned the full-strength version of GPT-2, to be used only for the purposes of this experiment. OpenAI made it possible for The New Yorker to log in to the New Yorker A.I. remotely.

We fed text from the end of each section in this article into the New Yorker A.I., and it generated what text would come next, including any quotations. The generative settings were consistent for each output, but we adjusted the slider for response length. In each case, we generated more than one response and selected the predicted text that follows each section in the article.

Published in the print edition of the October 14, 2019, issue, with the headline "The Next Word."

John Seabrook has been a contributor to The New Yorker since 1989 and became a staff writer in 1993. He has published four books, including, most recently, "The Song Machine: Inside the Hit Factory."

More: Writing Artificial Intelligence (A.I.) Google Neuroscience Linguistics The New Yorker Lillian Ross

READ MORE

A REPORTER AT LARGE

The Age of Robot Farmers

Picking strawberries takes speed, stamina, and skill. Can a robot do it?

By John Seabrook

DEPT. OF TECHNOLOGY

Network Insecurity

Are we losing the battle against cyber crime?

By John Seabrook

VIDEO

The Desire to Own Nothing

The filmmaker Sindha Agha explores her relationship to "stuff" and how a personal trauma led to an emotional version of the Kondo method.