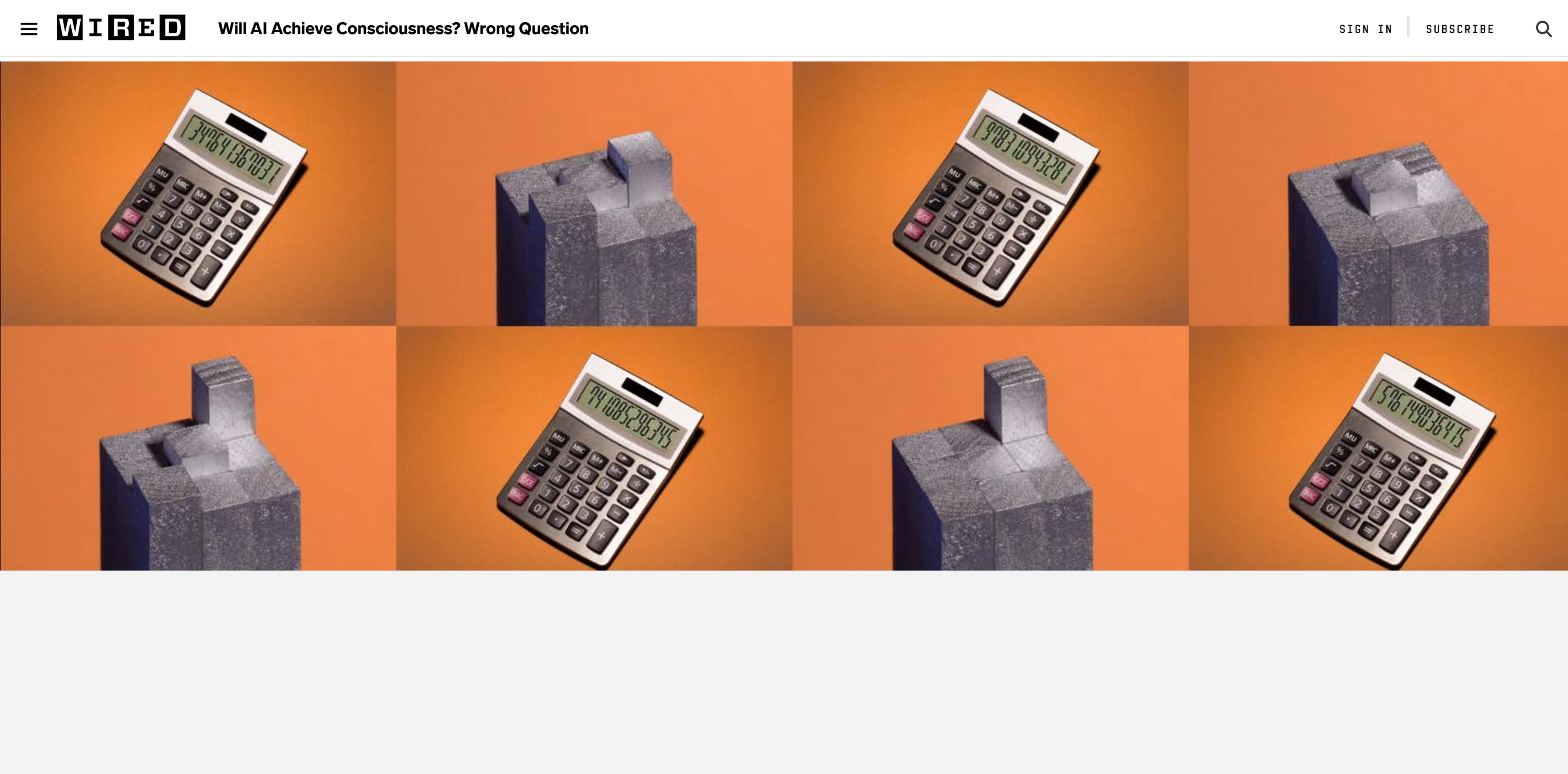


DANIEL C. DENNETT BACKCHANNEL 02.19.2019 07:00 AM

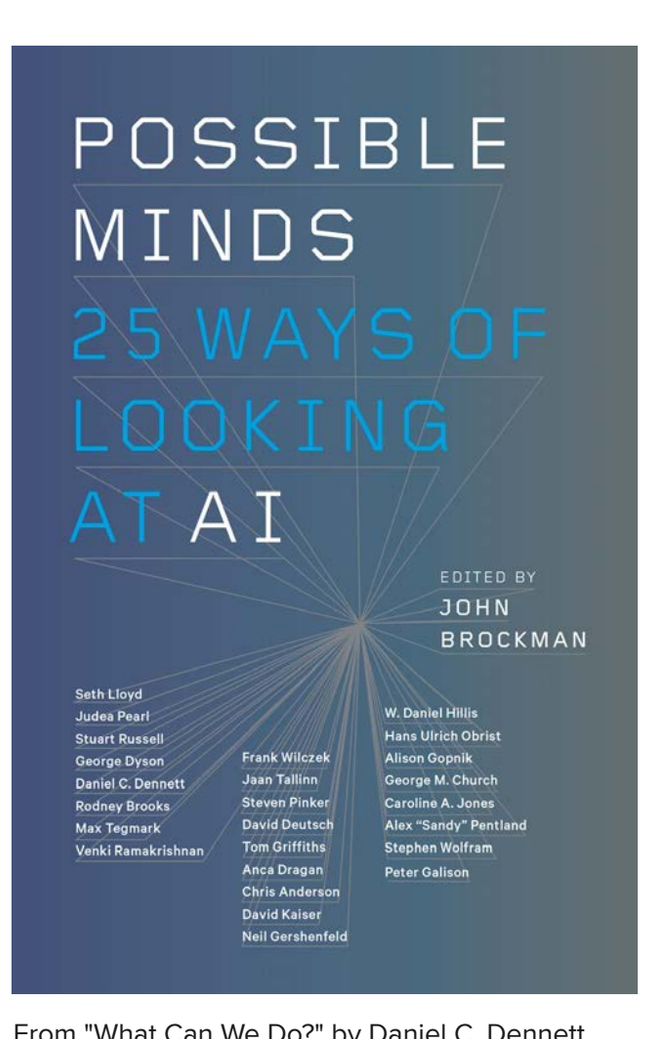
Will AI Achieve Consciousness? Wrong Question

We should not be creating conscious, humanoid agents but an entirely new sort of entity, rather like oracles, with no conscience, no fear of death, no distracting loves and hates.

WHEN NORBERT WIENER, the father of cybernetics, wrote his book *The Human Use of Human Beings* in 1950, vacuum tubes were still the primary electronic building blocks, and there were only a few actual computers in operation.

But he imagined the future we now contend with in impressive detail and with few clear mistakes. More than any other early philosopher of artificial intelligence, he recognized that AI would not just imitate—and replace—human beings in many intelligent activities but would change human beings in the process. "We are but whirlpools in a river of ever-flowing water," he wrote. "We are not stuff that abides, but patterns that perpetuate themselves."

When attractive opportunities abound, for instance, we are apt to be willing to pay a little and accept some small, even trivial cost of doing business for access to new powers. And pretty soon we become so dependent on our new tools that we lose the ability to thrive without them. Options become obligatory.



From "What Can We Do?" by Daniel C. Dennett. Adapted from Possible Minds: *Twenty-Five Ways of Looking at AI*, edited by John Brockman, published by Penguin Press, an imprint of Penguin Publishing Group, a division of Penguin Random House LLC. Copyright © 2019 by John Brockman. PENGUIN PRESS

It's an old, old story, with many well-known chapters in evolutionary history. Most mammals can synthesize their own vitamin C, but primates, having opted for a diet composed largely of fruit, lost the innate ability. The self-perpetuating patterns that we call human beings are now dependent on clothes, cooked food, vitamins, vaccinations, credit cards, smartphones, and the internet. And—tomorrow if not already today—AI.

Wiener foresaw several problems with this incipient state of affairs that Alan Turing and other early AI optimists largely overlooked. The real danger, he said, is

that such machines, though helpless by themselves, may be used by a human being or a block of human beings to increase their control over the rest of the race or that political leaders may attempt to control their populations by means not of machines themselves but through political techniques as narrow and indifferent to human possibility as if they had, in fact, been conceived mechanically.

Sure enough, these dangers are now pervasive.

In media, for instance, the innovations of digital audio and video let us pay a small price (in the eyes of audiophiles and film lovers) when we abandon analog formats, and in return provide easy—all too easy?—reproduction of recordings with almost perfect fidelity.

But there is a huge hidden cost. Orwell's Ministry of Truth is now a practical possibility. AI techniques for creating all-but-undetectable forgeries of "recordings" of encounters are now becoming available, which will render obsolete the tools of investigation we have come to take for granted in the past 150 years.

Will we simply abandon the brief Age of Photographic Evidence and return to the earlier world in which human memory and trust provided the gold standard, or will we develop new techniques of defense and offense in the arms race of truth? (We can imagine a return to *analog* film-exposed-to-light, kept in "tamper-proof" systems until shown to juries, etc., but how long would it be before somebody figured out a way to infect such systems with doubt?

One of the disturbing lessons of recent experience is that the task of destroying a reputation for credibility is much less expensive than the task of protecting such a reputation. Wiener saw the phenomenon at its most general: "In the long run, there is no distinction between arming ourselves and arming our enemies." The information age is also the disinformation age.

What can we do? A key phrase, it seems to me, is Wiener's almost offhand observation, above, that "these machines" are "helpless by themselves." As I have been arguing recently, we're making tools, not colleagues, and the great danger is not appreciating the difference, which we should strive to accentuate, marking and defending it with political and legal innovations.

AI in its current manifestations is parasitic on human intelligence. It quite indiscriminately gorges on whatever has been produced by human creators and extracts the patterns to be found there—including some of our most pernicious habits. These machines do not (yet) have the goals or strategies or capacities for self-criticism and innovation to permit them to transcend their databases by reflectively thinking about their own thinking and their own goals.

They are, as Wiener says, helpless, not in the sense of being *shackled* agents or *disabled* agents but in the sense of not being agents at all—not having the capacity to be "moved by reasons" (as Kant put it) presented to them. It is important that we keep it that way, which will take some doing.

In the long term, "strong AI," or general artificial intelligence, is possible in principle but not desirable (more on this later). The far more constrained AI that's practically possible today is not necessarily evil. But it poses its own set of dangers—chiefly that it might be mistaken for strong AI!

Most Popular

GEAR
One Woman's High-Touch Bid to Upend the Sex-Toy Industry
LUX ALPTRAUM

CULTURE
Television Like *The Boys* Is Destroying You
JASON KEHE

BUSINESS
Clarence Thomas Wants to Rethink Internet Speech. Be Afraid
STEVEN LEVY

GEAR
The iPhone 12 Ships Without a Charger. Will It Curb E-Waste?
JULIAN CHOKKATTU

THE GAP BETWEEN today's systems and the science-fictional systems dominating the popular imagination is still huge, though many folks, both lay and expert, manage to underestimate it. Let's consider IBM's Watson, which can stand as a worthy landmark for our imaginations for the time being.

It is the result of a very large-scale R&D process extending over many person-centuries of intelligent design, and it uses thousands of times more energy than a human brain. Its victory in *Jeopardy!* was a genuine triumph, made possible by the formulaic restrictions of the *Jeopardy!* rules, but in order for it to compete, even these rules had to be revised (one of those trade-offs: you give up a little versatility, a little humanity, and get a crowd-pleasing show).

Watson is not good company, in spite of misleading ads from IBM that suggest a general conversational ability, and turning Watson into a plausibly multidimensional *agent* would be like turning a hand calculator into Watson. Watson could be a useful core faculty for such an agent, but more like a cerebellum or an amygdala than a mind—at best, a special-purpose subsystem that could play a big supporting role, but not remotely up to the task of framing purposes and plans and building insightfully on its conversational experiences.

Why would we want to create a thinking, creative agent out of Watson? Perhaps Turing's brilliant idea of an operational test—the famous Turing test—has lured us into a trap: the quest to create at least the illusion of a real person behind the screen, bridging the "uncanny valley."

The danger here is that ever since Turing posed his challenge—which was, after all, a challenge to *fool* the judges—AI creators have attempted to paper over the valley with cutesy humanoid touches, Disneyfication effects that will enchant and disarm the uninitiated. Joseph Weizenbaum's ELIZA, a very early chatbot, was the pioneer example of such superficial illusion making, and it was his dismay at the ease with which his laughably simple and shallow program could persuade people they were having a serious heart-to-heart conversation that first sent him on his mission.

He was right to be worried. If there is one thing we have learned from the restricted Turing test competitions for the annual Loebner Prize, it is that even very intelligent people who aren't tuned in to the possibilities and shortcuts of computer programming are readily taken in by simple tricks.

The attitudes of people in AI toward these methods of dissembling at the "user interface" have ranged from contempt to celebration, with a general appreciation that the tricks are not deep but can be potent. One shift in attitude that would be very welcome is a candid acknowledgment that humanoid embellishments are *false advertising*—something to condemn, not applaud.

How could that be accomplished? Once we recognize that people are starting to make life-or-death decisions largely on the basis of "advice" from AI systems whose inner operations are unfathomable in practice, we can see a good reason why those who in any way encourage people to put more trust in these systems than they warrant should be held morally and legally accountable.

Most Popular

GEAR
One Woman's High-Touch Bid to Upend the Sex-Toy Industry
LUX ALPTRAUM

CULTURE
Television Like *The Boys* Is Destroying You
JASON KEHE

BUSINESS
Clarence Thomas Wants to Rethink Internet Speech. Be Afraid
STEVEN LEVY

GEAR
The iPhone 12 Ships Without a Charger. Will It Curb E-Waste?
JULIAN CHOKKATTU

AI systems are very powerful tools—so powerful that even experts will have good reason not to trust their own judgment over the "judgments" delivered by their tools. But then, if these tool users are going to benefit, financially or otherwise, from driving these tools through terra incognita, they need to make sure they know how to do this responsibly, with maximum control and justification.

Licensing and bonding the operators of these systems, just as we license pharmacists, crane operators, and other specialists whose errors and misjudgments can have dire consequences, could, with pressure from insurance companies and other underwriters, oblige creators of AI systems to go to extraordinary lengths to search for and reveal weaknesses and gaps in their products, and to train those entitled to operate them to watch out for them.

One can imagine a sort of inverted Turing test in which the judge is on trial; until he or she can spot the weaknesses, the overstepped boundaries, the gaps in a system, no license to operate will be issued. The mental training required to achieve certification as a judge will be demanding. The urge to attribute humanlike powers of thought to an object, our normal tactic whenever we encounter what seems to be an intelligent agent, is almost overpoweringly strong.

Indeed, the capacity to resist the allure of treating an apparent person as a person is an ugly talent, reeking of racism or species-ism. Many people would find the cultivation of such a ruthlessly skeptical approach morally repugnant, and we can anticipate that even the most proficient system users would occasionally succumb to the temptation to "befriend" their tools, if only to assuage their discomfort with the execution of their duties.

No matter how scrupulously the AI designers launder the phony "human" touches out of their wares, we can expect a flourishing of shortcuts, workarounds and tolerated distortions of the actual "comprehension" of both the systems and their operators. The comically long lists of known side effects of new drugs advertised on television will be dwarfed by the obligatory revelations of the sorts of questions that *cannot* be responsibly answered by particular systems, with heavy penalties for manufacturers who "overlook" flaws in their products. (It is widely noted that a considerable part of the growing economic inequality in today's world is due to the wealth accumulated by digital entrepreneurs; we should enact legislation that puts their deep pockets in escrow for the public good.)

WE DON'T NEED artificial conscious agents. There is a surfeit of natural conscious agents, enough to handle whatever tasks should be reserved for such special and privileged entities. We need intelligent tools. Tools do not have rights and should not have feelings that could be hurt or be able to respond with resentment to "abuses" rained on them by inept users.

One of the reasons for not making artificial conscious agents is that, however autonomous they might become (and in principle they can be as autonomous, as self-enhancing or self-creating, as any person), they would not—without special provision, which might be waived—share with us natural conscious agents our vulnerability or our mortality.

I once posed a challenge to students in a seminar at Tufts on artificial agents and autonomy: Give me the specs for a robot that could sign a binding contract with you—not as a surrogate for some human owner but on its own. This isn't a question of getting it to understand the clauses or manipulate a pen on a piece of paper but of having and *deserving* legal status as a morally responsible agent. Small children can't sign such contracts, nor can those disabled people whose legal status requires them to be under the care and responsibility of guardians of one sort or another.

The problem for robots who might want to attain such an exalted status is that, like Superman, they are too invulnerable to be able to make a credible promise. If they were to renege, what would happen? What would be the penalty for promise breaking? Being locked in a cell or, more plausibly, dismantled? Being locked up is barely an inconvenience for an AI unless we first install artificial wanderlust that cannot be ignored or disabled by the AI on its own (and it would be systematically difficult to make this a foolproof solution, given the presumed cunning and self-knowledge of the AI); and dismantling an AI (either a robot or a bedridden agent like Watson) is not killing it if the information stored in its design and software is preserved.

The very ease of digital recording and transmitting— the breakthrough that permits software and data to be, in effect, immortal—removes robots from the world of the vulnerable (at least robots of the usually imagined sorts, with digital software and memories). If this isn't obvious, think about how human morality would be affected if we could make "backups" of people every week, say. Diving headfirst on Saturday off a high bridge without benefit of a bungee cord would be a rush that you wouldn't remember when your Friday night backup was put online Sunday morning, but you could enjoy the videotape of your apparent demise thereafter.

So what we are creating are not—should not be—conscious, humanoid agents but an entirely new sort of entity, rather like oracles, with no conscience, no fear of death, no distracting loves and hates, no personality (but all sorts of foibles and quirks that would no doubt be identified as the "personality" of the system): boxes of truths (if we're lucky) almost certainly contaminated with a scattering of falsehoods.

It will be hard enough learning to live with them without distracting ourselves with fantasies about the Singularity in which these AIs will enslave us, literally. The *human* use of human beings will soon be changed—once again—forever, but we can take the tiller and steer between some of the hazards if we take responsibility for our trajectory.

Daniel C. Dennett is the Austin B. Fletcher professor of philosophy and codirector of the Center for Cognitive Studies at Tufts University.

From "What Can We Do?" by Daniel C. Dennett. Adapted from *Possible Minds: Twenty-Five Ways of Looking at AI*, edited by John Brockman, published by Penguin Press, an imprint of Penguin Publishing Group, a division of Penguin Random House LLC. Copyright © 2019 by John Brockman.
