Dan Dennett *is the philosopher of choice in the AI community. He is perhaps best known in cognitive science for his concept of intentional systems and his model of human consciousness, which sketches a computational architecture for realizing the stream of consciousness in the massively parallel cerebral cortex. That uncompromising computationalism has been opposed by philosophers such as John Searle, David Chalmers, and the late Jerry Fodor, who have protested that the most important aspects of consciousness—intentionality and subjective qualia—cannot be computed.*

Twenty-five years ago, I was visiting Marvin Minsky, one of the original AI pioneers, and asked him about Dan. "He's our best current philosopher—the next Bertrand Russell," said Marvin, adding that unlike traditional philosophers, Dan was a student of neuroscience, linguistics, artificial intelligence, computer science, and psychology: "He's redefining and reforming the role of the philosopher. Of course, Dan doesn't understand my Society-of-Mind theory, but nobody's perfect."

Dan's view of the efforts of AI researchers to create superintelligent AIs is relentlessly levelheaded. What, me worry? In this essay, he reminds us that AIs, above all, should be regarded—and treated—as tools and not as humanoid colleagues.

He has been interested in information theory since his graduate school days at Oxford. In fact, he told me that early in his career he was keenly interested in writing a book about Wiener's cybernetic ideas. As a thinker who embraces the scientific method, one of his charms is his willingness to be wrong. Of a recent piece entitled "What Is Information?" he has announced, "I stand by it, but it's under revision. I'm already moving beyond it and realizing there's a better way of tackling some of these issues." He will most likely remain cool and collected on the subject of AI research, although he has acknowledged, often, that his own ideas evolve—as anyone's ideas should.

WHAT CAN WE DO?

Daniel C. Dennett

Daniel C. Dennett *is University Professor and Austin B. Fletcher Professor of Philosophy and director of the Center for Cognitive Studies at Tufts University. He is the author of a dozen books, including* Consciousness Explained *and, most recently,* From Bacteria to Bach and Back: The Evolution of Minds.

Many have reflected on the irony of reading a great book when you are too young to appreciate it. Consigning a classic to the *already read* stack and thereby insulating yourself against any further influence while gleaning only a few ill-understood ideas from it is a recipe for neglect that is seldom benign. This struck me with particular force when I reread *The Human Use of Human Beings* more than sixty years after my juvenile encounter. We should all make it a regular practice to reread books from our youth, where we are apt to discover clear previews of some of our own later "discoveries" and "inventions," along with a wealth of insights to which we were bound to be impervious until our minds had been torn and tattered, exercised and enlarged by confrontations with life's problems.

Writing at a time when vacuum tubes were still the primary electronic building blocks and there were only a few actual computers in operation, Norbert Wiener imagined the future we now contend with in impressive detail and with few clear mistakes. Alan Turing's famous 1950 article "Computing Machinery and Intelligence," in the philosophy journal *Mind,* foresaw the development of AI, and so did Wiener, but Wiener saw farther and deeper, recognizing that AI would not just imitate—and replace—human beings in many intelligent activities but change human beings in the process.

> We are but whirlpools in a river of ever-flowing water. We are not stuff that abides, but patterns that perpetuate themselves. (p. 96)

When that was written, it could be comfortably dismissed as yet another bit of Heraclitean overstatement. Yeah, yeah, you can never step in the same river twice. But it contains the seeds of the revolution in outlook. Today we know how to think about complex adaptive systems, strange attractors, extended minds, and homeostasis, a change in perspective that promises to erase the "explanatory gap"[1] between mind and mechanism, spirit and matter, a gap that is still ardently defended by latter-day Cartesians who cannot bear the thought that we—we *ourselves*—are self-perpetuating patterns of information-bearing matter, not "stuff that abides." Those patterns are remarkably resilient and self-restoring but at the same time protean, opportunistic, selfish exploiters of whatever new is available to harness in their quest for perpetuation. And here is where things get dicey, as Wiener recognized. When attractive opportunities abound, 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.

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. We are now obligate ingesters of vitamin C, but not

---

[1] Joseph Levine, "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64, pp. 354-61 (1983).

obligate frugivores like our primate cousins, since we have opted for technology that allows us to make, and take, vitamins as needed. 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 the problems that Turing and the other optimists have 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. (p. 181)

The power, he recognized, lay primarily in the algorithms, not the hardware they run on, although the hardware of today makes practically possible algorithms that would have seemed preposterously cumbersome in Wiener's day. What can we say about these "techniques" that are "narrow and indifferent to human possibility"? They have been introduced again and again, some obviously benign, some obviously dangerous, and many in the omnipresent middle ground of controversy.

Consider a few of the skirmishes. My late friend Joe Weizenbaum, Wiener's successor as MIT's Jeremiah of hi-tech, loved to observe that credit cards, whatever their virtues, also provided an inexpensive and almost foolproof way for the government, or corporations, to track the travels and habits and desires of individuals. The anonymity of cash has been largely underappreciated, except by drug dealers and other criminals, and now it may be going extinct. This may make money laundering a more difficult technical challenge in the future, but the AI pattern finders arrayed against it have the side effect of making us all more transparent to any "block of human beings" that may "attempt to control" us.

Looking to the arts, the innovation of digital audio and video recording lets us pay a small price (in the eyes of all but the most ardent audiophiles and film lovers) when we abandon analog formats, and in return provides easy—all too easy?—reproduction of artworks 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 last hundred and fifty 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." (p. 129) The Information Age is also the Dysinformation Age.

What can we do? We need to rethink our priorities with the help of the passionate but flawed analyses of Wiener, Weizenbaum, and the other serious critics of our technophilia. 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.

Perhaps the best way to see what is being missed is to note that Alan Turing himself suffered an entirely understandable failure of imagination in his formulation of the famous Turing Test. As everyone knows, it is an adaptation of his "imitation game," in which a man, hidden from view and communicating verbally with a judge, tries to convince the judge that he is in fact a woman, while a woman, also hidden and communicating with the judge, tries to convince the judge that she is the woman. Turing reasoned that this would be a demanding challenge for a man (or for a woman pretending to be a man), exploiting a wealth of knowledge about how the other sex thinks and acts, what they tend to favor or ignore. Surely (*ding!*)[2], any man who could beat a woman at being perceived to be a woman would be an intelligent agent. What Turing did not foresee is the power of deep-learning AI to acquire this wealth of information in an exploitable form *without having to understand it.* Turing imagined an astute and imaginative (and hence conscious) agent who cunningly designed his responses based on his detailed "theory" of what women are likely to do and say. Top-down intelligent design, in short. He certainly didn't think that a man, winning the imitation game, would somehow *become* a woman; he imagined that there would still be a man's consciousness guiding the show. The hidden premise in Turing's almost-argument was: Only a conscious, intelligent *agent* could devise and control a winning strategy in the imitation game. And so it was persuasive to Turing (and others, including me, still a stalwart defender of the Turing Test) to argue that a "computing machine" that could pass as human in a contest with a human might not be conscious in just the way a human being is, but would nevertheless have to be a conscious agent of *some* kind. I think this is still a defensible position—the only defensible position—but you have to understand how resourceful and ingenious a judge would have to be to expose the shallowness of the façade that a deep-learning AI (a tool, not a colleague) could present.

What Turing didn't foresee is the uncanny ability of superfast computers to sift mindlessly through Big Data, of which the Internet provides an inexhaustible supply, finding probabilistic patterns in human activity that could be used to pop "authentic"-seeming responses into the output for almost any probe a judge would think to offer. Wiener also underestimates this possibility, seeing the tell-tale weakness of a machine in not being able to

> take into account the vast range of probability that characterizes the human situation.[1] (p.181)

But taking into account that range of probability is just where the new AI excels. The only chink in the armor of AI is that word "vast"; human possibilities, thanks to language and the culture that it spawns, are truly Vast.[3] No matter how many patterns we may find with AI in the flood of data that has so far found its way onto the Internet, there are Vastly more possibilities that have never been recorded there. Only a fraction (but not a Vanishing fraction) of the world's accumulated wisdom and design and repartee and silliness has made it onto the Internet, but probably a better tactic for the judge to adopt when confronting a candidate in the Turing Test is not to *search* for such items but to *create* them anew. AI in its current manifestations is

---

[2] The *surely alarm* (the habit of having a bell ring in your head whenever you see the word in an argument) is described and defended by me in *Intuition Pumps and Other Tools for Thinking* (2013).

[3] In *Darwin's Dangerous Idea,* 1995, p. 109, I coined the capitalized version, *Vast,* meaning *Very much more than ASTronomical,* and its complement, *Vanishing,* to replace the usual exaggerations *infinite* and *infinitesimal* for discussions of those possibilities that are not officially infinite but nevertheless infinite for all practical purposes.

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.[4] 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.

One of the flaws in Weizenbaum's book *Computer Power and Human Reason,* something I tried in vain to convince him of in many hours of discussion, is that he could never decide which of two theses he wanted to defend: *AI is impossible!* or *AI is possible but evil!* He wanted to argue, with John Searle and Roger Penrose, that "Strong AI" is impossible, but there are no good arguments for that conclusion. After all, everything we now know suggests that, as I have put it, we are robots made of robots made of robots. . . down to the motor proteins and their ilk, with no magical ingredients thrown in along the way. Weizenbaum's more important and defensible message was that we should not strive to create Strong AI and should be extremely cautious about the AI systems that we can create and have already created. As one might expect, the defensible thesis is a hybrid: *AI* (Strong AI) *is possible in principle but not desirable. The AI that's practically possible is not necessarily evil—unless it is mistaken for Strong AI!*

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 as George Church notes in these pages, it uses thousands of times more energy than a human brain (a technological limitation that, as he also notes, may be temporary). 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 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. Weizenbaum's ELIZA 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.

---

[4] Aylin Caliskan-Islam, Joanna J. Bryson & Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science,* 14 April 2017, 356: 6334, pp. 183-6. DOI: 10.1126/science.aal4230.

He was right to be worried. If there is one thing we have learned from the restricted Turing Test competitions for the 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. 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 operators, just as we license pharmacists (and crane operators!) and other specialists whose errors and misjudgments can have dire consequences, can, 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.

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 adopt the intentional stance, 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 novel habits of thought, conversational gambits and ruses, traps and bluffs to arise in this novel setting for human action. 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 can*not* be responsibly answered by particular systems, with heavy penalties for those 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. Some of the deepest pockets are voluntarily out in front of these obligations to serve society first and make money secondarily, but we shouldn't rely on good will alone.

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.[5] 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 I co-taught with Matthias Scheutz 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 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 entities, 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.

---

[5] Joanna J. Bryson, "Robots Should Be Slaves," in *Close Engagement with Artificial Companions*, Yorick Wilks, ed., (Amsterdam, The Netherlands: John Benjamins, 2010), pp. 63-74; http://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Slaves-Book09.html.

_____, "Patiency Is Not a Virtue: AI and the Design of Ethical Systems," https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Patiency-AAAISS16.pdf.