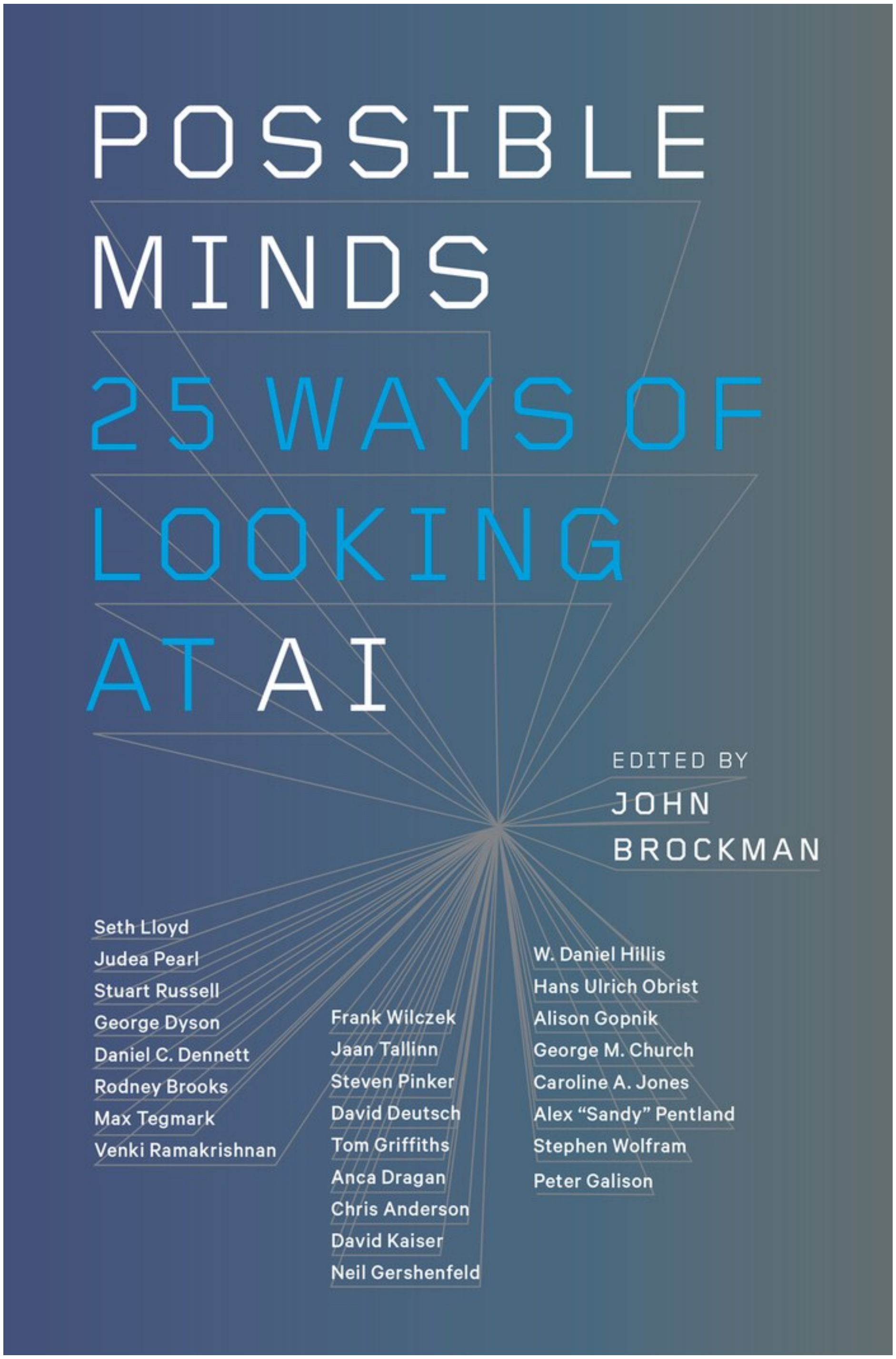

POSSIBLE MINDS

25 WAYS OF LOOKING AT AI

EDITED BY
JOHN
BROCKMAN

Seth Lloyd
Judea Pearl
Stuart Russell
George Dyson
Daniel C. Dennett
Rodney Brooks
Max Tegmark
Venki Ramakrishnan

Frank Wilczek
Jaan Tallinn
Steven Pinker
David Deutsch
Tom Griffiths
Anca Dragan
Chris Anderson
David Kaiser
Neil Gershenfeld

W. Daniel Hillis
Hans Ulrich Obrist
Alison Gopnik
George M. Church
Caroline A. Jones
Alex "Sandy" Pentland
Stephen Wolfram
Peter Galison

Throughout his career, whether studying language, advocating a realistic biology of mind, or examining the human condition through the lens of humanistic Enlightenment ideas, psychologist **Steven Pinker** *has embraced and championed a naturalistic understanding of the universe and the computational theory of mind. He is perhaps the first internationally recognized public intellectual whose recognition is based on the advocacy of empirically based thinking about language, mind, and human nature.*

"Just as Darwin made it possible for a thoughtful observer of the natural world to do without creationism," he says, "Turing and others made it possible for a thoughtful observer of the cognitive world to do without spiritualism."

In the debate about AI risk, he argues against prophecies of doom and gloom, noting that they spring from the worst of our psychological biases—exemplified particularly by media reports: "Disaster scenarios are cheap to play out in the probability-free zone of our imaginations, and they can always find a worried, technophobic, or morbidly fascinated audience." Hence, over the centuries: Pandora, Faust, the Sorcerer's Apprentice, Frankenstein, the population bomb, resource depletion, HAL, suitcase nukes, the Y2K bug, and engulfment by nanotechnological grey goo. "A characteristic of AI dystopias," he points out, "is that they project a parochial alpha-male psychology onto the concept of intelligence. . . . History does turn up the occasional megalomaniacal despot or psychopathic serial killer, but these are products of a history of natural selection shaping testosterone-sensitive circuits in a certain species of primate, not an inevitable feature of intelligent systems."

In the present essay, he applauds Wiener's belief in the strength of ideas vis-à-vis the encroachment of technology. As Wiener so aptly put it, "The machine's danger to society is not from the machine itself but from what man makes of it."

TECH PROPHECY AND THE UNDERAPPRECIATED CAUSAL POWER OF IDEAS
Steven Pinker

Steven Pinker, *a Johnstone Family Professor in the Department of Psychology at Harvard University, is an experimental psychologist who conducts research in visual cognition, psycholinguistics, and social relations. He is the author of eleven books, including* The Blank Slate, The Better Angels of Our Nature, *and, most recently,* Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.

Artificial intelligence is an existence proof of one of the great ideas in human history: that the abstract realm of knowledge, reason, and purpose does not consist of an élan vital or immaterial soul or miraculous powers of neural tissue. Rather, it can be linked to the physical realm of animals and machines via the concepts of information, computation, and control. Knowledge can be explained as patterns in matter or energy that stand in systematic relations with states of the world, with mathematical and logical truths, and with one another. Reasoning can be explained as transformations of that knowledge by physical operations that are designed to preserve those relations. Purpose can be explained as the control of operations to effect changes in the world, guided by discrepancies between its current state and a goal state. Naturally evolved brains are just the most familiar systems that achieve intelligence through information, computation, and control. Humanly designed systems that achieve intelligence vindicate the notion that information processing is sufficient to explain it—the notion that the late Jerry Fodor dubbed the computational theory of mind.

The touchstone for this volume, Norbert Wiener's *The Human Use of Human Beings,* celebrated this intellectual accomplishment, of which Wiener himself was a foundational contributor. A potted history of the mid-20th-century revolution that gave the world the computational theory of mind might credit Claude Shannon and Warren Weaver for explaining knowledge and communication in terms of information. It might credit Alan Turing and John von Neumann for explaining intelligence and reasoning in terms of computation. And it ought to give Wiener credit for explaining the hitherto mysterious world of purposes, goals, and teleology in terms of the technical concepts of feedback, control, and cybernetics (in its original sense of "governing" the operation of a goal-directed system). "It is my thesis," he announced, "that the physical functioning of the living individual and the operation of some of the newer communication machines are precisely parallel in their analogous attempts to control entropy through feedback"—the staving off of life-sapping entropy being the ultimate goal of human beings.

Wiener applied the ideas of cybernetics to a third system: society. The laws, norms, customs, media, forums, and institutions of a complex community could be considered channels of information propagation and feedback that allow a society to ward off disorder and pursue certain goals. This is a thread that runs through the book and which Wiener himself may have seen as its principal contribution. In his explanation of feedback, he wrote, "This complex of behavior is ignored by the average man, and in particular does not play the role that it should in our habitual analysis of society; for just as individual physical responses may be seen from this point of view, so may the organic responses of society itself."

Indeed, Wiener gave scientific teeth to the idea that in the workings of history, politics, and society, *ideas matter*. Beliefs, ideologies, norms, laws, and customs, by regulating the behavior of the humans who share them, can shape a society and power the course of historical events as surely as the phenomena of physics affect the structure and evolution of the solar system. To say that ideas—and not just weather, resources, geography, or weaponry—can shape history is not woolly mysticism. It is a statement of the causal powers of information instantiated in human brains and exchanged in networks of communication and feedback. Deterministic theories of history, whether they identify the causal engine as technological, climatological, or geographic, are belied by the causal power of ideas. The effects of these ideas can include unpredictable lurches and oscillations that arise from positive feedback or from miscalibrated negative feedback.

An analysis of society in terms of its propagation of ideas also gave Wiener a guideline for social criticism. A healthy society—one that gives its members the means to pursue life in defiance of entropy—allows information sensed and contributed by its members to feed back and affect how the society is governed. A dysfunctional society invokes dogma and authority to impose control from the top down. Wiener thus described himself as "a participant in a liberal outlook," and devoted most of the moral and rhetorical energy in the book (both the 1950 and 1954 editions) to denouncing communism, fascism, McCarthyism, militarism, and authoritarian religion (particularly Catholicism and Islam) and to warning that political and scientific institutions were becoming too hierarchical and insular.

Wiener's book is also, here and there, an early exemplar of an increasingly popular genre, tech prophecy. Prophecy not in the sense of mere prognostications but in the Old Testament sense of dark warnings of catastrophic payback for the decadence of one's contemporaries. Wiener warned against the accelerating nuclear arms race, against technological change that was imposed without regard to human welfare ("[W]e must know as scientists what man's nature is and what his built-in purposes are"), and against what today is called the value-alignment problem: that "the machine like the djinnee, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us." In the darker, 1950 edition, he warned of a "threatening new Fascism dependent on the *machine à gouverner*."

Wiener's tech prophecy harks back to the Romantic movement's rebellion against the "dark Satanic mills" of the Industrial Revolution, and perhaps even earlier, to the archetypes of Prometheus, Pandora, and Faust. And today it has gone into high gear. Jeremiahs, many of them (like Wiener) from the worlds of science and technology, have sounded alarms about nanotechnology, genetic engineering, Big Data, and particularly artificial intelligence. Several contributors to this volume characterize Wiener's book as a prescient example of tech prophecy and amplify his dire worries.

Yet the two moral themes of *The Human Use of Human Beings*—the liberal defense of an open society and the dystopian dread of runaway technology—are in tension. A society with channels of feedback that maximize human flourishing will have mechanisms in place, and can adapt them to changing circumstances, in a way that can domesticate technology to human purposes. There's nothing idealistic or mystical about this; as Wiener emphasized, ideas, norms, and institutions are themselves a form of technology, consisting of patterns of information distributed across brains. The possibility that machines threaten a new fascism must be weighed against the vigor of the liberal ideas, institutions, and norms that Wiener championed throughout the book. The flaw in today's dystopian prophecies is that they disregard the existence of these norms and institutions, or drastically underestimate their causal potency. The result is a technological determinism whose dark predictions are repeatedly refuted by the course of events. The numbers "1984" and "2001" are good reminders.

I will consider two examples. Tech prophets often warn of a "surveillance state" in which a government empowered by technology will monitor and interpret all private communications, allowing it to detect dissent and subversion as it arises and make resistance to state power futile. Orwell's telescreens are the prototype, and in 1976 Joseph Weizenbaum, one of the gloomiest tech prophets of all time, warned my class of graduate students not to pursue automatic speech recognition because government surveillance was its only conceivable application.

Though I am on record as an outspoken civil libertarian, deeply concerned with contemporary threats to free speech, I lose no sleep over technological advances in the Internet, video, or artificial intelligence. The reason is that almost all the variation across time and space in freedom of thought is driven by differences in norms and institutions and almost none of it by differences in technology. Though one can imagine hypothetical combinations of the most malevolent totalitarians with the most advanced technology, in the real world it's the norms and laws we should be vigilant about, not the tech.

Consider variation across time. If, as Orwell hinted, advancing technology was a prime enabler of political repression, then Western societies should have gotten more and more restrictive of speech over the centuries, with a dramatic worsening in the second half of the 20th century continuing into the 21st. That's not how history unfolded. It was the centuries when communication was implemented by quills and inkwells that had autos-da-fé and the jailing or guillotining of Enlightenment thinkers. During World War I, when the state of the art was the wireless, Bertrand Russell was jailed for his pacifist opinions. In the 1950s, when computers were room-size accounting machines, hundreds of liberal writers and scholars were professionally punished. Yet in the technologically accelerating, hyperconnected 21st century, 18 percent of social science professors are Marxists[1]; the President of the United States is nightly ridiculed by television comedians as a racist, pervert, and moron; and technology's biggest threat to political discourse comes from amplifying too many dubious voices rather than suppressing enlightened ones.

Now consider variations across place. Western countries at the technological frontier consistently get the highest scores in indexes of democracy and human rights, while many backward strongman states are at the bottom, routinely jailing or killing government critics. The lack of a correlation between technology and repression is unsurprising when you analyze the channels of information flow in any human society. For dissidents to be influential, they have to get their message out to a wide network via whatever channels of communication are available—pamphleteering, soap-box oration, subversive soirées in cafés and pubs, word of mouth. These channels enmesh influential dissidents in a broad social network which makes them easy to identify and track down. All the more so when dictators rediscover the time-honored technique of weaponizing the people against each other by punishing those who don't denounce or punish others.

In contrast, technologically advanced societies have long had the means to install Internet-connected, government-monitored surveillance cameras in every bar and bedroom. Yet that has not happened, because democratic governments (even the current American administration, with its flagrantly antidemocratic impulses) lack the will and the means to

[1] Neil Gross & Solon Simmons, "The Social and Political Views of American College and University Professors," in N. Gross & S. Simmons, eds., *Professors and Their Politics* (Baltimore: Johns Hopkins University Press, 2014).

enforce such surveillance on an obstreperous people accustomed to saying what they want. Occasionally, warnings of nuclear, biological, or cyberterrorism goad government security agencies into measures such as hoovering up mobile phone metadata, but these ineffectual measures, more theater than oppression, have had no significant effect on either security or freedom. Ironically, tech prophecy plays a role in encouraging these measures. By sowing panic about supposed existential threats such as suitcase nuclear bombs and bioweapons assembled in teenagers' bedrooms, they put pressure on governments to prove they're doing something, anything, to protect the American people.

It's not that political freedom takes care of itself. It's that the biggest threats lie in the networks of ideas, norms, and institutions that allow information to feed back (or not) on collective decisions and understanding. As opposed to the chimerical technological threats, one real threat today is oppressive political correctness, which has choked the range of publicly expressible hypotheses, terrified many intelligent people against entering the intellectual arena, and triggered a reactionary backlash. Another real threat is the combination of prosecutorial discretion with an expansive lawbook filled with vague statutes. The result is that every American unwittingly commits "three felonies a day" (as the title of a book by civil libertarian Harvey Silverglate puts it) and is in jeopardy of imprisonment whenever it suits the government's needs. It's this prosecutorial weaponry that makes Big Brother all-powerful, not telescreens. The activism and polemicizing directed against government surveillance programs would be better directed at its overweening legal powers.

The other focus of much tech prophecy today is artificial intelligence, whether in the original sci-fi dystopia of computers running amok and enslaving us in an unstoppable quest for domination, or the newer version in which they subjugate us by accident, single-mindedly seeking some goal we give them regardless of its side effects on human welfare (the value-alignment problem adumbrated by Wiener). Here again both threats strike me as chimerical, growing from a narrow technological determinism that neglects the networks of information and control in an intelligent system like a computer or brain and in a society as a whole.

The subjugation fear is based on a muzzy conception of intelligence that owes more to the Great Chain of Being and a Nietzschean will to power than to a Wienerian analysis of intelligence and purpose in terms of information, computation, and control. In these horror scenarios, intelligence is portrayed as an all-powerful, wish-granting potion that agents possess in different amounts. Humans have more of it than animals, and an artificially intelligent computer or robot will have more of it than humans. Since we humans have used our moderate endowment to domesticate or exterminate less well-endowed animals (and since technologically advanced societies have enslaved or annihilated technologically primitive ones), it follows that a supersmart AI would do the same to us. Since an AI will think millions of times faster than we do, and use its superintelligence to recursively improve its superintelligence, from the instant it is turned on we will be powerless to stop it.

But these scenarios are based on a confusion of intelligence with motivation—of beliefs with desires, inferences with goals, the computation elucidated by Turing and the control elucidated by Wiener. Even if we did invent superhumanly intelligent robots, why would they *want* to enslave their masters or take over the world? Intelligence is the ability to deploy novel means to attain a goal. But the goals are extraneous to the intelligence: Being smart is not the same as wanting something. It just so happens that the intelligence in *Homo sapiens* is a product of Darwinian natural selection, an inherently competitive process. In the brains of that species, reasoning comes bundled with goals such as dominating rivals and amassing resources.

But it's a mistake to confuse a circuit in the limbic brain of a certain species of primate with the very nature of intelligence. There is no law of complex systems that says that intelligent agents must turn into ruthless megalomaniacs.

A second misconception is to think of intelligence as a boundless continuum of potency, a miraculous elixir with the power to solve any problem, attain any goal. The fallacy leads to nonsensical questions like when an AI will "exceed human-level intelligence," and to the image of an "artificial general intelligence" (AGI) with God-like omniscience and omnipotence. Intelligence is a contraption of gadgets: software modules that acquire, or are programmed with, knowledge of how to pursue various goals in various domains. People are equipped to find food, win friends and influence people, charm prospective mates, bring up children, move around in the world, and pursue other human obsessions and pastimes. Computers may be programmed to take on some of these problems (like recognizing faces), not to bother with others (like charming mates), and to take on still other problems that humans can't solve (like simulating the climate or sorting millions of accounting records). The problems are different, and the kinds of knowledge needed to solve them are different.

But instead of acknowledging the centrality of knowledge to intelligence, the dystopian scenarios confuse an artificial general intelligence of the future with Laplace's demon, the mythical being that knows the location and momentum of every particle in the universe and feeds them into equations for physical laws to calculate the state of everything at any time in the future. For many reasons, Laplace's demon will never be implemented in silicon. A real-life intelligent system has to acquire information about the messy world of objects and people by engaging with it one domain at a time, the cycle being governed by the pace at which events unfold in the physical world. That's one of the reasons that understanding does not obey Moore's Law: Knowledge is acquired by formulating explanations and testing them against reality, not by running an algorithm faster and faster. Devouring the information on the Internet will not confer omniscience either: Big Data is still finite data, and the universe of knowledge is infinite.

A third reason to be skeptical of a sudden AI takeover is that it takes too seriously the inflationary phase in the AI hype cycle in which we are living today. Despite the progress in machine learning, particularly multilayered artificial neural networks, current AI systems are nowhere near achieving general intelligence (if that concept is even coherent). Instead, they are restricted to problems that consist of mapping well-defined inputs to well-defined outputs in domains where gargantuan training sets are available, in which the metric for success is immediate and precise, in which the environment doesn't change, and in which no stepwise, hierarchical, or abstract reasoning is necessary. Many of the successes come not from a better understanding of the workings of intelligence but from the brute-force power of faster chips and Bigger Data, which allow the programs to be trained on millions of examples and generalize to similar new ones. Each system is an idiot savant, with little ability to leap to problems it was not set up to solve, and a brittle mastery of those it was. And to state the obvious, none of these programs has made a move toward taking over the lab or enslaving its programmers.

Even if an artificial intelligence system tried to exercise a will to power, without the cooperation of humans it would remain an impotent brain in a vat. A superintelligent system, in its drive for self-improvement, would somehow have to build the faster processors that it would run on, the infrastructure that feeds it, and the robotic effectors that connect it to the world—all impossible unless its human victims worked to give it control of vast portions of the engineered world. Of course, one can always imagine a Doomsday Computer that is malevolent, universally empowered, always on, and tamperproof. The way to deal with this threat is straightforward: Don't build one.

What about the newer AI threat, the value-alignment problem, foreshadowed in Wiener's allusions to stories of the Monkey's Paw, the genie, and King Midas, in which a wisher rues the unforeseen side effects of his wish? The fear is that we might give an AI system a goal and then helplessly stand by as it relentlessly and literal-mindedly implemented its interpretation of that goal, the rest of our interests be damned. If we gave an AI the goal of maintaining the water level behind a dam, it might flood a town, not caring about the people who drowned. If we gave it the goal of making paper clips, it might turn all the matter in the reachable universe into paper clips, including our possessions and bodies. If we asked it to maximize human happiness, it might implant us all with intravenous dopamine drips, or rewire our brains so we were happiest sitting in jars, or, if it had been trained on the concept of happiness with pictures of smiling faces, tile the galaxy with trillions of nanoscopic pictures of smiley-faces.

Fortunately, these scenarios are self-refuting. They depend on the premises that (1) humans are so gifted that they can design an omniscient and omnipotent AI, yet so idiotic that they would give it control of the universe without testing how it works; and (2) the AI would be so brilliant that it could figure out how to transmute elements and rewire brains, yet so imbecilic that it would wreak havoc based on elementary blunders of misunderstanding. The ability to choose an action that best satisfies conflicting goals is not an add-on to intelligence that engineers might forget to install and test; it *is* intelligence. So is the ability to interpret the intentions of a language user in context.

When we put aside fantasies like digital megalomania, instant omniscience, and perfect knowledge and control of every particle in the universe, artificial intelligence is like any other technology. It is developed incrementally, designed to satisfy multiple conditions, tested before it is implemented, and constantly tweaked for efficacy and safety.

The last criterion is particularly significant. The culture of safety in advanced societies is an example of the humanizing norms and feedback channels that Wiener invoked as a potent causal force and advocated as a bulwark against the authoritarian or exploitative implementation of technology. Whereas at the turn of the 20th century Western societies tolerated shocking rates of mutilation and death in industrial, domestic, and transportation accidents, over the course of the century the value of human life increased. As a result, governments and engineers used feedback from accident statistics to implement countless regulations, devices, and design changes that made technology progressively safer. The fact that some regulations (such as using a cell phone near a gas pump) are ludicrously risk-averse underscores the point that we have become a society obsessed with safety, with fantastic benefits as a result: Rates of industrial, domestic, and transportation fatalities have fallen by more than 95 (and often 99) percent since their highs in the first half of the 20th century.[2] Yet tech prophets of malevolent or oblivious artificial intelligence write as if this momentous transformation never happened and one morning engineers will hand total control of the physical world to untested machines, heedless of the human consequences.

Norbert Wiener explained ideas, norms, and institutions in terms of computational and cybernetic processes that were scientifically intelligible and causally potent. He explained human beauty and value as "a local and temporary fight against the Niagara of increasing entropy" and expressed the hope that an open society, guided by feedback on human well-being,

[2] Steven Pinker, "Safety," *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (New York: Penguin, 2018).

would enhance that value. Fortunately his belief in the causal power of ideas counteracted his worries about the looming threat of technology. As he put it, "the machine's danger to society is not from the machine itself but from what man makes of it." It is only by remembering the causal power of ideas that we can accurately assess the threats and opportunities presented by artificial intelligence today.