Summer 2017

Artificial Intelligence and Public Trust

Shannon Vallor
Santa Clara University, svallor@scu.edu

Artificial Intelligence and Public Trust

A future with artificial intelligence is no longer a sci-fi fantasy. But how do we ensure that it is shaped with moral intelligence?

WORDS BY SHANNON VALLOR
ILLUSTRATIONS BY JOSH COCHRAN

The future is here. With the exploding commercial market for high-powered, cloud-computing AI services provided by the likes of Amazon, Microsoft, and Google, the reach of artificial intelligence technologies is virtually unlimited. What does this mean for humans? How will we adapt to a world in which we increasingly find ourselves in economic, creative, and cognitive competition with machines? Will we embrace these new technologies with the same fervor as we embraced televisions and smartphones? Will we trust them? Should we trust them?

Popular essays and news articles about an AI-driven future often highlight grim warnings of science and technology luminaries like Elon Musk and Stephen Hawking, who raise the specter of the emergence of "superintelligent" machines that could threaten human survival or assume control of our future. Yet most AI researchers regard this prospect as highly unlikely, for it presupposes the emergence of artificial general intelligence (AGI)—the kind of flexible, self-aware, and fairly comprehensive understanding of the world that humans enjoy. The AI that we have today (and will be seeing a lot more of) is of an entirely different kind, one that fundamentally lacks the capacities needed for AGI. For the foreseeable future, humans will navigate a world populated by artificial agents that possess no general understanding of the world—or of us, or of themselves, or much of anything at all, really. What they will have is exceptional skill and speed at performing specific, well-defined tasks that used to require human intelligence. This kind of AI, powered by large datasets combined with advances in machine learning techniques, doesn't recreate or even imitate our kind of smarts at all. It bypasses it—and does smart things without it. Although this kind of AI may seem far less scary than a self-aware Skynet that decides to wipe out human pests, the risks of this more mundane species of AI are nearly as profound.

One obvious risk: a new wave of AI-driven technological unemployment. Although economists' predictions vary, an oft-cited 2013 study from the Oxford Martin School estimates that as many as 47 percent of American jobs could be at risk from AI-driven automation within a few decades. Even if artificial agents cannot wholly replace most human workers in the short term, the emergence of task-specific artificial intelligence across a broad range of new industries and social contexts is already rapidly transforming every domain of human activity, from commerce and transportation to education and medicine. Every system that makes, sells, or distributes goods and services to human beings has the opportunity to benefit—and to be radically destabilized by—the new wave of machine automation and decision support that task-specific AI makes possible.

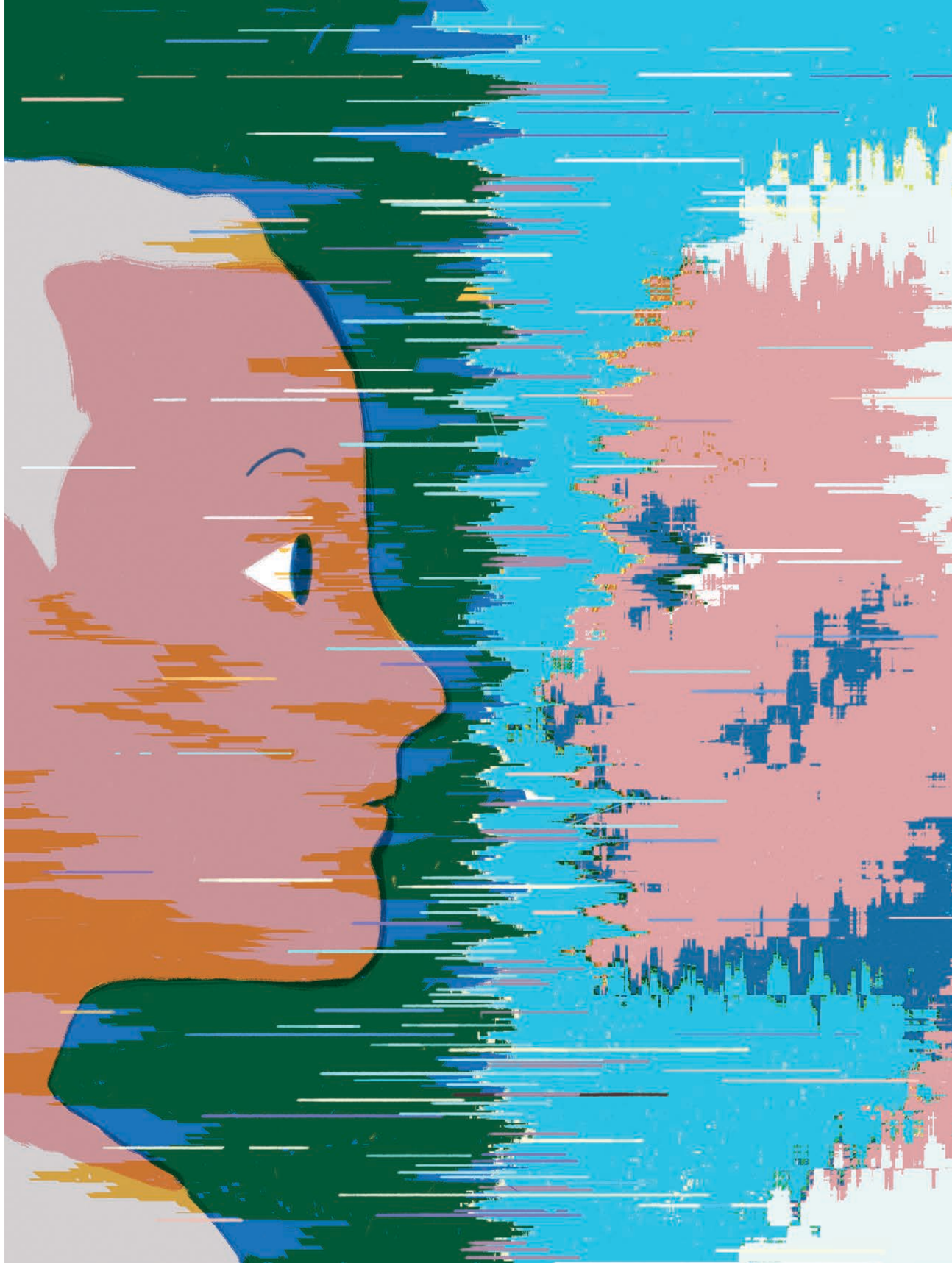
Lack of transparency in some "deep learning" algorithms means that even system programmers lack a clear view of how or why the system reached its conclusion.

CAN WE TRUST AI?

Today, AI-powered software is used to identify terrorist threats and targets in voice, image, email, social media, and SMS data; to assign criminal defendants risk scores for judges to use in making bail, sentencing, and parole decisions; to tell your local law enforcement where they are most likely to encounter certain crimes; and to diagnose cancers and recommend personalized treatment plans. Task-specific AI algorithms are calculating how likely you are to "fit" into the corporate culture or remain with the company to which you have applied, how close a "match" a stranger is to your romantic preferences, how likely you are to repay the loan you applied for, or the chances that your kid will thrive at the selective private school you want her to attend. These decisions govern how well or how poorly our lives go: whether we live or die, whether we work or are unemployed, whether we are free or unfree. What would it take for you to trust a machine to make such life-changing decisions for you—or for your employer, loan officer, doctor, insurance company, or your child's college admissions committee? In many cases, it's already happening.

There is a common saying that commands prudence in matters of social reliance: "trust, but verify." Consider this: In virtually none of these artificial decision support systems can you, as an ordinary person affected by the outcome, know how the algorithmic decision process is carried out, or what salient factors drove the algorithm's result in your particular case. In many cases—due to the lack of transparency in "deep learning" algorithms that work without showing their internal logic—even the system's programmers and administrators lack a clear view of how or why the system reached its conclusion. So who, what, and how do we verify? And if we cannot verify, can we still trust?

One might think that careful regimes of inspection can easily ensure that artificial agents are operating properly, and that what's "under the hood" is not broken or poorly designed. Yet what's under the hood in many such systems is not a set of clear, stable rules and inferences that we can examine and test for their validity, but rather a tangled mess of artificial neural networks arranged in complex layers with nodes and weightings that constantly rearrange



themselves based on changing inputs and outputs. Verification of such a system's accuracy and reliability, or reconstruction of a machine's pattern of reasoning, is often impossible in individual cases. At best we can say that as a statistical matter, over a large number of trials, the system produces acceptable results at least as often as a human would. In fact, the impressive power of many machine-learning techniques results from designs that simultaneously make it impossible to guarantee an accurate result in any particular case. In such systems, it is inevitable that they will sometimes, however rarely, produce "inappropriate" solutions—even wildly inappropriate, just because AI agents "reason" so differently from human intellects.

Ironically, at other times algorithmic systems will produce harmful and unfair outcomes for the opposite reason—that is, because their decisions will not be different *enough* from ours, if they are trained on human-generated data that infects them with our own harmful biases and falsehoods. Examples include racial bias found in criminal risk-score algorithms widely relied upon by U.S. judges, algorithms which produce the illusion of "neutral," "objective" analysis but in fact reproduce unjust human prejudices by mislabeling black defendants as high-risk reoffenders at far higher rates than similar white defendants are mislabeled. A less grave but still ugly example was Microsoft's notorious "Tay" teen chatbot that in 2016 began "learning" to adopt white supremacist slurs and conspiracy theories within hours of its release on Twitter.

WHO'S RESPONSIBLE?

One might be tempted at this point to say, "Well then, so much the worse for AI—let's just get rid of it and go back to relying on our own mental horsepower!" But this kind of neo-Luddite response to AI would be throwing the baby out with the bathwater. Due to the immense speed, adaptability, and computational power of these new software tools, they hold the promise of helping us solve countless urgent problems that human minds are just too slow, too distractible, or too constrained by evolutionary pressures to solve alone. Would you be willing to forgo—or forgo for your children and grandchildren—a cure for Alzheimer's, or cleaner and vastly more efficient power systems, or reliable weather and global climate forecasts, or better responses to drought and famine? Then we cannot afford to reject artificial intelligence out of hand.

This creates an unprecedented ethical imperative for AI researchers, designers, users, and companies and institutions that employ them. Artificial intelligence is immensely powerful, but it is not *magic*. It does not run without human intelligence—including, even chiefly, our moral intelligence. The future of an AI-driven world depends less upon new breakthroughs in machine learning algorithms and big data than it does upon the choices that humans make in how AI gets integrated into our daily lives and institutions and how its risks and effects are managed.

This imperative falls within the realm of ethics because core human goods and values are at stake. An artificial agent that ruins the rest of your life by falsely labeling you a high-risk defendant, or that denies you a home or a job because of a random algorithmic quirk that no one can see, is implicated in an injustice, especially when it is relied upon by other humans in ways that deny you due process or meaningful remedies. We cannot sit by and allow compassion, justice, liberty, and respect for human dignity to be sacrificed at the altar of algorithmic efficiency. Every AI-enabled decision process is still a human responsibility, all the way down to its deepest, darkest, most inscrutable layers.

We cannot sit by and allow compassion, justice, liberty, and respect for human dignity to be sacrificed at the altar of algorithmic efficiency.

Things can be done to foster and earn the public's trust in artificial intelligence. First, companies that develop and market AI-driven technologies need to cultivate a sincere public conscience and internal corporate culture, supported by incentive structures, that reflect awareness of the unprecedented social power of these tools. Respect for human life and dignity is not incompatible with healthy commerce and reliance on markets. It's essential to it. If we don't tolerate profit-driven recklessness and contempt for public health and safety from companies that build and operate nuclear reactors or airliners, we cannot tolerate it from companies that build and operate AI, especially when they impact critical human systems and institutions.

Second, the public needs to adopt a more critical, questioning relationship with technology and its social effects. We each need to become better educated about the promise and the limits of artificial intelligence, and to actively demand and participate in AI governance and oversight, in both formal regulatory structures and informal citizen-driven structures. From the person who is asked by their doctor or employer to surrender genetic data to an AI-driven cloud platform, to the HR manager who downloads an AI hiring assistant to sort résumés or evaluate interview responses, to the juror or judge presented with an AI-generated risk score, we all need to ask reasonable questions and demand reasonable answers about AI-driven systems, such as: "What are appropriate uses of this tool? What are common inappropriate uses/misuses of this tool?" "What human biases could have skewed the data this system was trained on, and what measures were taken to identify or mitigate biased results?" "What kind of errors will this system most likely make, when it makes them?" "What auditing processes are in place to identify individual errors or harmful/unjust patterns in the results?" "What steps can I or my organization take to ensure that independent human checks and other due-process measures are available when an algorithmic decision is contested by an affected party?"

Third, institutions that rely heavily upon AI-driven solutions, especially those institutions that protect fundamental human goods such as education and health, need to develop institutional structures and incentives that ensure that fundamental human values central to the mission of the institutions are not lost or sacrificed to the rule of algorithmic "efficiency" and its opaque authority. Human judgment must remain in the loop in such a way that the vigor of human intellect, the virtues of moral wisdom, and an ethos of personal responsibility are preserved and given ample opportunities to be practiced and honed. Artificial intelligence can even be enlisted in this effort as artificial helpers and tutors that encourage and support the ongoing cultivation and refinement of human intelligence, rather than demoting or degrading it to a lesser status.

Artificial intelligence is already one of humanity's sharpest tools. But like any very sharp tool we have crafted for ourselves, it must be treated with care and discernment. We must know where and when it is safe to use, and where and when it is not. We must know with whom to entrust its use, and with whom to not. We must know how to keep its power from injuring or enfeebling ourselves, or those we love. And we must know that the tool and its power is always the responsibility of the one who trusts it.

SHANNON VALLOR is the William J. Rewak, S.J., Professor of Philosophy and the author of *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press).