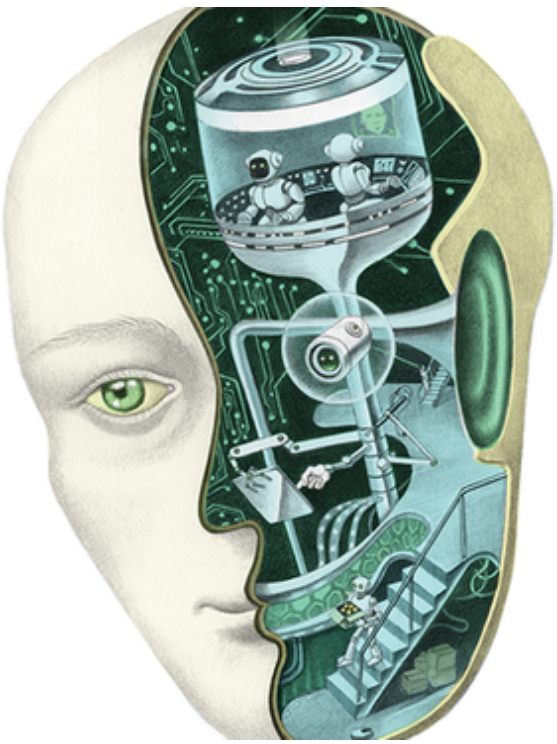
ENGINEERING

Artificial Intelligence Will Serve Humans, Not Enslave Them

AI will serve our species, not control it

By Pedro Domingos | Scientific American September 2018 Issue



Credit: Armando Veve

IN BRIEF

The pursuit of artificial intelligence can be seen as part of human evolution. The next stage of automation will require the creation of a so-called master algorithm. It would integrate the five main ways that machines currently learn into a single, unified paradigm.

Technology is simply an extension of human capabilities. Machines do not have free will, only goals that we give to them. It is the misuse of the technology by people that we should be worried about, not a robot takeover.

A more plausible near-term scenario for AI is the proliferation of "digital doubles"—virtual models of ourselves that will interact with each other in countless simulations to help us make faster, more informed choices in our daily lives.

Humans are the only animals that build machines. By doing so, we expand our capabilities beyond our biological limits. Tools turn our hands into more versatile appendages. Cars let us travel faster, and airplanes give us wings. Computers endow us with bigger brains and memory capacity, and smartphones orchestrate daily life. Now we are creating technology that can evolve on its own by encoding into it an ability to learn through data and effort. Will it ultimately supplant us? Or will it augment our abilities, enhancing our humanness in unprecedented ways?

Machine learning started in the 1950s with the work of pioneering scientists such as Frank Rosenblatt, who built an electronic neuron that learned to recognize digits, and Arthur Samuel, whose checkers program learned by playing against itself until it could beat some humans. But it is only in the past decade that the field has truly taken off, giving us self-driving cars, virtual assistants that understand our commands (up to a point) and countless other applications.

Every year we invent thousands of new algorithms, which are sequences of instructions telling a computer what to do. The hallmark of learning machines, however, is that instead of programming them in detail, we give them general goals such as "learn to play checkers." Then, like humans, they improve with experience. These learning algorithms tend to fall into five main categories, each inspired by a different scientific field. Unsurprisingly, one way that machines learn is by mimicking natural selection, through evolutionary algorithms. In the Creative Machines Lab at Columbia University, primitive robots try to crawl or fly, and the specifications of those that perform best are periodically mixed and mutated to 3-D print the next generation. Starting with randomly assembled bots that can barely move, this process eventually produces creatures such as robot spiders and dragonflies after thousands or tens of thousands of generations.

But evolution is slow. Deep learning, currently the most popular machine-learning paradigm, takes inspiration from the brain. We start with a highly simplified mathematical model of how a neuron works and then build a network from thousands or millions of these units and let it learn by gradually strengthening the connections between neurons that fire together when looking at data. These neural networks can recognize faces, understand speech and translate languages with uncanny accuracy. Machine learning also draws on psychology. Like humans, these analogy-based algorithms solve new problems by finding similar ones in memory. This ability allows for the automation of customer support, as well as e-commerce sites that recommend products based on your tastes.

Machines may also learn by automating the scientific method. To induce a new hypothesis, symbolic learners invert the process of deduction: If I know that Socrates is human, what else do I need to infer that he is mortal? Knowing that humans are mortal would suffice, and this hypothesis can then be tested by checking if other humans in the data are also mortal. Eve, a biologist robot at the University of Manchester in England, has used this approach to discover a potential new malaria drug. Starting with data about the disease and basic knowledge of molecular biology, Eve formulated hypotheses about what drug compounds might work, designed experiments to test them, carried out the experiments in a robotic lab, revised or discarded the hypotheses, and repeated until it was satisfied.

Finally, learning can rely purely on mathematical principles, the most important of which is Bayes's theorem. The theorem says that we should assign initial probabilities to hypotheses based on our knowledge, then let the hypotheses that are consistent with the data become more probable and those that are not become less so. It then makes predictions by letting all the hypotheses vote, with the more probable ones carrying more weight. Bayesian learning machines can do some medical diagnoses more accurately than human doctors. They are also at the heart of many spam filters and of the system that Google uses to choose which ads to show you.

Each of these five kinds of machine learning has its strengths and weaknesses. Deep learning, for example, is good for perceptual problems such as vision and speech recognition but not for cognitive ones such as acquiring commonsense knowledge and reasoning. With symbolic learning, the reverse is true. Evolutionary algorithms are capable of solving harder problems than neural

networks, but it can take a very long time to solve them. Analogical methods can learn from just a small number of instances but are liable to get confused when given too much information about each. Bayesian learning is most useful for dealing with small amounts of data but can be prohibitively expensive with big data.

These vexing trade-offs are why machine-learning researchers are working toward combining the best elements of all the paradigms. In the same way that a master key opens all locks, our goal is to create a so-called master algorithm—one that can learn everything that can be extracted from data, deriving all possible knowledge from it.

The challenge on us now is similar to the one faced by physicists: quantum mechanics is effective at describing the universe at the smallest scales and general relativity at the largest scales, but the two are incompatible and need to be reconciled. And in the same way that James Clerk Maxwell first unified light, electricity and magnetism before the Standard Model of particle physics could be developed, different research groups, including mine at the University of Washington, have proposed ways to unify two or more of the machine-learning paradigms. Because scientific progress is not linear and instead happens in fits and starts, it is difficult to predict when the full unification of the master algorithm might be complete. Regardless, achieving this goal will not usher in a new, dominant race of machines. Rather, it will accelerate human progress.

MACHINE TAKEOVER?

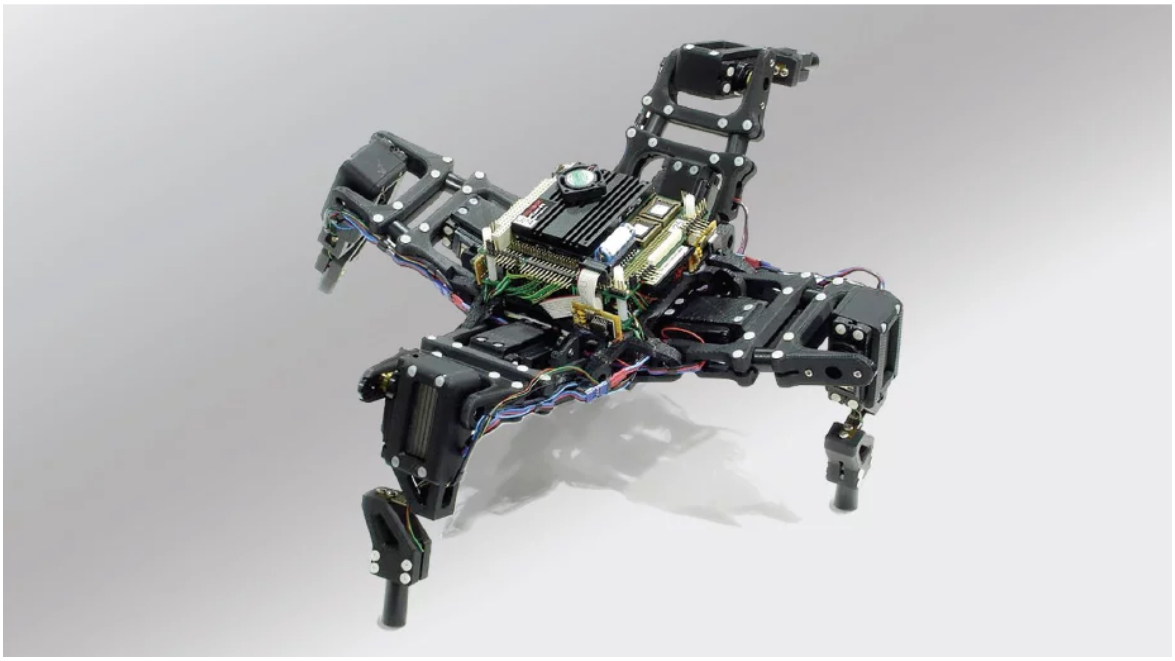
Once we attain the master algorithm and feed it the vast quantities of data each of us produce, artificial-intelligence systems will potentially be able to learn very accurate and detailed models of individual people: our tastes and habits, strengths and weaknesses, memories and aspirations, beliefs and personalities, the people and things we care about, and how we will respond in any given situation. That models of us could essentially predict the choices we will make is both exciting and disquieting.

Many worry that machines with these capabilities will use their newfound knowledge to take all our jobs, enslave us or even exterminate us. But that is unlikely to happen because they have no will of their own. Essentially all AI algorithms are driven by goals that we program, such as “find the shortest route

from the hotel to the airport." What distinguishes these algorithms from ordinary ones is that they have a lot of flexibility in figuring out how to reach the goals we set for them rather than needing to execute a predefined series of steps. Even as they get better at the task with experience, the goals remain unchanged. Solutions that do not make progress toward the goal are automatically discarded. Plus, humans get to check that what the machines produce does indeed satisfy our objectives. We are also able to verify that the machines do not violate any of the constraints we put on them, such as "obey the rules of the road."

When we envision an AI, though, we tend to project onto it human qualities such as volition and consciousness. Most of us are also more familiar with humanlike AIs, such as home robots, than with the myriad other types that do their work behind the scenes. Hollywood compounds this perception by depicting robots and AIs as humans in disguise—an understandable tactic that makes for a more compelling story. Artificial intelligence is just the ability to solve hard problems—a task that does not require free will. It is no more likely to turn against us than your hand is to slap you. Like any other technology, AIs will always be extensions of us. The more powerful we can make them, the better.

What, then, might our AI-enabled future look like? Intelligent machines will indeed supplant many jobs, but the effects on society will likely be similar to previous forms of automation. Two hundred years ago the majority of Americans were farmers. Yet today machines have replaced almost all of them without causing massive unemployment. Doomsayers argue that this time is different because machines are replacing our brains, not just our brawn, leaving nothing for humans to do. But the day that AIs can carry out all the tasks we can is still very distant, if it ever comes. For the foreseeable future, AIs and humans will be good at different things. Machine learning's primary effect will be to greatly lower the cost of intelligence. This democratization will increase the variety of economically feasible uses of that intelligence, generating new jobs and transforming old ones to accomplish more with the same amount of human labor.
Then there is the "singularity" scenario, popularized by futurist Ray Kurzweil. It is one of ever accelerating technological progress: machines learn to make better machines, which in turn make even better ones, and so on. But we know that this cannot continue forever because the laws of physics place strict limits on how powerful even a quantum computer can be, and in some aspects, we are not far from hitting them. The progress of AI, like the progress of everything else, will eventually plateau.



SMART BOT: This sea star uses evolutionary algorithms to learn how to simulate itself. These algorithms are one type of machine learning that could be unified with others into a “master algorithm,” a singularly powerful human tool. Credit: Victor Zykov and Josh Bongard

Another vision popular among futurists is that computer models of us will become so good that they will be practically indistinguishable from the real thing. In this scenario, we could upload ourselves to the cloud and live on forever as pieces of software, free of the pesky constraints of the physical world. One problem with this scenario is that it may not be biologically feasible. To upload yourself, you would presumably need an accurate model of each of your neurons, complete with the memories they store. It would have to be captured so reliably that the model's predictions would not rapidly diverge from the behavior of the real neurons—a tall order indeed. But even if this were a realistic option, would you really upload yourself if you had the chance? How could you know for sure that your model was not missing some essential part of you—or that it was conscious at all? What if a thief stole your identity in the most absolute and complete sense of the word? I believe that people will opt to hang on to their squishy, carbon-based selves—the “wetware,” as computer scientists jokingly call it—for as long as they can and then call it quits.

CHERCHEZ L'HUMAIN

AI—machine learning in particular—is really just the continuation of human evolution. In *The Extended Phenotype*, Richard Dawkins shows how common it is for animals' genes to control the environment beyond their bodies, from cuckoo eggs to beaver dams. (Dawkins serves on *Scientific American*'s board of advisers.) Technology is the extended phenotype of humans, and what we are building today is another layer of our technological exoskeleton. I think the most likely scenario for how humans will use AI is more fascinating than the usual speculations.

Within a decade each one of us will probably have a "digital double," an AI companion that will be even more indispensable than our smartphones are today. Your digital double will not need to physically move around with you; most likely it will live somewhere in the cloud, just as much of your data already does. We can see its beginnings in virtual assistants such as Siri, Alexa and Google Assistant. At the heart of your digital double will be a model of you, learned from all the data you have ever generated in your interactions with the digital world, from desktop computers and Web sites to wearable devices and sensors in the environment such as smart speakers, thermostats, cell-phone towers and video cameras.

The better our learning algorithms become and the more personal data we feed them, the more accurate our digital doubles will get. Once we have the master algorithm and then couple it with continuous capture of your sensorimotor stream via an augmented reality headset and other personal sensors, your double will grow to know you better than your best friend.
The model and data will be maintained by a "data bank," not unlike a traditional bank that stores and invests your money. Many existing companies would surely like to provide that service for you. Google co-founder Sergey Brin has said that Google wants to be "the third half of your brain," but you probably would not want part of your brain to subsist by showing you ads. You might be better served by a new kind of company with fewer conflicts of interest or by a data union you form with like-minded people.

After all, the central worry about AI is not that it will spontaneously turn evil but that the humans who control it will misuse it (*cherchez l'humain*, as the French might say—"look to the human"). So your data bank's first duty will be to ensure that your model is never used against your interests. Both you and the data bank must be vigilant about monitoring AI crime because this technology will empower bad actors as much as anyone. We will need AI police (the Turing police, as William Gibson called it in his 1984 book *Neuromancer*) to catch the AI criminals.

If you have the misfortune of living under an authoritarian regime, this scenario could usher in unprecedented dangers because it will allow the government to monitor and restrain you like never before. Given the speed at which machine learning is progressing and the predictive policing systems already in use, the *Minority Report* scenario—where people are preemptively arrested when they are about to commit a crime—no longer seems far-fetched. Then there are the implications of inequality as the world adapts to the speed of life with digital doubles before all of us are able to afford one.

Our first duty, as individuals, will be not to become complacent and trust our digital doubles beyond their years. It is easy to forget that AIs are like autistic savants and will remain so for the foreseeable future. From the outside, AIs may seem objective, even perfect, but inside they are as flawed as we are or more, just in different ways. For example, AIs lack common sense and can easily make errors that a human never would, such as mistaking a person crossing the street for a windblown plastic bag. They are also liable to take our instructions too literally, giving us precisely what we asked for instead of what we actually wanted. (So think twice before telling your self-driving car to get you to the airport on time at all costs.)

Practically speaking, your digital double will be similar enough to you to take your place in all kinds of virtual interactions. Its job will not be to live your life for you but rather to make all the choices you do not have the time, patience or knowledge for. It will read every book on Amazon and recommend the few that you are most likely to want to read yourself. If you need a car, it will research the options and haggle with the car dealer's bots. If you are job hunting, it will interview itself for all the positions that fit your needs and then schedule live interviews for you for the most promising ones. If you get a cancer diagnosis, it will try all potential treatments and recommend the most effective ones. (It will be your ethical duty to use your digital double for the greater good by letting it take part in medical research, too.) And if you are seeking a romantic partner, your double will go on millions of virtual dates with all eligible doubles. The pairs that hit it off in cyberspace can then go on a date in real life.
Essentially your double will live out countless probable lives in cyberspace so that the single one you live in the physical world is likely to be the best version. Whether your simulated lives are somehow “real” and your cyberselves have a kind of consciousness (as portrayed in the plots of some *Black Mirror* episodes, for instance) are interesting philosophical questions.

Some people worry that this means that we are handing over control of our lives to computers. But it actually gives us more control, not less, because it allows us to make choices we could not before. Your model will also learn from the results of each virtual experience (Did you enjoy the date? Do you like your new job?) so that over time, it will become better at suggesting the things you would choose for yourself.

In fact, we are already accustomed to most of our decision making taking place without our conscious intervention because that is what our brains do now. Your digital double will be like a greatly expanded subconscious, with one key difference: Whereas your subconscious lives alone inside your skull, your digital double will continuously interact with those of other people and organizations. Everyone's doubles will keep trying to learn models of one another, and they will form a society of models, living at computer speeds, branching out in all directions, figuring out what we would do if we were there. Our machines will be our scouts, blazing a trail into the future for us as individuals and as a species. Where will they lead us? And where will we choose to go?

This article was originally published with the title "Our Digital Doubles"

MORE TO EXPLORE

The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Pedro Domingos. Basic Books, 2015.

The Digital Mind: How Science Is Redefining Humanity. Arlindo Oliveira. MIT Press, 2017.

FROM OUR ARCHIVES

Self-Taught Robots. Diana Kwon; March 2018.

ABOUT THE AUTHOR(S)

Pedro Domingos

Pedro Domingos is a professor of computer science at the University of Washington and author of *The Master Algorithm* (Basic Books, 2015). A fellow of the Association for the Advancement of Artificial Intelligence, he lives near Seattle.

Credit: Nick Higgins

Recent Articles

Why Businesses Embrace Machine Learning [Excerpt]

LATEST NEWS

BIOLOGY

Survey the Wildlife of the 'Great Indoors'

1 hour ago — Karen Hopkin

NATURAL DISASTERS

How to Evacuate Cities before Dangerous Hurricanes

5 hours ago — Leonardo Dueñas-Osorio, Devika Subramanian and Robert M. Stein

AUTOMOTIVE

This Federal Lab Works to Make Cars More Efficient, As Trump Pumps the Brakes