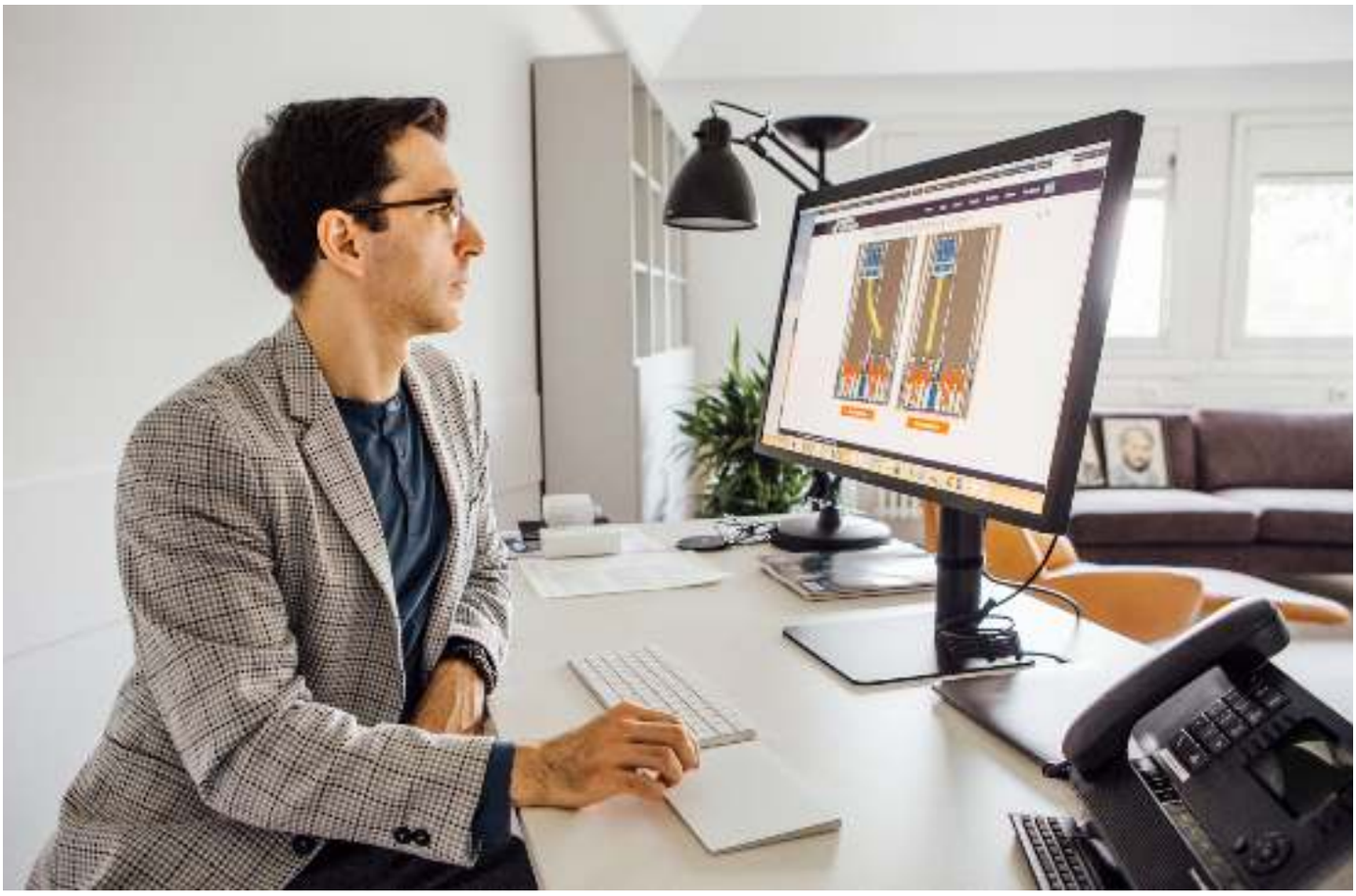
The Anthropologist of Artificial Intelligence

By John PavlusAugust 26, 2019

Q&A

The algorithms that underlie much of the modern world have grown so complex that we can't always predict what they'll do. Iyad Rahwan's radical idea: The best way to understand them is to observe their behavior in the wild.



Iyad Rahwan, the director of the Center for Humans and Machines at the Max Planck Institute for Human Development, argues that "machines impact our lives, and with AI, increasingly those machines have agency."

How do new scientific disciplines get started? For Iyad Rahwan, a computational social scientist with self-described "maverick" tendencies, it happened on a sunny afternoon in Cambridge, Massachusetts, in October 2017. Rahwan and Manuel Cebrian, a colleague from the MIT Media Lab, were sitting in Harvard Yard discussing how to best describe their preferred brand of multidisciplinary research. The rapid rise of artificial intelligence technology had generated new questions about the relationship between people and machines, which they had set out to explore. Rahwan, for example, had been exploring the question of ethical behavior for a self-driving car — should it swerve to avoid an oncoming SUV, even if it means hitting a cyclist? — in his Moral Machine experiment.

"I was good friends with Iain Couzin, one of the world's foremost animal behaviorists," Rahwan

said, “and I thought, ‘Why isn’t he studying online bots? Why is it only computer scientists who are studying AI algorithms?’

“All of a sudden,” he continued, “it clicked: We’re studying behavior in a new ecosystem.”

Two years later, Rahwan, who now directs the Center for Humans and Machines at the Max Planck Institute for Human Development, has gathered 22 colleagues — from disciplines as diverse as robotics, computer science, sociology, cognitive psychology, evolutionary biology, artificial intelligence, anthropology and economics — to publish a paper in *Nature* calling for the inauguration of a new field of science called “machine behavior.”

Directly inspired by the Nobel Prize-winning biologist Nikolaas Tinbergen’s four questions — which analyzed animal behavior in terms of its function, mechanisms, biological development and evolutionary history — machine behavior aims to empirically investigate how artificial agents interact “in the wild” with human beings, their environments and each other. A machine behaviorist might study an AI-powered children’s toy, a news-ranking algorithm on a social media site, or a fleet of autonomous vehicles. But unlike the engineers who design and build these systems to optimize their performance according to internal specifications, a machine behaviorist observes them from the outside in — just as a field biologist studies flocking behavior in birds, or a behavioral economist observes how people save money for retirement.

“The reason why I like the term ‘behavior’ is that it emphasizes that the most important thing is the observable, rather than the unobservable, characteristics of these agents,” Rahwan said.

He believes that studying machine behavior is imperative for two reasons. For one thing, autonomous systems are touching more aspects of people’s lives all the time, affecting everything from individual credit scores to the rise of extremist politics. But at the same time, the “behavioral” outcomes of these systems — like flash crashes caused by financial trading algorithms, or the rapid spread of disinformation on social media sites — are difficult for us to anticipate by examining machines’ code or construction alone.

“There’s this massively important aspect of machines that has nothing to do with how they’re built,” Rahwan said, “and has everything to do with what they do.”

Quanta spoke with Rahwan about the concept of machine behavior, why it deserves its own branch of science, and what it could teach us. The interview has been condensed and edited for clarity.

Why are you calling for a new scientific discipline? Why does it need its own name?

This is a common plight of interdisciplinary science. I don’t think we’ve invented a new field so much as we’ve just labeled it. I think it’s in the air for sure. People have recognized that machines impact our lives, and with AI, increasingly those machines have agency. There’s a greater urgency to study how we interact with intelligent machines.

Naming this emerging field also legitimizes it. If you’re an economist or a psychologist, you’re a serious scientist studying the complex behavior of people and their agglomerations. But people might consider it less important to study machines in those systems as well.

So when we brought together this group and coined this term “machine behavior,” we’re basically telling the world that machines are now important actors in the world. Maybe they don’t have free will or any legal rights that we ascribe to humans, but they are nonetheless actors that impact the world in ways that we need to understand. And when people of high stature in those fields sign up [as co-authors] to this paper, that sends a very strong signal.

You mentioned free will. Why even call this phenomenon "behavior," which seems to unnecessarily invite that association? Why not use a term like "functionality" or "operation"?

Some people have a problem with giving machines agency. For instance, Joanna Bryson from the University of Bath, she's always outspoken against giving machines agency, because she thinks that then you're removing agency and responsibility from human actors who may be misbehaving.

But for me, behavior doesn't mean that it has agency [in the sense of free will]. We can study the behavior of single-celled organisms, or ants. "Behavior" doesn't necessarily imply that a thing is super intelligent. It just means that our object of study isn't static — it's the dynamics of how this thing operates in the world, and the factors that determine these dynamics. So, does it have incentives? Does it get signals from the environment? Is the behavior something that is learned over time, or learned through some kind of copying mechanism?

Rahwan in his office at the Max Planck Institute, where he moved this summer. Behind him are portraits of John von Neumann (left), one of the founders of computer science, and Stanley Milgram, a key figure in social psychology.

Don't the engineers who design these agents make those decisions? Aren't they deterministically defining this behavior in advance?

They build the machines, program them, build the architecture of the neural networks and so on. They're engineering, if you like, the "brain" and the "limbs" of these agents, and they do study the behavior to some extent, but only in a very limited way. Maybe by looking at how accurate they are at classifying things, or by testing them in a controlled environment. You build the machine to perform a particular task, and then you optimize your machine according to this metric.

But its behavior is an open-ended aspect. And it's an unknown quantity. There are behaviors that manifest themselves across different timescales. So [when you're building it] maybe you focus on short timescales, but you can only know that long-timescale behavior once you deploy these machines.

Imagine that machine behavior is suddenly a mature field. What does it let us understand or do better?

I think we would be able to better diagnose emergent [technological] problems, and maybe anticipate them. For example, for somebody designing Facebook and its news feed algorithm, maybe we would have known early enough that this was going to lead to a far more polarized political sphere and a lot of spreading of misinformation. And maybe we'd have been able to build immunity into the system, so that it could self-correct.

Today, we're sort of figuring things out as we go. So the companies build a thing and then they launch it and then say, "Oh, wow, all of a sudden there's spam everywhere" or "there's misinformation everywhere." Or "people seem to hate each other and are yelling at each other all the time on Twitter."

Maybe there are lessons in nature that would have allowed us not just to engineer solutions, but also to detect those problems a little bit earlier. So, what's the [machine behavior] equivalent of colony collapse? Or what's an equivalent of speciation? Is there an analogy that we could use to anticipate problems that would undermine democracy, freedom of speech, and other values that we hold dear, as technology is introduced? That's the broad goal.

Why is this biology-inspired framework necessary?

There are lots of nature-inspired algorithms. For instance, an algorithm similar to swarming is being developed to allow inter-vehicle communication [between autonomous cars], so if there's an accident on the side of the road, they can smooth things out and you don't end up with traffic jams. So they'll go to nature and look at animal behaviors for inspiration, but then they'll just let the thing loose.

A behaviorist would notice things that emerge once you've let these cars into the wild. This isn't happening now, but imagine that this system that vehicles use to signal each other for minimizing traffic jams interacts with a car's reinforcement learning algorithm for optimizing its own behavior and causes some coordination pattern that wasn't preprogrammed. I can imagine an ecologist saying, "Oh, I know this, I've seen this kind of species of bees do this."

What happens when machines exhibit behavioral patterns that are completely new — that aren't analogous to anything biological or ecological?

That's going to be very important. Maybe [the field] will begin with a kind of "categorizing butterflies" phase, where we say, "Here are these types of machines, and they fall into these classes." Eventually we'd develop some kind of theory of how they change, and maybe also an understanding of how they feed off each other.

Does a washing machine have "behavior"? Is there some lower bound of autonomy or intelligence that makes a machine suitable for this kind of study?

Intelligence is such a hard word to describe. What we mean by that is any machine exhibiting behavior that is not completely derivable from its design. A washing machine has behavior —

sometimes it malfunctions, and of course it's very frustrating. But it's a fairly closed system in which all the kinds of functions and malfunctions can be predicted and described with precision, to a large degree. So you could say it has uninteresting behavior.

But then you could have a very simple algorithm that just repeats or retweets things. It could be rule based, but it can still exhibit interesting behaviors once it interacts with the world, with humans and with other algorithms. So while it's not very intelligent, the consequences of this simple behavior are much harder to foresee or to understand or to describe just based on its design.

How would a machine behaviorist study, say, self-driving cars differently than an engineer?

An engineer who is trying to improve the car toward some performance objective would, for example, test the car under different driving conditions, and they would test different components of the vehicle. The focus is very much on performance. But once this whole thing is built, you've got an agent — an actual, physical agent — moving around in the world and doing things. And you could do all kinds of things from the behavioral perspective when you're looking at this agent.

For example, let's assume that, overall, autonomous vehicles have managed to eliminate some 90% of fatal accidents. Let's assume that among those remaining fatalities, one carmaker is just killing far fewer passengers — but twice as many pedestrians. If we're not taking a behavioral perspective on autonomous cars, we wouldn't be looking for these things. We'd just be certifying that the car's systems were performing adequately according to certain benchmarks, but we would be missing this kind of emergent behavior that may be problematic. That's the kind of thing that economists or very highly quantitative social scientists could do, with absolutely no knowledge of the underlying engineering mechanisms of the vehicle.

What's the goal of machine behavior — to make these "agents" as predictable as washing machines?

Well, why do we care about understanding animal behavior? Part of it is pure curiosity, of course. But as a society we support science also because we want to understand the mechanisms that drive the world and use this understanding to predict these phenomena. We'd like the ecosystem to be healthy, but we also want to thrive economically, and so on.

I would say the same thing holds for machine behavior. Machines promise to vastly increase our economic productivity and bring information to our fingertips and improve our lives. But a lot of people are afraid of what these machines might do. Maybe they could impede our autonomy.

So again, I think an objective scientific understanding of the behavior of those machines is important and should inform a discussion about how we want to control those machines. Are they improving fast? Should we regulate them? Not regulate them? How exactly do we regulate them? And so on. So I see the study of machine behavior as the scientific task that complements the broader societal task of thriving with the machines.