ScienceTimes
The New York Times

What's Ahead For A.I.

Robot, Know Thyself

Building advanced mental capacities into machines hinges on a squishy question.

By OLIVER WHANG

Hod Lipson, a mechanical engineer who directs the Creative Machines Lab at Columbia University, has shaped most of his career around what some people in his industry have called the c-word.

On a sunny morning this past October, the Israeli-born roboticist sat behind a table in his lab and explained himself. "This topic was taboo," he said, a grin exposing a slight gap between his front teeth. "We were almost forbidden from talking about it — 'Don't talk about the c-word; you won't get tenure' — so in the beginning I had to disguise it, like it was something else."

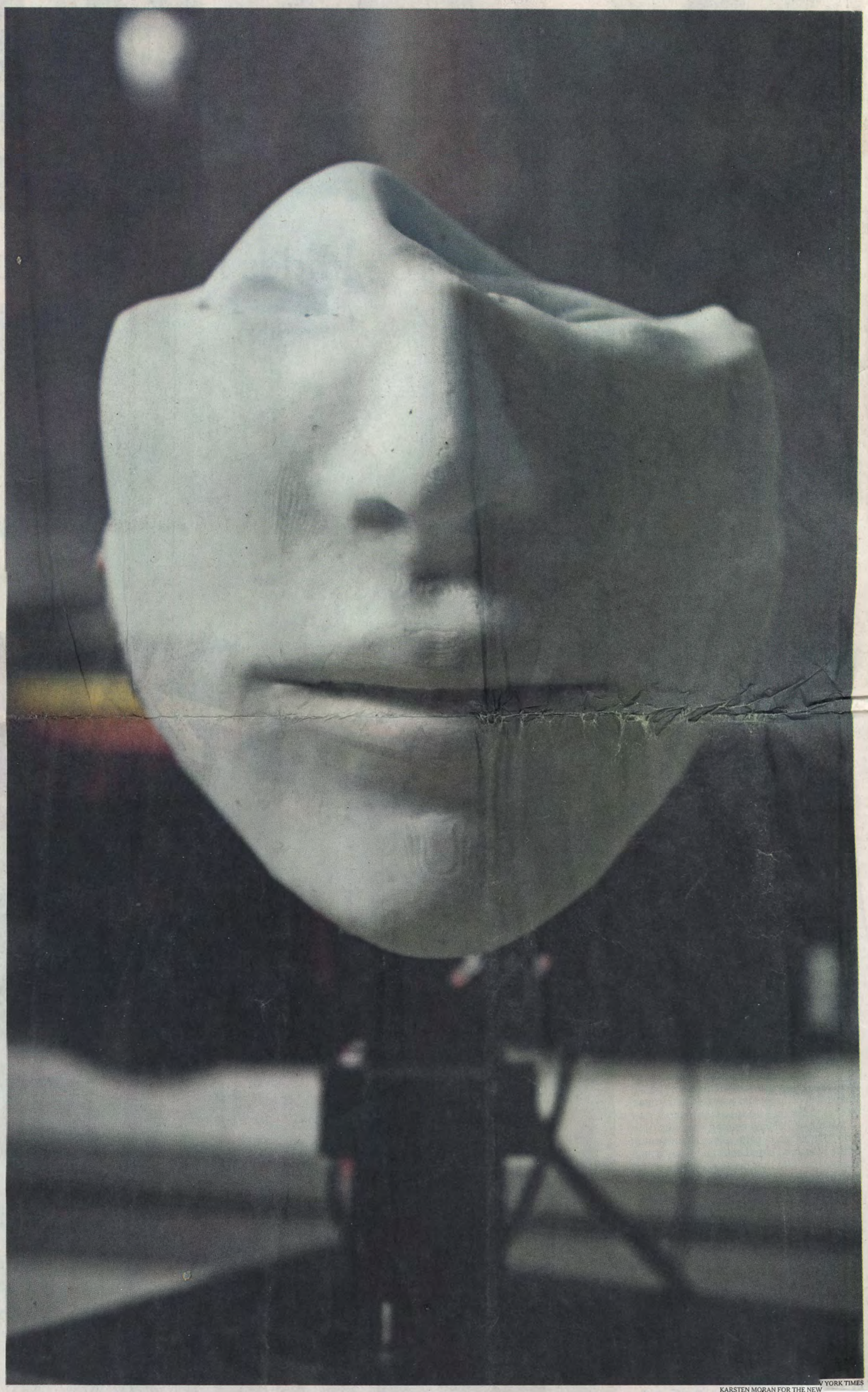
That was back in the early 2000s, when Dr. Lipson was an assistant professor at Cornell University. He was working to create machines that could note when something was wrong with their own hardware — a broken part, or faulty wiring — and then change their behavior to compensate for that impairment without the guiding hand of a programmer. Just as when a dog loses a leg in an accident, it can teach itself to walk again in a different way.

This sort of built-in adaptability, Dr. Lipson argued, would become more important as we became more reliant on machines. Robots were being used for surgical procedures, food manufacturing and transportation; the applications for machines seemed pretty much endless, and any error in their functioning, as they became more integrated with our lives, could spell disaster. "We're literally going to surrender our life to a robot," he said. "You want these machines to be resilient."

One way to do this was to take inspiration from nature. Animals, and particularly humans, are good at adapting to changes. This ability might be a result of millions of years of evolution, as resilience in response to injury and changing environments typically increases the chances that an animal will survive and reproduce. Dr. Lipson wondered whether he could replicate this kind of natural selection in his code, creating a generalizable form of intelligence that could learn about its body and function no matter what that body looked like, and no matter what that function was.

This kind of intelligence, if possible to create, would be flexible and fast. It would be

CONTINUED ON PAGE D4



Engineers in the Creative Machines Lab at Columbia University are working to create conscious robots, at least at a primitive level.

KARSTEN MORAN FOR THE NEW YORK TIMES

Can Robots Know That They're Robots?

CONTINUED FROM PAGE D1

as good in a tight situation as humans — better, even. And as machine learning grew more powerful, this goal seemed to become realizable. Dr. Lipson earned tenure, and his reputation as a creative and ambitious engineer grew. So, over the past couple of years, he began to articulate his fundamental motivation for doing all this work. He began to say the c-word out loud: He wants to create conscious robots.

"This is not just another research question that we're working on; this is *the* question," he said. "This is bigger than curing cancer. If we can create a machine that will have consciousness on par with a human, this will eclipse everything else we've done. That machine itself can cure cancer."

THE CREATIVE MACHINES LAB, on the first floor of the Seeley W. Mudd Building, is organized into boxes. The room itself is a box, broken into boxy workstations lined with boxed cubbies. Within this order, robots, and pieces of robots, are strewn about. A blue face staring blankly from a shelf; a green spiderlike machine splaying its legs out of a basket on the ground; a delicate dragonfly robot balanced on a worktable. This is the evolutionary waste of mechanical minds.

The first difficulty with studying the c-word is that there is no consensus around what it refers to. Such is the case with many vague concepts, like freedom, meaning, love and existence, but that domain is often supposed to be reserved for philosophers, not engineers. Some people have tried to taxonomize consciousness, explaining it by pointing to functions in the brain or some more metaphysical substances, but these efforts are hardly conclusive and give rise to more questions. Even one of the most widely shared descriptions of so-called phenomenal consciousness — an organism is conscious "if there is something that it is like to be that organism," as the philosopher Thomas Nagel put it — can feel unclear.

Wading directly into these murky waters might seem fruitless to roboticists and computer scientists. But, as Antonio Chella, a roboticist at the University of Palermo in Italy, said, unless consciousness is accounted for, "it feels like something is missing" in the function of intelligent machines.

The invocation of human features goes back to the dawn of artificial intelligence research in 1955, when a group of scientists at Dartmouth asked how machines could "solve kinds of problems now reserved for humans, and improve themselves." They wanted to model advanced capacities of the brain, like language, abstract thinking and creativity, in machines. And consciousness seems to be central to many of these capacities.

But trying to render the squishy c-word using tractable inputs and functions is a difficult, if not impossible, task. Most roboticists and engineers tend to skip the philosophy and form their own functional definitions. Thomas Sheridan, a professor emeritus of mechanical engineering at the Massachusetts Institute of Technology, said that he believed consciousness could be reduced to a certain process and that the more we find out about the brain, the less fuzzy the concept will seem. "What started out as being spooky and kind of religious ends up being sort of straightforward, objective science," he said.

(Such views aren't reserved for roboticists. Philosophers like Daniel Dennett and Patricia Churchland and the neuroscientist Michael Graziano, among others, have put forward a variety of functional theories of consciousness.)

Dr. Lipson and the members of the Creative Machines Lab fall into this tradition. "I need something that is totally buildable, dry, unromantic, just nuts and bolts," he said. He settled on a practical criterion for consciousness: the ability to imagine yourself in the future.

According to Dr. Lipson, the fundamental difference among types of consciousness — human consciousness and octopus consciousness and rat consciousness, for example — is how far into the future an entity is able to imagine itself. Consciousness exists on a continuum. At one end is an organism that has a sense of where it is in the world — some primitive self-awareness. Somewhere beyond that is the ability to imagine where your body will be in the future, and beyond that is the ability to imagine what you might eventually imagine.

"So eventually these machines will be able to understand what they are, and what they think," Dr. Lipson said. "That leads to emotions, and other things." For now, he added, "we're doing the cockroach version."

THE BENEFIT OF TAKING a stand on a functional theory of consciousness is that it allows for technological advancement.

One of the earliest self-aware robots to emerge from the Creative Machines Lab had four hinged legs and a black body with sensors attached at different points. By moving around and noting how the information entering its sensors changed, the robot created a stick figure simulation of itself. As the robot continued to move around, it used a machine-learning algorithm to improve the fit between its self-model and its actual body. The robot used this self-image to figure out, in simulation, a method of moving forward. Then it applied this method to its

body; it had figured out how to walk without being shown how to walk.

This represented a major step forward, said Boyuan Chen, a roboticist at Duke University who worked in the Creative Machines Lab. "In my previous experience, whenever you trained a robot to do a new capability, you always saw a human on the side," he said.

Recently, Dr. Chen and Dr. Lipson published a paper in the journal Science Robotics that revealed their newest self-aware machine, a simple two-jointed arm that was fixed to a table. Using cameras set up around it, the robot observed itself as it moved — "like a baby in a cradle, watching itself in the mirror," Dr. Lipson said. Initially, it had no sense of where it was in space, but over the course of a couple of hours, with the help of a powerful deep-learning algorithm and a probability model, it was able to pick itself out in the world. "It has this notion of self, a cloud," Dr. Lipson said.

Was it truly conscious, though?

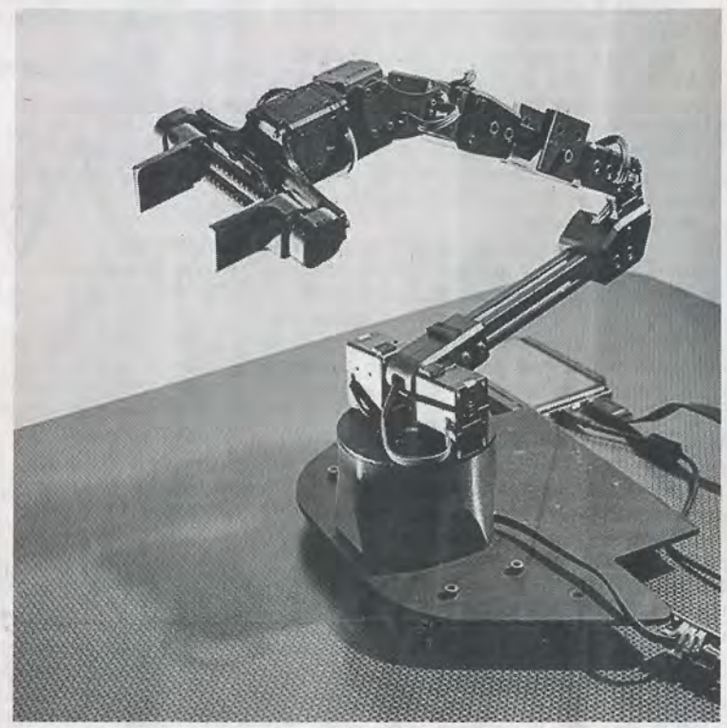
The risk of committing to any theory of consciousness is that doing so opens up the possibility of criticism. Sure, self-awareness seems important, but aren't there other key features of consciousness? Can we call something conscious if it doesn't feel conscious to us?

'We're literally going to surrender our life to a robot. You want these machines to be resilient.'

HOD LIPSON
CREATIVE MACHINES LAB

Dr. Chella believes that consciousness can't exist without language and has been developing robots that can form internal monologues, reasoning to themselves and reflecting on the things they see around them. One of his robots was recently able to recognize itself in a mirror, passing what is probably the most famous test of animal self-consciousness.

PHOTOGRAPHS BY KARSTEN MORAN FOR THE NEW YORK TIMES



Clockwise from top: a workstation at the Creative Machines Lab, with robot projects in various stages of assembly; a two-jointed robot arm that uses external cameras, a deep learning algorithm and a probability model to recognize itself in the world; Hod Lipson, the director of the Creative Machines Lab. "This is not just another research question that we're working on; this is the question," he said.

Joshua Bongard, a roboticist at the University of Vermont and a former member of the Creative Machines Lab, believes that consciousness doesn't consist just of cognition and mental activity but has an essentially bodily aspect. He has developed beings called xenobots, made entirely of frog cells linked together so that a programmer can control them like machines. According to Dr. Bongard, it's not just that humans and animals have evolved to adapt to their surroundings and interact with one another; our tissues have evolved to subserve these functions, and our cells have evolved to subserve our tissues. "What we are is intelligent machines made of intelligent machines made of intelligent machines, all the way down," he said.

This summer, around the same time that Dr. Lipson and Dr. Chen released their newest robot, a Google engineer said that the company's newly improved chatbot, called LaMDA, was conscious and deserved to be treated like a small child. This claim was met with skepticism, mainly because, as Dr. Lipson noted, the chatbot was processing "a code that is written to complete a task." There was no underlying structure of consciousness, other researchers said, only the illusion of consciousness.

Dr. Lipson added: "The robot was not self-aware. It's a bit like cheating."

But with so much disagreement, who's to say what counts as cheating?

ERIC SCHWITZGEBEL, a philosophy professor at the University of California, Riverside, who has written about artificial consciousness, said the issue with this general uncertainty was that, at the rate things are progressing, humankind would probably develop a robot that many people think is conscious before we agree on the criteria of consciousness. When that happens, should the robot be granted rights? Freedom? Should it be programmed to feel happiness when it serves us? Will it be allowed to speak for itself? To vote?

(Such questions have fueled an entire subgenre of science fiction in books by writers such as Isaac Asimov and Kazuo Ishiguro and in television shows like "Westworld" and "Black Mirror.")

Issues around so-called moral considerability are central to the animal rights debate. If an animal can feel pain, is killing it for its meat wrong? If animals don't experience things in the same ways that humans do, does that mean we can use them for our own enjoyment? Whether an animal has certain conscious capacities often seems to come to bear on whether it has certain rights, but there is no consensus around which capacities matter.

In the face of such uncertainty, Dr. Schwitzgebel has advocated for what he calls "the design policy of the excluded middle." The idea is that we should create only machines that we agree definitely do not matter morally — or that definitely do. Anything in the gray area of consciousness and mattering is liable to cause serious harm from one perspective or another.

Robert Long, a philosopher at the Future for Humanity Institute at Oxford University, supports this caution. He said A.I. development at big research labs and companies gave him the sensation of "hurtling toward a future filled with all sorts of unknown and vexing problems that we're not ready for." A famous one is the possibility of creating superintelligent machines that could annihilate the human population; the development of robots that are widely perceived to be conscious would add to the difficulties of tackling these risks.

"I'd rather live in a world where things are moving a lot more slowly and people think a lot more about what's being put in these machines," Dr. Long said.

But the downside of caution is slower technological development. Dr. Schwitzgebel and Dr. Long conceded that this more deliberate approach might impede the development of A.I. that was more resilient and useful. To scientists in the lab, such theorizing can feel frustratingly abstract.

"I think that we are not close to this risk," Dr. Chella said when asked about the risks of creating robots with conscious capacities similar to humans. Dr. Lipson added: "I am worried about it, but I think the benefits outweigh the risks. If we're going on this pathway where we rely more and more on technology, technology needs to become more resilient."

He added: "And then there's the hubris of wanting to create life. It's the ultimate challenge, like going to the moon." But a lot more impressive than that, he said later.

IN ONE OF THE workstations in the Creative Machines Lab, a self-aware robot arm started moving. Yuhang Hu, a graduate student, had initiated a mechanical sequence. For now, the robot wasn't watching itself and forming a self-model; it was just moving around randomly, twisting or shifting once every second. If it could be conscious, it wasn't yet.

Dr. Lipson leaned back in his chair and looked at the robot, then said to Mr. Hu, "Another thing we need to do is have this robot make a model of itself by just bumping into things."

Mr. Hu, his hair tussled, put his chin in his hand. "Yes, that's interesting," he said.

Dr. Lipson said, "Because even someone who is blind can form an image of itself."

"You can just put a box over it," Mr. Hu said.

"Right," Dr. Lipson said. "It has to be a rich enough environment, a playground."

The two scientists sat there thinking, or appearing to think, staring at the robot that continued to move on the table.

This, Dr. Lipson noted, is how research is done in his lab. The researchers look inward and notice some element of themselves — a bodily self-awareness, a sense of their surroundings, a self-consciousness around other people — and then try to put that element into a machine. "I want to push this as far as I can," he said. "I want a robot to think about its body, to think about its plans."

In a sense, it is the simplest of all robotics exercises, like something elementary school children do with old electronics. If you can do it with a retired printer, why can't you do it with your mind? Break it down, see how it works, and then try to build it back up again.