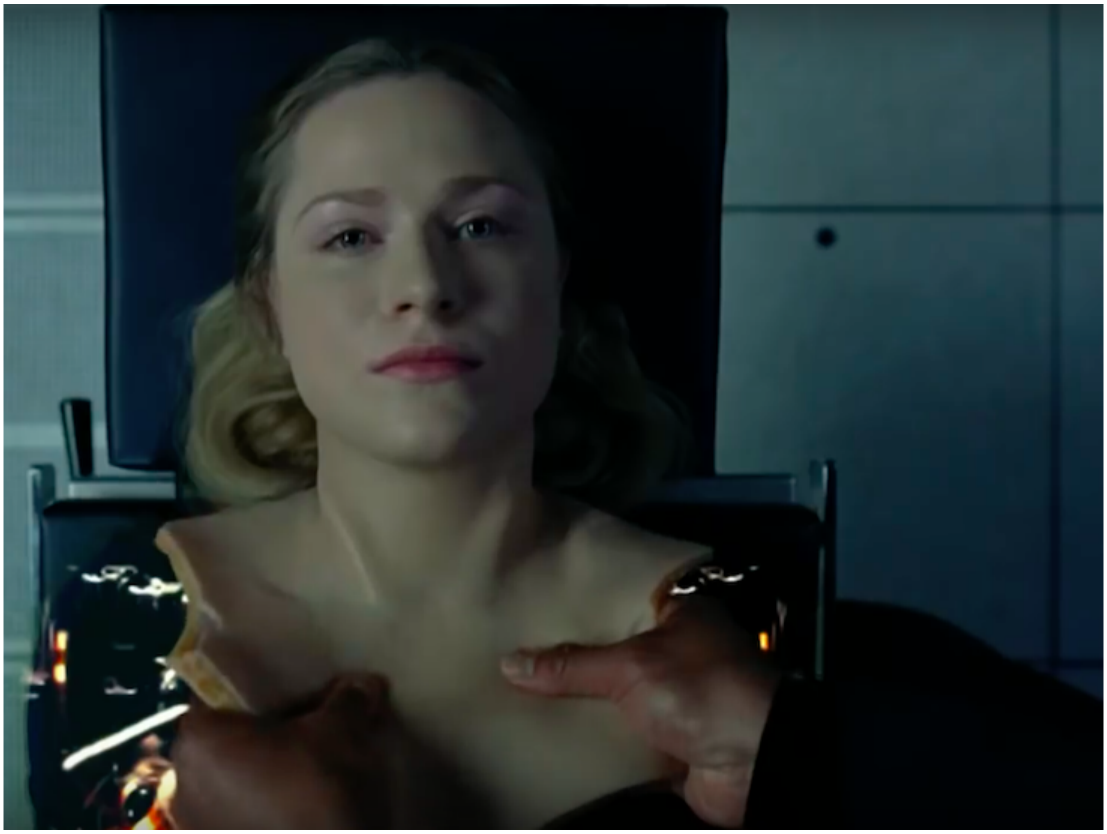
NAUTILUS

FACTS SO ROMANTIC ON IDEAS

Westworld Is Strikingly Real: AI Could Be Conscious and Unpredictable

POSTED BY CODY DELISTRATY ON DEC 11, 2016

ADD A COMMENT FACEBOOK TWITTER EMAIL SHARING



Photograph by HBO / YouTube

Westworld recently wrapped its first season with a few stunning twists and a stunning statistic: With a 12-million-viewer average, it was the most-watched first season of an original HBO show in the network's history. *Westworld* concerns a perverse theme park, styled in the fashion of the American Old West. The park's "hosts," artificially intelligent beings physically indistinguishable from humans, begin to remember the horrifying experiences inflicted on them by the park's "guests," the humans who pay to visit and do as they please, including raping and killing hosts.

Robert Ford (Anthony Hopkins), the fictional cofounder of Westworld, built the park's hosts with the ability to improvise and make decisions based on their environment—a vision of AI strikingly similar to the one held by Simon Stringer, the director of the Oxford Centre for Theoretical Neuroscience and Artificial Intelligence. Stringer is one of the field's leading thinkers, and like Ford, he says machines with some internalized spatial and causal model of the world could achieve an intuitive, human-like intelligence.

In my conversation with Stringer about *Westworld*, we discussed what makes AI seem human, the potential threat AI poses to humans, the role of self-modifying programming, and the importance of the Turing Test.

When might we have a *Westworld* of our own?

Well, I can tell you, no one is going to produce anything of that level within the next century. That's very long-term research. Anyone from my lab will know how long this kind of research takes. You can spend a lifetime working on one small area of the brain, trying to make sense of it.

Will AI ever achieve consciousness?

With adequate funding, I think within the next 20 years we'll see a move toward consciousness. We can get to basic cognitive systems that might be somewhere between a mouse and a rat, and that would be extraordinary to achieve because you would've built the first basic artificially conscious systems.

It's been shown back in the 1930s and '40s that even the brains of rats can learn a causal model of the world, and then they can exploit this world knowledge to guide their behavior. If a rat learns about the paths in its maze and then you block a path, it can take the next alternative route to get to a food reward. That sort of behavior flexibility, which is why our human brains learn about how the world works, means we can juggle this information to correct normal behavioral sequences. In other words, a core part of our own intelligence is that we learn a causal model of the world, and we exploit this world knowledge to guide our behaviors. So it's both memory and flexibility that will play a key role in future machines that can display genuine artificial intelligence and machine consciousness.

Will it be realistic to give AI biological components? This is the expectation of the *Westworld* writers, who have the inactive hosts being stored in refrigerated facilities.

I think that's going to come, but in a different way. At Oxford, people are developing artificial retinas. People are already developing neural implants to cope with, for example, sensory processing, and I think we'll see other sorts of implants that go directly into the brain that might help with memory and motor functions and so on. I think we're going to see the development of cyborgs, but we're going to see humans morphing into cyborgs, rather than AIs morphing into cyborgs. It's happening now. There are people walking around Oxford who were blind and now they literally have bionic eyes. I was growing up in 1970s when we had *The Six Million Dollar Man* and Steve Austin and it was all science fiction. Now they're actually creating bionic eyes, and they work. It's amazing.

Will it be possible to set strict limitations on AI while allowing them to access memories and self-modify their programming?

It's difficult to say what sort of control we might have. The one other thing about human beings and consciousness is that it gives us this great autonomy. That's what we're aiming for in machine consciousness. We want systems that really understand the world. We've got people like Stephen Hawking today expressing concerns about an existential threat to humanity by artificial intelligence. I can tell you: None of the systems today pose any such threat because they don't begin to understand their world. They will only represent a threat to us if they begin to understand the world and in the common sense way like the human brain.

You may be familiar with the work coming out of DeepMind. They rose to prominence when they developed some software that learned how to play Atari computer games. I was listening to someone being interviewed on the BBC World Service, an eminent professor of machine learning; I won't say which university he was from. He was saying that they were using what's called "model-free reinforcement learning." Through this kind of technique, the program doesn't actually learn about how the world works. It simply learns that if I see this sort of arrangement of pixels on the screen, I just do this. Its mapping is optimized by reward symbols. It's an algorithmic approach that doesn't learn.

Will machines ever be able to understand the world?

It's only the kinds of systems that display consciousness itself that would ever begin to understand the world. These machines really will display behavioral autonomy, and they will be unpredictable. They will be clever.

All the *Westworld* hosts looklike humans, but will AI beings in the real world necessarily looklike people?

Not at all. They could even have entirely virtual lives inside a virtual world. I think they have to exist in some sort of physical space, a simulacrum in a physical space, and they'll need some sort of morphology, so they'll need some sort of body with senses and drive in order for the system to self-organize, like a brain. But I think we'll be entirely free to design our own morphology and drives and the system will adapt.

In *Westworld*, the park's success is predicated on the fact that the hosts are indistinguishable from guests—what is it that makes artificial beings seemhuman?

I think we look first at their behavior. Neurological models give us deep insight into the very nature of consciousness itself. When you look at the world, you see millions of features at every spatial scale. If I look at a lamp stand opposite me, I see all the boundary elements—the curvature and how all the elements link together at every spatial scale. These visual pictures are thought to be represented at successive stages of the visual system. In other words, the simplest features are extracted at the early layers of the visual cortex. More complex features like whole objects and faces are extracted at later stages of processing. We now can see how all these features are related to each other; that's the binding problem that we have now solved. It's incredibly rich. We really just find ourselves now standing at the foot of a mountain of complexity.

But whenever you develop models of brain function, there are two halves of the coin. There's the actual neural-level architecture and then there's the sensory world. When you combine these two things, the emerging complexity is almost overwhelming. That's what we will be exploring in the years ahead. That's why it's so hard to say what is human, what is not.

Is non-physical AI the likely next step for the discipline? Like Microsoft's Xiaoice and other AI virtual chat bots?

You know what? That is such a good question. Let me give you an interesting answer to it. I think the Turing Test is a disaster for artificial intelligence. A lot of people who design chat bots are thinking about the Turing Test. Can we get this chat bot to replicate human speech so well that it would fool people? I think the way to true, genuine artificial intelligence and machine consciousness has to be to build up to that level by simulating simple life forms like mice and rats. For example, with rats, there's a lot of neurophysiology there to help build the models in the first place. I think trying to build a chat bot today that could solve the Turing Test is a dead end, and I think what we need to do is develop simple, artificial life forms. Just work it out step-by-step, gradually increasing the complexity and intelligence of these systems.

So physicality is key then for AI?

Yes, exactly.

One of the arguments made in favor of pursuing AI consciousness—which Ford makes—is that it's the next stage of human evolution. What do you think?

You look at humanity over tens of thousands of years; it's spread all over the surface of the Earth. I think that's inevitable because we're natural explorers. In the same vein, science is an exploratory activity; it's creative; it's a natural human instinct to try to understand the world and see what we can do and build wonderful engineering structures and wonderful art. The most extraordinary thing is the brain itself and consciousness itself. Our own sense of spiritual self-worth is based on this notion of consciousness. I think we'll want to understand that always. I think that's the driver that's driving neuroscience today, including theoretical neuroscience, which is computer simulation of the brain.

Additionally, we're trying to understand neurological disorders like autism and schizophrenia, depression, anxiety. That means that we want to understand brain functions so that we can make medical interventions. That, in itself, is what will drive and improve the understanding of brain function, which can then be translated to machine learning and robotics. I think it's inevitable that we will continue to make progress.

Cody Delistraty is a writer and historian based in Paris. He writes on books, culture, and interesting humans for places like The New York Times*,* The New Yorker, The Paris Review*, and* Aeon*. Follow him on Twitter @Delistraty.*

Get the Nautilus *newsletter*

The newest and most popular articles delivered right to your inbox!

Your name

Email address

SUBSCRIBE

☑ **New Chapters** *Thursdays* ☑ **Editor's Picks** *Sundays*

Watch: Ken Goldberg, a roboticist at UC Berkeley, explains the appeal of human-like AI.

2 COMMENTS - JOIN THE DISCUSSION

RELATED ISSUE

043: HEROES

OUR BETTER SELVES

SEE FULL ISSUE

RELATED FACTS SO ROMANTIC

IDEAS

This Simple Philosophical Puzzle Shows How Difficult It Is to Know Something

In the 1960s, the American philosopher Edmund Gettier devised a thought experiment that has become known as a

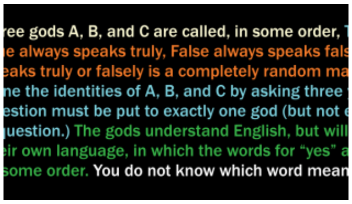
IDEAS

It's Time to Retire the "Trolley Problem"

In the 1960s, the moral philosopher Philippa Foot devised a thought experiment that would revolutionize her field. This ethical puzzle, today known

IDEAS

How to Solve the World's Hardest Logic Puzzle

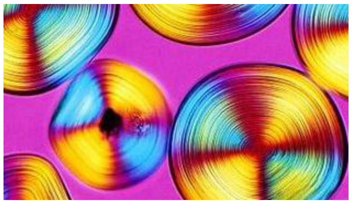


ree gods A, B, and C are called, in some order, T
e always speaks truly, False always speaks fals
eaks truly or falsely is a completely random ma
ne the identities of A, B, and C by asking three
estion must be put to exactly one god (but not
uestion.) The gods understand English, but will
eir own language, in which the words for "yes" a
some order. You do not know which word mean

While a doctoral student at Princeton University in 1957, studying under a founder of theoretical computer science, Raymond Smullyan would occasionally...

IDEAS

Look Through This Microscope and Tell Us What You See



As you look closer and closer at the world, you find more and more levels of organization. And at many of those steps, “Gettier case.” It shows that something’s...

READ MORE

as...

READ MORE

READ MORE

the view is fantastic. From butterfly...

READ MORE

NAUTILUS: SCIENCE CONNECTED

Nautilus is a different kind of science magazine. We deliver big-picture science by reporting on a single monthly topic from multiple perspectives. Read a new chapter in the story every Thursday.

© 2016 NautilusThink Inc, All rights reserved. Matter, Biology, Numbers, Ideas, Culture, *Connected*