"The Duck Test" (from "When Empires Fall II")

Excerpt from article by Dave Nilsen, February 1993, Challenge Magazine #69

The age-old duck test declares, 'If it walks like a duck and quacks like a duck, it's a duck.' Unfortunately, there is nothing resembling such a straightforward test for sentience.

The main reason for this is the emotional priority given to preexistence. There is a clear popular prejudice that something artificially created or manufactured cannot be sentient; that it is somehow instead merely a simulation or mimicry of sentience. On the other hand, a pre-existing organism that is discovered to have sophisticated mental processes is much more likely to be perceived as legitimately sentient.

What this demonstrates is a deep-seated psychological or spiritual belief that life or intelligence can only be created by something greater than we are, something which is infinite or eternal. Just as we finite creatures are an order of magnitude less substantial than the infinite (shadows of the eternal forms, as Plato would have it), our creations must again be an order of magnitude lower than we are, and the notion that we can rise this lesser order of energy up to our level is something that we emotionally resist. Whether this belief is true or not is arguable; that it is unconsciously subscribed to by a substantial portion of humans is not. Thus we find ... the persistent rejection of increasingly sophisticated robots and computers as anything other than machines.

A powerful adjunct to this belief is the 'threshold' issue. If we accept the premise that we can manufacture sentient life, then we should be able to define the threshold at which a mere thing becomes a *being*, something with *a priori* value. Then it follows that we can simply add the ingredient that takes something across the line into beinghood, we can take it back and pull it back across into thinghood. Once we know what that thing is, we can identify the people that were born without it, or the point at which a person loses it and becomes merely disposable. We don't really want to know what that thing is; after all, we have a vested interest in maintaining our anointed status as beings.

- ✓ See previously Plato's "Allegory of the Cave" (Second Coversheet, *supra*), see also the concept of uplifting lower forms or races common in science fiction.
- ✓ <u>Rhetorical Question</u>: Do you think that the source being a 1993 RPG magazine article somehow lessens the thought-value of the argument? Is there anything in the quote that makes you say, "Oh, that's just from some silly space game"? Like we are talking about Martian invaders?

The Dark Secret at the Heart of AI

Excerpts from article by Will Knight, April 11, 2017, MIT Technology Review

Last year, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors, and it showed the rising power of artificial intelligence. The car didn't follow a single instruction provided by an engineer or programmer. Instead, it relied entirely on an algorithm that had taught itself to drive by watching a human do it.

Getting a car to drive this way was an impressive feat. But it's also a bit unsettling, since it isn't completely clear how the car makes its decisions. Information from the vehicle's sensors goes straight into a huge network of artificial neurons that process the data and then deliver the commands required to operate the steering wheel, the brakes, and other systems. The result seems to match the responses you'd expect from a human driver. But what if one day it did something unexpected – crashed into a tree, or sat at a green light? As things stand now, it might be difficult to find out why. The system is so complicated that even the engineers who designed it may struggle to isolate the reason for any single action. And you can't ask it: there is no obvious way to design such a system so that it could always explain why it did what it did.

The mysterious mind of this vehicle points to a looming issue with artificial intelligence. The car's underlying AI technology, known as deep learning, has proved very powerful at solving problems in recent years, and it has been widely deployed for tasks like image captioning, voice recognition, and language translation. There is now hope that the same techniques will be able to diagnose deadly diseases, make million-dollar trading decisions, and do countless other things to transform whole industries.

But this won't happen – or shouldn't happen – unless we find ways of making techniques like deep learning more understandable to their creators and accountable to their users. Otherwise it will be hard to predict when failures might occur – and it's inevitable they will. That's one reason Nvidia's car is still experimental.

Already, mathematical models are being used to help determine who makes parole, who's approved for a loan, and who gets hired for a job. If you could get access to these mathematical models, it would be possible to understand their reasoning. But banks, the military, employers, and others are now turning their attention to more complex machine-learning approaches that could make automated decision-making altogether inscrutable. Deep learning, the most common of these approaches, represents a fundamentally different way to program computers. "It is a problem that is already relevant, and it's going to be much more relevant in the future," says Tommi Jaakkola, a professor at MIT who works on applications of machine learning. "Whether it's an investment decision, a medical decision, or maybe a military decision, you don't want to just rely on a 'black box' method."

There's already an argument that being able to interrogate an AI system about how it reached its conclusions is a fundamental legal right. Starting in the summer of 2018, the European Union may require that companies be able to give users an explanation for decisions that automated systems reach. This might be impossible, even for systems that seem relatively simple on the surface, such as the apps and websites that use deep learning to serve ads or recommend songs. The computers that run those services have programmed themselves, and they have done it in ways we cannot understand. Even the engineers who build these apps cannot fully explain their behavior.

This raises mind-boggling questions. As the technology advances, we might soon cross some threshold beyond which using AI requires a leap of faith. Sure, we humans can't always truly explain our thought processes either – but we find ways to intuitively trust and gauge people. Will that also be possible with machines that think and make decisions differently from the way a human would? We've never before built machines that operate in ways their creators don't understand. How well can we expect to communicate – and get along with – intelligent machines that could be unpredictable and inscrutable?

***

The resulting program, which the researchers named Deep Patient, was trained using data from about 700,000 individuals, and when tested on new records, it proved incredibly good at predicting disease. Without any expert instruction, Deep Patient had discovered patterns hidden in the hospital data that seemed to indicate when people were on the way to a wide range of ailments, including cancer of the liver. There are a lot of methods that are "pretty good" at predicting disease from a patient's records, says Joel Dudley, who leads the Mount Sinai team. But, he adds, "this was just way better."

At the same time, Deep Patient is a bit puzzling. It appears to anticipate the onset of psychiatric disorders like schizophrenia surprisingly well. But since schizophrenia is notoriously difficult for physicians to predict, Dudley wondered how this was possible. He still doesn't know. The new tool offers no clue as to how it does this. If something like Deep Patient is actually going to help doctors, it will ideally give them the rationale for its prediction, to reassure them that it is accurate and to justify, say, a change in the drugs someone is being prescribed. "We can build these models," Dudley says ruefully, "but we don't know how they work."

***

But it was not until the start of this decade, after several clever tweaks and refinements, that very large – or "deep" – neural networks demonstrated dramatic improvements in automated perception. Deep learning is responsible for today's explosion of AI. It has given computers extraordinary powers, like the ability to recognize spoken words almost as well as a person could, a skill too complex to code into the machine by hand. Deep learning has transformed computer vision and dramatically improved machine translation. It is now being used to guide all sorts of key decisions in medicine, finance, manufacturing – and beyond.

The workings of any machine-learning technology are inherently more opaque, even to computer scientists, than a hand-coded system. This is not to say that all future AI techniques will be equally unknowable. But by its nature, deep learning is a particularly dark black box.

You can't just look inside a deep neural network to see how it works. A network's reasoning is embedded in the behavior of thousands of simulated neurons, arranged into dozens or even hundreds of intricately interconnected layers. The neurons in the first layer each receive an input, like the intensity of a pixel in an image, and then perform a calculation before outputting a new signal. These outputs are fed, in a complex web, to the neurons in the next layer, and so on, until an overall output is produced. Plus, there is a process known as back-propagation that tweaks the calculations of individual neurons in a way that lets the network learn to produce a desired output.

The many layers in a deep network enable it to recognize things at different levels of abstraction. In a system designed to recognize dogs, for instance, the lower layers recognize simple things like outlines or color; higher layers recognize more complex stuff like fur or eyes; and the topmost layer identifies it all as a dog. The same approach can be applied, roughly speaking, to other inputs that lead a machine to teach itself: the sounds that make up words in speech, the letters and words that create sentences in text, or the steering-wheel movements required for driving.

- Commentary: Imagine that a bank is sued by a group of Plaintiffs for discrimination in hiring or promotion decisions or in loan and mortgage decisions – don't think a moment that all of these have not happened many times already. The bank defends the claims stating that the decisions were made by Deep Learning AI software and that, even though the results may appear to be a pattern of discrimination, in fact the bank had no such intention. See, it was all the computer? And the bank cannot explain how the software made these decisions. You cannot depose the AI software to ask, and the person-most-knowledgeable (PMK) of the subject sent over for deposition by the bank doesn't know and cannot explain it either. What can the Plaintiffs do? The pattern of discrimination, the appearance of discrimination, may be enough to settle the case favorably for the Plaintiffs, but direct testimony, documentary evidence like e-mails and texts, leading to intent would be a much better case. Except that it was all the computer, see? Might the Court conclude then that just as an owner is responsible if their dog bites someone – whether they loosed the dog on the victim or not – that the bank is responsible for the pattern, appearance and actual damage of seeming discrimination due to the AI? You cannot say, legally, well my dog bit him, I didn't bite him so I am not responsible, and we can only ever speculate why the dog bit someone, we cannot ask him, depose him. This would be a fascinating case, I would happily work on such case of first impression.

AI Is Inventing Languages Humans Can't Understand. Should We Stop It?
Excerpts from article by Mark Wilson, July 14, 2017, fastcodesign.com

Bob: "I can can I I everything else."

Alice: "Balls have zero to me to me to me to me to me to me to me to me to."

To you and I, that passage looks like nonsense. But what if I told you this nonsense was the discussion of what might be the most sophisticated negotiation software on the planet? Negotiation software that had learned, and evolved, to get the best deal possible with more speed and efficiency – and perhaps, hidden nuance – than you or I ever could? Because it is.

This conversation occurred between two AI agents developed inside Facebook. At first, they were speaking to each other in plain old English. But then researchers realized they'd made a mistake in programming.

"There was no reward to sticking to English language," says Dhruv Batra, visiting research scientist from Georgia Tech at Facebook AI Research (FAIR). As these two agents competed to get the best deal – a very effective bit of AI vs. AI dogfighting researchers have dubbed a "generative adversarial network" – neither was offered any sort of incentive for speaking as a normal person would. So they began to diverge, eventually rearranging legible words into seemingly nonsensical sentences.

"Agents will drift off understandable language and invent codewords for themselves," says Batra, speaking to a now-predictable phenomenon that's been observed again, and again, and again. "Like if I say 'the' five times, you interpret that to mean I want five copies of this item. This isn't so different from the way communities of humans create shorthands."

***

We Teach Bots To Talk, But We'll Never Learn Their Language

Facebook ultimately opted to require its negotiation bots to speak in plain old English. "Our interest was having bots who could talk to people," says Mike Lewis, research scientist at FAIR. Facebook isn't alone in that perspective. When I inquired to Microsoft about computer-to-computer languages, a spokesperson clarified that Microsoft was more interested in human-to-computer speech. Meanwhile, Google, Amazon, and Apple are all also focusing incredible energies on developing conversational personalities for human consumption. They're the next wave of user interface, like the mouse and keyboard for the AI era.

The other issue, as Facebook admits, is that it has no way of truly understanding any divergent computer language. "It's important to remember, there aren't bilingual speakers of AI and human languages," says Batra. We already don't generally understand how complex AIs think because we can't really see inside their thought process. Adding AI-to-AI conversations to this scenario would only make that problem worse.

But at the same time, it feels shortsighted, doesn't it? If we can build software that can speak to other software more efficiently, shouldn't we use that? Couldn't there be some benefit?

Because, again, we absolutely can lead machines to develop their own languages. Facebook has three published papers proving it. "It's definitely possible, it's possible that [language] can be compressed, not just to save characters, but compressed to a form that it could express a sophisticated thought," says Batra. Machines can converse with any baseline building blocks they're offered. That might start with human vocabulary, as with Facebook's negotiation bots. Or it could start with numbers, or binary codes. But as machines develop meanings, these symbols become "tokens"–they're imbued with rich meanings. As Dauphin points out, machines might not think as you or I do, but tokens allow them to exchange incredibly complex thoughts through the simplest of symbols. The way I think about it is with algebra: If A + B = C, the "A" could encapsulate almost anything. But to a computer, what "A" can mean is so much bigger than what that "A" can mean to a person, because computers have no outright limit on processing power.

"It's perfectly possible for a special token to mean a very complicated thought," says Batra. "The reason why humans have this idea of decomposition, breaking ideas into simpler concepts, it's because we have a limit to cognition." Computers don't need to simplify concepts. They have the raw horsepower to process them.

***

Why We Should Let Bots Gossip

But how could any of this technology actually benefit the world, beyond these theoretical discussions? Would our servers be able to operate more efficiently with bots speaking to one another in shorthand? Could microsecond processes, like algorithmic trading, see some reasonable increase? Chatting with Facebook, and various experts, I couldn't get a firm answer.

However, as paradoxical as this might sound, we might see big gains in such software better understanding *our* intent. While two computers speaking their own language might be more opaque, an algorithm predisposed to learn new languages might chew through strange new data we feed it more effectively. For example, one researcher recently tried to teach a neural net to create new colors and name them. It was terrible at it, generating names like Sudden Pine and Clear Paste (that clear paste, by the way, was labeled on a light green). But then they made a simple change to the data they were feeding the machine to train it. They made everything lowercase–because lowercase and uppercase letters were confusing it. Suddenly, the color-creating AI was working, well, pretty well! And for whatever reason, it preferred, and performed better, with RGB values as opposed to other numerical color codes.

Why did these simple data changes matter? Basically, the researcher did a better job at speaking the computer's language. As one coder put it to me, "Getting the data into a format that makes sense for machine learning is a huge undertaking right now and is more art than science. English is a very convoluted and complicated language and not at all amicable for machine learning."

***

Given that our connected age has been a bit of a disappointment, given that the internet of things is mostly a joke, given that it's no easier to get a document from your Android phone onto your LG TV than it was 10 years ago, maybe there is something to the idea of letting the AIs of our world just talk it out on our behalf. Because our corporations can't seem to decide on anything. But these adversarial networks? They get things done.