INTELLIGENCE IN THE UNEXPECTED

MOLLY WRIGHT STEENSON

In 1976, the architect Cedric Price designed an intelligent arts retreat centre for a site in Florida; it was never built. Generator was composed of 150 cubes, 12 feet on each side, and other components that could be moved around by mobile cranes according to the desires of Generator's users. Four years later, the programmer-architects John and Julia Frazer proposed four computer programmes for Generator. The Boredom programme, for instance, would redesign Generator's layouts if the parts had not been moved in a while. "If you kick a system, the very least you would expect it to do is kick you back," John Frazer wrote in his proposal to Price. In a handwritten postscript, he added, "You seemed to imply that we were only useful if we produced results that you did not expect. I think this leads to some definition of computer aids in general. At least one thing that you would expect from any half decent program is that it should produce at least one plan which you did not expect."[1]

At least one plan which you did not expect. The unexpected is central to our very idea of what intelligence is, whether human or artificial. As Marvin Minsky wrote in 1960, "To me 'intelligence' seems to denote little more than the complex of performances which we happen to respect, but do not understand."[2] One might observe loops and subroutines, but no "locus of intelligence."[3] His claim—that "we cannot assign all the credit to its programmer if the operation of a system comes to reveal structures not recognizable or anticipated by the programmer"—could have come from an engineer today who cannot explain why a deep learning algorithm works the way it does.

The unexpected results of algorithms press our assumptions about the worlds we've created. Janelle Shane, an electrical engineer, trains neural networks to do silly things. She discovered how the Microsoft Azure computer vision algorithm insinuates sheep (or "hallucinates" them) into green, rocky and foggy landscapes, even when none are present—clearly because the training data showed sheep on green pastures. "Bring sheep indoors, and they're labeled as cats. Pick up a sheep (or a goat) in your arms, and they're labeled as dogs," she writes. When she colours them orange, the algorithm parses them as flowers. As Shane explains, "If life plays by the rules, image recognition works well. But as soon as people—or sheep—do something unexpected, the algorithms show their weaknesses."[4]

That very outcome could be seen in 2016 when researchers at OpenAI used their Universe platform to train an AI agent to play CoastRunners, a boat race video game.[5] Typically, players complete clockwise laps in a small lagoon and pick up targets along the way. But the AI agent player ran its boat backwards, continuously caught itself on fire, smashed into other boats, never completed a normal lap—and got 20% more points than its human competitors. Winning! As the researchers note, their experiment is a cautionary tale for reinforcement learning: it's hard to get an agent to do exactly what you want it to do, and the outcomes could be not only unexpected but dangerous.

The unexpected and unwelcome are where most people direct their fears of AI—the drone that misstrikes, the AI that develops superintelligence and becomes uncharitable toward the humans that spun it into existence. But other unexpected, surreal responses—the video game boat that careens its way to a high score, the sheep that befuddle the algorithm—provide us with ways to understand the boundaries and permeability of machine learning, to understand how algorithms see the world, or us, or whether there's any difference.

1 John Frazer (2017) 'Letter to Cedric Price', in Molly Wright Steenson (ed.) Architectural Intelligence: How Designers and Architects Created the Digital Landscape . Cambridge: MIT Press, pp. 160.

2 Marvin Minsky (1961)., 'Steps toward Artificial Intelligence'. Proceedings of the I.R.E. 49. https://web.media.mit.edu/~minsky/papers/steps.html. Accessed 19th August 2018.

3 ibid

4 Janelle Shane (2018) Do neural nets dream of electric sheep?, Available at: http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep.(Accessed: 19th August 2018).

5 Jack Clark, Dario Amodei (2016) Faulty Reward Functions in the Wild, Available at: https://blog.openai.com/faulty-reward-functions/. (Accessed: 19th August 2018).