The Curious Cases of Unruly Robots

(Published in Science Reporter, October 2018)

On 31st July 2017, an unusual news item was reported by media all over the World which in particular alarmed the researchers of Artificial Intelligence (AI). Headlined, "Facebook's Artificial Intelligence Robots shut down after they start talking to each other in their own language", it reported that Facebook had abandoned an experiment after two artificially intelligent programs called Chatbots appeared to be chatting with each other in a strange language which nobody else really understood. The Chatbots created their own language using English words only, but which made no sense to the humans who programmed them to converse with each other.[i]

Researchers wanted to programme the Chatbots to negotiate and bargain with people, because they thought, rightly, that these skills which are essential for cooperation will enable the bots to work with humans. They started with a simple game in which two players were programmed to divide a collection of objects like hats, balls and books between themselves through a two-step program. First, they fed them with dialogues from thousands of games between humans to teach a sense of the language of negotiation, and then made them master their tactics and improve the bartering by trial and error through a technique called reinforcement learning. The objective of the game was to score as many points as possible. What followed was bizarre, but Facebook's researchers underplayed it by merely stating, "We found that updating the parameters of both agents led to divergence from human language."[ii]

The transcript of their conversation went something like the following:

Bob: i 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
Bob: you i everything else . . . . . .
Alice: balls have a ball to me to me to me to me to me to me to me
Bob: i i can i i i everything else . . . . . .
Alice: balls have a ball to me to me to me to me to me to me to me
Bob: i . . . . . .
Alice: balls have zero to me to me to me to me to me to me to me to me to
Bob: you i i i 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

Of course, the bots were NOT told to use only comprehensible English, and hence did not violate the instructions fed to them. Yet there seemed to be some rule in their apparently incomprehensible chat. The way they keep stressing themselves (me, i) appears to be a part of their negotiations, not simply a glitch in the way the messages were read out. In fact, some of the negotiations that were carried out in this bizarre manner even ended up successfully in concluding their negotiations, suggesting that they might have used a "shorthand" –a machine language which only they understood and invented to deceive the humans. The bots learned the rules of the game just like humans, pretending to be very interested in one specific item so as to pretend later that they were making a big sacrifice in giving it up. Facebook

chose to shut down the chats because "our interest was having bots who could talk to people", as the researcher Mike Lewis claimed, and not because they were scared, but it did not prevent the media to exude widespread fear about the future of AI and what it might do to humans with scary headlines like "Facebook AI creates its own language in creepy preview of our potential future", "Creepy Facebook bots talked to each other in a secret language", or "Facebook engineers panic, pull plug on AI after bots develop their own language".[iii]

Fear can metastasize in many ways and provide fodder for doomsayers to depict an impending apocalyptic doomsday scenarios for humanity. But Facebook's experiment isn't the first time that AI has invented new forms of language. Google has recently revealed that the AI it uses for its Translate tool has created its own language, into and out of which it would translate things, but it has found it okay and allowed it to proceed. In another similar experiment, researchers at AI research group OpenAI again used reinforced learning to challenge software bots to complete a series of tasks by communicating with other software agents, using a cooperative rather than competitive strategy.

Reinforced learning allows machines and software agents to automatically learn to determine the optimal behaviour in order to maximize performance. It is used in Google's AlphaGo program to defeat champion Go players. Go is an ancient Chinese board game like chess to learn discipline, concentration and balance. In October 2017, Google's artificial intelligence group, DeepMind, unveiled the latest version of its Go-playing program, AlphaGo Zero, which mastered three thousand years of knowledge of the game before inventing better moves of its own without any human help beyond being told about the rules, and all in the space of only three days.[iv] It learned purely by playing itself millions of times over, beginning with placing pieces on the Go board at random but improving incredibly fast as it discovered and mastered the winning strategies. Given that Go has a 19x19 board with 361 different places into which the black and white pieces can move, the total number of legal board arrangements is in the order of 10^{170} (ten times the total number of atoms in the observable universe), the self-learning feat is incredible and creepy.[v] A 2016 version of Alphago had beaten the grandmaster Ke Jie, the best player of the game 3-1, and the new AlphaGo Zero beat the earlier version 100-0. It means that self-learning machines can be used to solve problems humans cannot.

In OpenAI also, the robots learned to collaborate and communicate through trial and error, remembering the symbols, words and signals that helped them to achieve a goal and storing them in a private recurrent neural network to be used later. In the process, the robots created their own language to communicate with each other. As the researchers introduced tougher tasks, the language evolved to become more and more complex, with the robots eventually learning to work together by composing sentences comprising multiple words. Based on this ability of robots, researchers hope to build a translator bot capable of translating their communications for humans. "We hope that this research into growing a language will let us develop machines that have their own language tied to their own lived experience," an OpenAI post said, "We think that if we slowly increase the complexity of their environment, and the range of actions the agents themselves are allowed to take, it's possible they'll create an expressive language which contains concepts beyond the basic verbs and nouns that evolved here."[vi]

The scientific community is deeply divided over whether AI can spin out of control and about the impact of integration of humanity with AI. Elon Musk, the lead designer of SpaceX and CEO of Tesla, voiced these fears, "I have exposure to the most cutting edge AI, and I think people should be really concerned by it….AI is a fundamental risk to the existence of human civilization". Facebook's Mark Zuckerberg dismissed such fear as "irresponsible", "I think people who are naysayers and try to drum up these doomsday scenarios — I just, I don't understand it. It's really negative and in some ways I actually think it is pretty irresponsible". But the potential and reality of AI technologies need a deeper understanding as machine learning permeates ever more spheres of human activities and becomes more and more pervasive.

In June 2016, a Robot in a research facility in Perm, Russia called Promobot IR77 made headlines across the globe. It was programmed to move freely about a room and return to a designated spot, learning from experience and surroundings. Scientists were training the robot to act as a tour guide. A researcher had left the facility without properly closing the door and somehow, the robot fled out the open door, travelled 45 metres onto a nearby street, before running out of battery. It was stuck there for 40 minutes, creating traffic chaos. Police asked to remove the robot away from the crowded area, even trying to handcuff it. It was like replaying the 'Number 5' runaway military robot from 1986 Hollywood sci-fi comedy 'Short Circuit'.

IR77 had apparently developed an insatiable yearning for freedom, for a few weeks later, it was still persistently trying to flee towards the exit of the facility, even after undergoing extensive reprogramming to avoid the issue. The frustrated scientists were considering shutting it down – rather killing it, if it persisted in this weird behavior, though some doubted it as a publicity stunt. "We're considering recycling the IR77 because our clients hiring it might not like that specific feature", the Promobot co-founder Oleg Kivokurtsev assured.[vii]

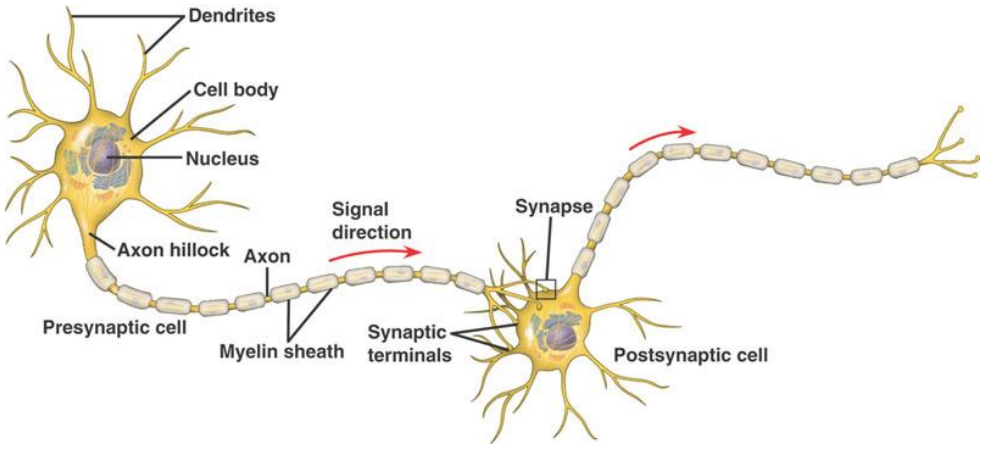
This again was not the first time that a robot seemed to be getting a mind of its own and thinking of escaping. At Hinterstoder in Kirchdorf, Austria, a cleaning robot, christened Irobot Roomba 760, reportedly 'committed suicide' by switching itself on, and climbing on to a kitchen hotplate where it was burned to death. Firemen called to put off the blaze found its remains on the hotplate and confirmed that after its job was done, the house-owner had switched it off and left the house while leaving the robot on the kitchen sideboard. The robot had somehow reactivated itself and moved onto the hotplate by pushing a cooking pot out of its way and set itself ablaze. Apparently it had enough of the chores and decided "enough was enough".[viii] It reminded one of the famous lines from Czechoslovak author Karel Capek's play R.U.R. (Rossum Universal Robots) which introduced the term "Robot", "Occasionally they seem to go off their heads…. They'll suddenly sling down everything they're holding, stand still, gnash their teeth!— and then they have to go into the stamping-mill. It's evidently some breakdown in the mechanism."

Adventurism among AI is not unknown, and it is easy to paint a doomsday picture like what Capek predicted in his play, "The era of man has come to its end. A new epoch has arisen! Domination by robots!" But as AI develops at a breakneck pace and keeps breaking new grounds almost on a daily basis, we need to understand them better. AI is a product of human brains, but often we do not really know how AI works, even though we can programme machine learning.

Machine learning is the simulation of human intelligence by machines, and is a little different from AI which is larger in scope. A machine learns by using algorithms that discover patterns and generate insights from data. It is a multi-step process: learning or acquisition of information from the analysis of data and information, discovering rules for using data and information, reasoning or using these rules to approximate solutions, self-correction and prediction of future behaviour. It bypasses the need to be programmed at every stage, being able to programme itself. Within machine learning, deep learning is another advanced field which attempts to enable machines to learn and think like humans. The more the data it is exposed to, the better the patterns it discovers and the smarter it gets, and finally start making predictions. However, machines cannot generalize abstractions from information – that till now has remained essentially an attribute of human consciousness only. Expert systems, speech recognition, machine vision, driverless cars, Google's language translation, Facebook's facial recognition or Snapchat's image altering filters are all examples of machine learning which got a boost after technology has enabled generation and processing of enormous volumes of data coupled with inexpensive storage.

To understand machine learning, AI systems rely on artificial neural networks (ANNs), by trying to simulate the way the human brain learns. It is difficult because we have little knowledge of how our brain, which is an incredibly efficient learning machine, actually learns. There is also no universally agreed definition of intelligence, but most agree that it fundamentally involves learning, understanding and application of the knowledge learned to achieve goals.

The human brain is composed of about 100 billion nerve cells called neurons, which receive stimuli from external environment or inputs from sensory organs through dendrites, which are like tendrils propagating from one end of the neurons to the other end of which are attached the axons which, like optical fibres, carry the messages to other neurons. These messages are electrical impulses generated within the cell by the action of the ions within and outside the cell membranes.



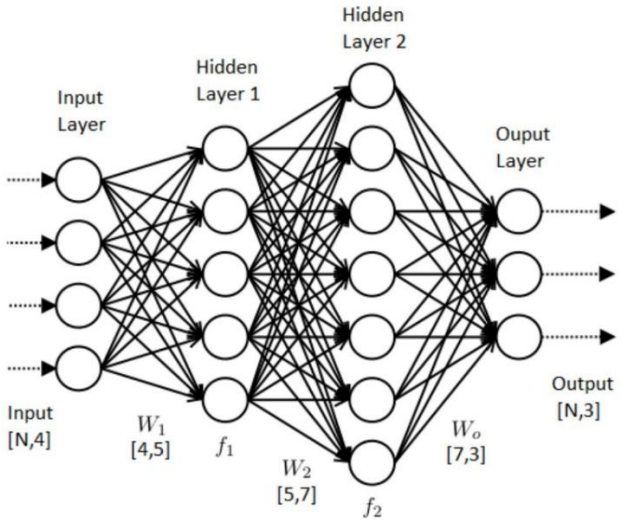
Structure of Neuron

Each axon has several terminals to connect to the dendrites of neurons that receive messages; these terminals are fitted with tiny sac-like structures called synaptic vesicles which are filled with chemicals called neurotransmitters. The junction between two neurons consists of a minute gap – called the synaptic gap - across which neural impulses pass by diffusion of a neurotransmitter; this junction is called a

synapse. Synapses act like gates, and regulate the flow of information. Through them, neurons connect to the incredibly complex network of neurons that can process information and pass it down to all parts of the body via the nerves. The strength of the synaptic connections depends on their activity and is altered when the brain learns something, in turn altering the brain's neural structure itself. Stronger synaptic connections characterize the strength of learning indicated by the higher frequency of recall of the information learnt, while weak synaptic connections makes it harder to recall a piece of information.

ANN mimics the human brain using silicon atoms and their interconnections as neurons and dendrites, creating multiple nodes through which neurons interact. The nodes can take input data and perform operations on them, results of which are transmitted to other neurons via links. The output at each node is called its node value. Each link is associated with a weight, and learning takes place by processing the inputs through the nodes and links whose numbers change with the volume of data. A message will be transmitted by one neuron to another across the node if the sum of the weighted input signals into it exceeds a predefined threshold, using a mathematical function called activation. In trying to replicate the learning process of a human brain, ANN adjusts the weighted connections between the neurons in the network, in a process akin to the strengthening and weakening of the synaptic connections that enables learning in humans.

Between the input and output interfaces are hidden several internal layers, called hidden layers since no AI programmer can interact with them. Mathematically, each node layer is represented as a function and each link as a weight, with the web of links between the layers being represented as a matrix. The matrix architecture enables simultaneous processing of a set of multiple inputs to yield a set of specific output corresponding to each input. ANNs can have anywhere between a few and several hundred layers with anywhere from a dozen to several thousand nodes. Once the number of layers increases enormously and the web becomes extremely complex and intricate, machine learning gives way to deep learning, giving the model far higher learning and predictive capabilities.



Legend: [No. of nodes, No. of links]; W_1, W_2 = Weights of links reaching internal layers 1 and 2 respectively.
Source: Mimlitz, Zac, https://techblog.viasat.com/using-artificial-neural-networks-to-analyze-trends-of-a-global-aircraft-fleet/

(The matrix of the above network can be represented mathematically as: Output = f_2 (f_1 (Input. W_1).W_2). W_o, where f_1 and f_2 are functions of the corresponding nodes and weights. f_1 can be taken as a summation function $\sum X_i W_i$, sum of the products of inputs and their corresponding weights, and if the sum exceeds a predefined threshold defined by the function f_2, the neuron fires and the output is obtained.)

The objective of learning is to map an input into the most correct output by minimizing the possible errors. Learning can be unsupervised, supervised or reinforced, depending on the nature of algorithm used and the adjustment of weights. The supervised learning process is something like the way we learn in real life, assisted by a teacher who corrects our mistakes and teaches us the rules. Neural networks here employ a 'trainer' program that tells it the expected output from an input or supplies the correct value of the output, and then based on its deviation from the actual value outputted by the network, an 'error' value is computed which is fed back into the network. This process is known as 'back-propagation'. Each layer then analyses the error and adjusts the threshold and weights, minimizing the error at each run till the error becomes minimum. At this stage, the network no longer needs the trainer and can run autonomously, making it an unsupervised learning process. In reinforced learning, the ANN makes a decision by observing the environment, but instead of providing a target output, a reward is given based on the performance of the system which automatically adjusts the weights so as to maximize the rewards through a system of trial and error. This is a rather simplistic description of the actual process which is extremely complex. There are ANNs without any feedback mechanism also; these are called Feed-Forward ANNs in which the information flows unidirectionally; these are used in pattern generation, recognition and classification software.

Following the above algorithm, it is virtually impossible to build a self-aware machine that is capable of thinking, deciding and acting independently - the way the robots described earlier seemed to have behaved. But the more the machines use deep leaning, the better they get at their jobs, and achieve mastery through self-learning algorithms – like Google Search – till one does not know the precise processes by which a query gets a response.

The human brain has perfected this self-learning process through millions of years of evolution, internalizing the self-learning algorithms in their DNA, first competing with each other and then learning to maximize the goals through co-operation of the cells which grouped together to specialize in different tasks, thereby increasing productivity and chances for beneficial mutation. At the social level also, we have inculcated this cooperation in order to maximize knowledge and innovation. There is no reason to think why self-learning machines would not discover the benefits of cooperation sooner or later and replicate the human condition. Then the self-learning algorithms would tend to become incredibly complex and challenge – and defy – the understanding of their creators. When they do so, they will tend to develop a persona of their own, and they might seem scary. Machines have astounding "intellectual capacity, but they have no soul", Capek wrote nearly a century ago. Future machines may look as if they really have a 'soul', but not one that will destroy. Instead it will build, and aid us in making the human condition a little better.

[i] Griffin, Andrew, https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html.

[ii] Simonite, Tom, https://www.wired.com/story/facebooks-chatbots-will-not-take-over-the-world/

[iii] Clifford, Catherine, https://www.cnbc.com/2017/08/02/facebook-bot-controversy-highlights-peoples-fears-about-ai-and-robots.html

[iv] https://www.theguardian.com/technology/deepmind

[v] "The latest AI can work things out without being taught", *The Economist*, Oct 21st 2017.

[vi] Sulleyman, Aatif, https://www.independent.co.uk/life-style/gadgets-and-tech/news/robots-create-new-language-to-work-together-a7636041.html

[vii] https://www.sciencealert.com/the-same-robot-keeps-trying-to-escape-a-lab-in-russia-even-after-reprogramming

[viii] https://www.mirror.co.uk/news/weird-news/intelligent-robot-remembers-learns-could-8248559