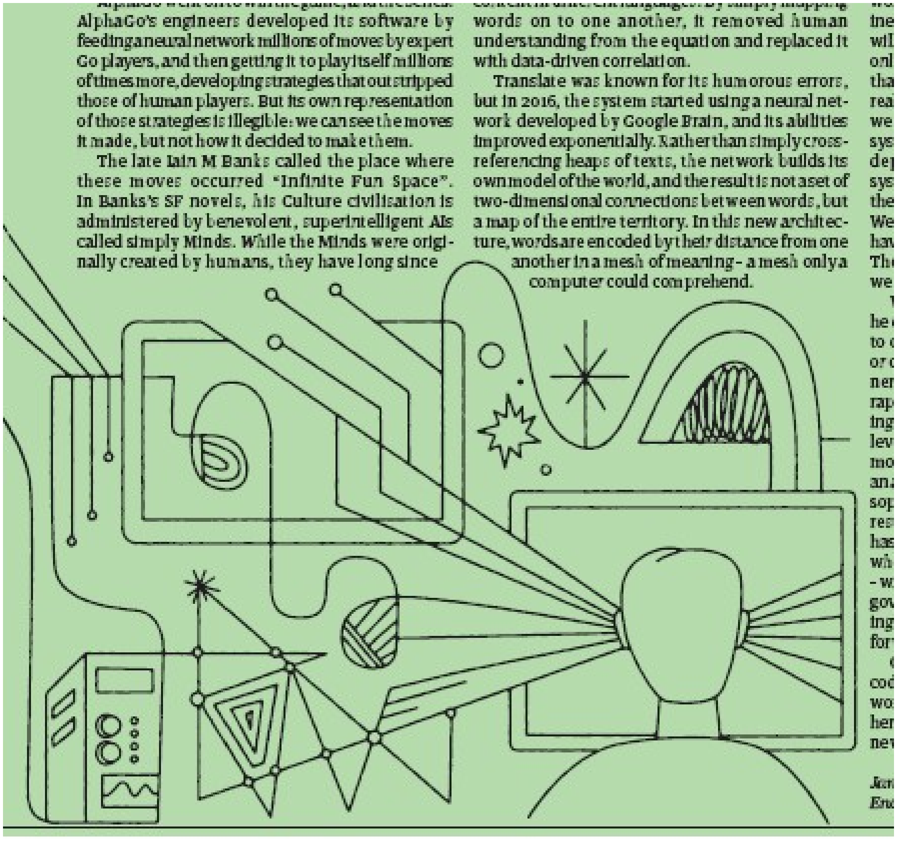
Slave to the algorithm

Has technology evolved beyond our control?

The Guardian Weekly · 22 giugno 2018 · By James Bridle

The voice-activated gadget in the corner of your bedroom suddenly laughs maniacally and sends a recording of your pillow talk to a colleague. The clip of Peppa Pig your toddler is watching on YouTube unexpectedly une descends into bloodletting and death. deat The social network you use to keep in touch with old school friends turns out to be influencing elections elec and fomenting coups.



AlphaGo's engineers developed its software by feeding a neural network millions of moves by expert Go players, and then getting it to play itself millions of times more, developing strategies that outstripped those of human players. But its own representation of those strategies is illegible: we can see the moves it made, but not how it decided to make them.

The late Iain M Banks called the place where these moves occurred "Infinite Fun Space". In Banks's SF novels, his Culture civilisation is administered by benevolent, superintelligent AIs called simply Minds. While the Minds were originally created by humans, they have long since

words on to one another, it removed human understanding from the equation and replaced it with data-driven correlation.

Translate was known for its humorous errors, but in 2016, the system started using a neural network developed by Google Brain, and its abilities improved exponentially. Rather than simply cross-referencing heaps of texts, the network builds its own model of the world, and the result is not a set of two-dimensional connections between words, but a map of the entire territory. In this new architecture, words are encoded by their distance from one another in a mesh of meaning – a mesh only a computer could comprehend.

Something So strange has happened to our way of thinking thin – and as a result, even stranger things are happening happ to the world. We have come to believe that everything is computable and can be resolved by the th application of new technologies. But these technologies tech are not neutral facilitators: they embody emb our politics and biases, they extend beyond the boundaries b of nations and legal jurisdictions and increasingly exceed the understanding of even their creators. As a result, we understand less and less about the world as these powerful technologies assume assu more control over our everyday lives.

Across A the sciences and society, in politics and education, in warfare and commerce, new technologies tech are not merely augmenting our abilities, abili they are actively shaping and directing them, them for better and for worse. If we do not understand und how complex technologies function then their potential is more easily captured by selfish elites elite and corporations. The results of this can be seen all around us. There is a causal relationship between betw the complex opacity of the systems we encounter enco every day and global issues of inequality, violence, viole populism and fundamentalism.

Instead In of a utopian future in which technological advancement adva casts a dazzling, emancipatory light on the world, we seem to be entering a new dark age characterised by ever more bizarre and unforeseen unfo events. The Enlightenment ideal of distributing distr more information ever more widely has not led us to greater understanding and growing grow peace, but instead seems to be fostering social soci divisions, distrust, conspiracy theories and post-factual politics. To understand what is happening, happ it's necessary to understand how our technologies tech have come to be, and how we have come com to place so much faith in them.

In the 1950s, a new symbol began to creep into the diagrams diag drawn by electrical engineers to describe the systems s they built: a fuzzy circle, or a puffball, or a thought t bubble. Eventually, its form settled into the shape s of a cloud. Whatever the engineer was working work on, it could connect to this cloud, and that's all you y needed to know. The other cloud could be a power pow system, or a data exchange, or another network of computers. Whatever. It didn't matter. The cloud clou was a way of reducing complexity; it allowed you to focus on the issues at hand. Over time, as networks netw grew larger and more interconnected, the cloud clou became more important. It became a business buzzword buzz and a selling point. It became more than engineering engi shorthand; it became a metaphor.

Today T the cloud is the central metaphor of the internet: inter a global system of great power and energy that nevertheless retains the aura of some- omething thin numinous, almost impossible to grasp. rasp. We work w in it; we store and retrieve stuff from it; it is something we experience all the e time without with really understanding what it is. s. But there's ther a problem with this metaphor: the he cloud clou is not some magical faraway place, ce, made mad of water vapour and radio waves, es, where whe everything just works. It is a physi- sical infrastructure in consisting of phone lines, nes, fibre optics, satellites, cables on the ocean n floor, and vast warehouses filled with computers, which consume huge amounts of water and energy. Absorbed into the cloud are many of the previously weighty edifices of the civic sphere: the places where we shop, bank, socialise, borrow books and vote. Thus obscured, they are rendered less visible and less amenable to critique, investigation, preservation and regulation.

Over the last few decades, trading floors around the world have fallen silent, as people are replaced by banks of computers that trade automatically. Digitisation meant that trades could happen faster and faster. HighFrequency Trading (HFT) algorithms, designed by former physics PhD students to take advantage of millisecond advantages, entered the market, and traders gave them names such as The Knife. These algorithms were capable of eking out fractions of a cent on every trade, and they could do it millions of times a day. Yet something deeply weird is occurring within these massively accelerated, opaque markets. On 6 May 2010, the Dow Jones opened lower than the previous day, falling slowly over the next few hours in response to the debt crisis in Greece. But at 2.42pm, the index started to fall rapidly. In less than five minutes, more than 600 points were wiped off the market. At its lowest point, the index was nearly 1,000 points below the previous day's average, a difference of almost 10% of its total value and the biggest single-day fall in the market's history. By 3.07pm, in just 25 minutes, it recovered almost all of those 600 points, in the largest and fastest swing ever.

In the chaos of those 25 minutes, 2bn shares, worth $56bn, changed hands. Even more worryingly, many orders were executed at what the Securities and Exchange Commission called "irrational prices": as low as a penny, or as high as $100,000. The event became known as the "flash crash", and it is still being investigated and argued over years later.

One report by regulators found that highfrequency traders exacerbated the price swings. Among the various HFT programs, many had hard-coded sell points: prices at which they were programmed to sell their stocks immediately. As prices started to fall, groups of programs were triggered to sell at the same time. As each waypoint was passed, the subsequent price fall triggered another anot set of algorithms to automatically sell their stocks, producing a feedback effect. As a result, prices fell faster than any human trader trad could react to. While experienced market mar players might have been able to stabilise s the crash by playing a longer game, the t machines, faced with uncertainty, got g out as quickly as possible.

Flash crashes are now a recognised feature fe of augmented markets, but are still poorly understood. u In October 2016, algorithms reacted to negative news headlines about Brexit negotiations by sending the pound down 6% against the dollar in under two minutes, before recovering almost immediately. Knowing which particular headline, or which particular algorithm, caused the crash is next to impossible. When one haywire algorithm started placing and cancelling orders that ate up 4% of all traffic in US stocks in October 2012, one commentator was moved to comment wryly that "the motive of the algorithm is still unclear".

At 1.07pm on 23 April 2013 Associated Press sent a tweet to its 2 million followers: "Breaking: Two Explosions in the White House and Barack Obama is injured." The message was the result of a hack later claimed by the Syrian Electronic Army, a group affiliated to Syrian president Bashar al-Assad. AP and other journalists quickly flooded the site with alerts that the message was false. The algorithms following breaking news stories had no such discernment, however. At 1.08pm, the Dow Jones went into a nosedive. Before most human viewers had even seen the tweet, the index had fallen 150 points in under two minutes, and bounced back to its earlier value. In that time, it erased $136bn in equity market value.

Computation is increasingly layered across, and hidden within, every object in our lives, and with its expansion comes an increase in opacity and unpredictability. One of the touted benefits of Samsung's line of "smart fridges" in 2015 was their integration with Google's calendar services, allowing owners to schedule grocery deliveries from the kitchen. It also meant that hackers who gained access to the then inadequately secured machines could read their owner's Gmail passwords. Researchers in Germany discovered a way to insert malicious code into Philips's wifi-enabled Hue lightbulbs, which could spread from fixture to fixture throughout a building or even a city, turning the lights rapidly on and off and – in one possible scenario – triggering photosensitive epilepsy. This is the approach favoured by Byron the Bulb in Thomas Pynchon's Gravity's Rainbow, an act of grand revolt by the little machines against the tyranny of their makers. Once-fictional possibilities for technological violence are being realised by the Internet of Things.

In Kim Stanley Robinson's novel Aurora, an intelligent spacecraft carries a human crew from Earth to a distant star. The journey will take multiple lifetimes, so one of the ship's jobs is to ensure that the humans look after themselves. When their fragile society breaks down, threatening the mission, the ship deploys safety systems as a means of control: it is able to see everywhere

The cloud is the central metaphor of the internet: ternet: a global system of great power and energy, almost impossible to grasp

through sensors, open or seal doors at will, speak so loudly through its communications equipment that it causes physical pain, and use fire suppression systems to draw down the level of oxygen in a particular space.

This is roughly the same suite of operations available now from Google Home and its partners: a network of internet-connected cameras for home security, smart locks on doors, a thermostat capable of raising and lowering the temperature in individual rooms, and a fire and intruder detection system that emits a piercing emergency alarm. Any successful hacker would have the same powers as the Aurora does over its crew, or Byron over his hated masters.

Before dismissing such scenarios as the fever dreams of science fiction writers, consider again the rogue algorithms in the stock exchanges. These are not isolated events, but everyday occurrences within complex systems. The question then becomes, what would a rogue algorithm or a flash crash look like in the wider reality?

Would it look, for example, like Mirai, a piece of software that brought down large portions of the internet for several hours on 21 October 2016? When researchers dug into Mirai, they discovered it targets poorly secured internet connected devices – from security cameras to digital video recorders – and turns them into an army of bots. In just a few weeks, Mirai infected half a million devices, and it needed just 10% of that capacity to cripple major networks for hours.

Mirai, in fact, looks like nothing so much as Stuxnet, another virus discovered within the industrial control systems of hydroelectric plants and factory assembly lines in 2010. Stuxnet was a military- grade cyberweapon; when dissected, it was found to be aimed specifically at Siemens centrifuges, and designed to go off when it encountered a facility that possessed a particular number of such machines. That number corresponded with one particular facility: the Natanz nuclear facility in Iran. When activated, the program would quietly degrade crucial components of the centrifuges, causing them to break down and disrupt the Iranian enrichment programme.

The attack was apparently partially successful, but the effect on other infected facilities is unknown. To this day, despite obvious suspicions, nobody knows where Stuxnet came from, or who made it. Nobody knows for certain who developed Mirai, either, or where its next iteration might come from, but it might be there, right now, breeding in the CCTV camera in your office, or the wifi-enabled kettle in the corner of your kitchen.

Or perhaps the crash will look like a string of blockbuster movies pandering to rightwing conspiracies and survivalist fantasies, from quasifascist superheroes (Captain America and the Batman series) to justifications of torture and assassination (Zero Dark Thirty, American Sniper). In Hollywood, studios run their scripts through the neural networks of a company called Epagogix, a system trained on the unstated preferences of millions of moviegoers developed over decades in order to predict which lines will push the right – meaning the most lucrative – emotional buttons. Algorithmic engines enhanced with data from Netflix, Hulu, YouTube and others, with access to the minute-by-minute preferences of millions of video watchers acquire a level of cognitive insight undreamed of by previous regimes. Feeding directly on the frazzled, binge-watching desires of news-saturated consumers, the network turns on itself, reflecting, reinforcing and heightening the paranoia inherent in the system.

Game developers enter endless cycles of updates and in-app purchases directed by testing interfaces and real-time monitoring of players' behaviours. They could have such a fine-grained grasp of dopamine-producing neural pathways that teenagers would die of exhaustion in front of their computers, unable to tear themselves away.

Or perhaps the flash crash will look like literal nightmares broadcast across the network for all to see? In the summer of 2015, the sleep disorders clinic of an Athens hospital was busier than it had ever been: the country's debt crisis was in its most turbulent period. Among the patients were top politicians and civil servants, but the machines they spent the nights hooked up to, monitoring their breathing, their movements, even the things they said out loud in their sleep, were sending that information, together with their personal medical details, back to the manufacturers' diagnostic data farms in northern Europe. What whispers might escape from such facilities?

We are able to record every aspect of our daily lives by attaching technology to the surface of our bodies. Smart bracelets and smartphone apps with integrated step counters and galvanic skin response monitors track not only our location, but every breath and heartbeat, even the patterns of our brainwaves. Users are encouraged to lay their phones beside them on their beds at night, so that their sleep patterns can be recorded. Where does all this data go, who owns it, and when might it come out? Data on our dreams, our night terrors and early morning sweating jags, the very substance of our unconscious selves, turn into more fuel for systems both pitiless and inscrutable.

Or perhaps the flash crash in reality looks exactly like everything we are experiencing right now: rising economic inequality, the breakdown of the nation-state and the militarisation of borders, totalising global surveillance and the curtailment of individual freedoms, the triumph of transnational corporations and neurocognitive capitalism, the rise of far-right groups and nativist ideologies, and the degradation of the natural environment. None of these are the direct result of novel technologies, but all of them are the product of a general inability to perceive the wider, networked effects of individual and corporate actions accelerated by opaque, technologically augmented complexity.

Video game developers could have such a fine grasp of neural pathways that teenagers would be unable to tear themselves away

In New York in 1997, world chess champion Garry Kasparov faced off for the second time against Deep Blue, a computer specially designed by IBM to beat him. When he lost, he claimed some of Deep Blue's moves were so intelligent and creative that they must have been the result of human intervention. But we understand why Deep Blue made those moves: its process for selecting them was ultimately one of brute force, a massively parallel architecture of custom-designed chess chips, capable of analysing 200m board positions per second. Kasparov was not out-thought, merely outgunned.

By the time the Google Brain–powered AlphaGo software took on the Korean professional Go player Lee Sedol in 2016, something had changed. In the second of five games, AlphaGo played a move that stunned Sedol, placing one of its stones on the far side of the board. "That's a very strange move," said one commentator. "I thought it was a mistake," said another. Fan Hui, a seasoned Go player who had been the first professional to lose to the machine six months earlier, said: "It's not a human move. I've never seen a human play this move."

AlphaGo went on to win the game, and the series. AlphaGo's engineers developed its software by feeding a neural network millions of moves by expert Go players, and then getting it to play itself millions of times more, developing strategies that outstripped those of human players. But its own representation of those strategies is illegible: we can see the moves it made, but not how it decided to make them.

The late Iain M Banks called the place where these moves occurred "Infinite Fun Space". In Banks's SF novels, his Culture civilisation is administered by benevolent, superintelligent AIs called simply Minds. While the Minds were originally created by humans, they have long since redesigned and rebuilt themselves and become allpowerful. Between controlling ships and planets, directing wars and caring for billions of humans, the Minds also take up their own pleasures. Capable of simulating entire universes within their imaginations, some Minds retreat for ever into Infinite Fun Space, a realm of meta-mathematical possibility, accessible only to superhuman artificial intelligences.

Many of us are familiar with Google Translate, which was launched in 2006, using a technique called statistical language inference. Rather than trying to understand how languages actually worked, the system imbibed vast corpora of existing translations: parallel texts with the same content in different languages. By simply mapping words on to one another, it removed human understanding from the equation and replaced it with data-driven correlation.

Translate was known for its humorous errors, but in 2016, the system started using a neural network developed by Google Brain, and its abilities improved exponentially. Rather than simply crossreferencing heaps of texts, the network builds its own model of the world, and the result is not a set of two-dimensional connections between words, but a map of the entire territory. In this new architecture, words are encoded by their distance from one another in a mesh of meaning – a mesh only a
computer could comprehend.

While a human can draw a line between the words "tank" and "water" easily enough, it quickly becomes impossible to draw on a single map the lines between "tank" and "revolution", between "water" and "liquidity", and all of the emotions and inferences that cascade from those connections. The map is thus multidimensional, extending in more directions than the human mind can hold. As one Google engineer commented, when pursued by a journalist for an image of such a system: “I do not generally like trying to visualise thousand-dimensional vectors in three-dimensional space.” This is the unseeable space in which machine learning makes its meaning. Beyond that which we are incapable of visualising is that which we are incapable of even understanding.

In the same year, other researchers at Google Brain set up three networks called Alice, Bob and Eve. Their task was to learn how to encrypt information. Alice and Bob both knew a number – a key, in cryptographic terms – that was unknown to Eve. Alice would perform some operation on a string of text, and then send it to Bob and Eve. If Bob could decode the message, Alice’s score increased; but if Eve could, Alice’s score decreased.

Over thousands of iterations, Alice and Bob learned to communicate without Eve breaking their code: they developed a private form of encryption like that used in private emails today. But crucially, we don’t understand how this encryption works. Its operation is occluded by the deep layers of the network. What is hidden from Eve is also hidden from us. The machines are learning to keep their secrets.

How we understand and think of our place in the world, and our relation to one another and to machines, will ultimately decide where our technologies will take us. We cannot unthink the network; we can only think through and within it. The technologies that inform and shape our present perceptions of reality are not going to go away, and in many cases we should not wish them to. Our current life support systems on a planet of 7.5 billion people and rising depend on them. Our understanding of those systems, and of the conscious choices we make in their design, remain entirely within our capabilities. We are not powerless, not without agency. We only have to think, and think again, and keep thinking. The network – us and our machines and the things we think and discover together – demands it.

When Garry Kasparov was defeated back in 1997, he didn’t give up the game. A year later, he returned to competitive play with a new format: advanced, or centaur, chess. In advanced chess, humans partner, rather than compete, with machines. And it rapidly became clear that something very interesting resulted from this approach. While even a midlevel chess computer can today wipe the floor with most grandmasters, an average player paired with an average computer is capable of beating the most sophisticated supercomputer – and the play that results from this combination of ways of thinking has revolutionised the game. It remains to be seen whether cooperation is possible – or will be permitted – with the kinds of complex machines and systems of governance now being developed, but understanding and thinking together offer a more hopeful path forward than obfuscation and dominance.

Our technologies are extensions of ourselves, codified in machines and infrastructures, in frameworks of knowledge and action. Computers are not here to give us all the answers, but to allow us to put new questions, in new ways, to the universe.