# 11

# Intellectual Debt

## *With Great Power Comes Great Ignorance*

*Jonathan Zittrain**

The boxes for prescription drugs typically include an insert of tissue-thin paper folded as tight as origami. For the bored or the preternaturally curious who unfurl it, there's a sketch of the drug's molecular structure using a notation that harkens to high school chemistry, along with 'Precautions' and 'Dosage and Administration' and 'How Supplied'. And for many drugs, under 'Clinical Pharmacology', one finds a sentence like this one for the wakefulness drug Provigil, after the subheading 'Mechanism of Action': 'The mechanism(s) through which modafinil promotes wakefulness is unknown.'[1] That sentence alone might provoke wakefulness without assistance from the drug. How is it that something could be so studied and scrutinized to find its way to regulatory approval and widespread prescribing, while we don't know how it works?

The answer is that industrial drug discovery has long taken the form of trial-and-error testing of new substances in, say, mice. If the creatures' condition is improved with no obvious downside, the drug may be suitable for human testing. Such a drug can then move through a trial process and earn approval. In some cases, its success might inspire new research to fill in the blanks on mechanism of action. For example, aspirin was discovered in 1897, and an explanation of how it works followed in 1995.[2] That, in turn, has spurred some research leads on making better pain relievers through something other than trial and error.

This kind of discovery – answers first, explanations later – accrues what I call 'intellectual debt'. We gain insight into what works without knowing why it works. We can put that insight to use immediately, and then tell ourselves we'll figure out the details later. Sometimes we pay off the debt quickly; sometimes, as with aspirin, it takes a century; and sometimes we never pay it off at all.

Be they of money or ideas, loans can offer great leverage. We can get the benefits of money – including use as investment to produce more wealth – before we've actually earned it, and we can deploy new ideas before having to plumb them to bedrock truth.

Indebtedness also carries risks. For intellectual debt, these risks can be quite profound, both because we are borrowing as a society, rather than individually, and because new technologies of Artificial Intelligence (AI) – specifically, machine learning – are bringing the old model of drug

* This chapter is based on an essay found at https://perma.cc/CN55-XLCW?type=image. A derivative version of it was published in *The New Yorker*, 'The Hidden Costs of Automated Thinking' (23 July 2019) www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.

[1] RxList, 'Provigil', (*RxList*, 16 June 2020) www.rxlist.com/provigil-drug.htm.

[2] 'How Aspirin Works', (1995) 15(1) *The University of Chicago Chronicle* http://chronicle.uchicago.edu/950817/aspirin.shtml.



https://doi.org/10.1017/9781009207898.014 Published online by Cambridge University Press

discovery to a seemingly unlimited number of new areas of inquiry. Humanity's intellectual credit line is undergoing an extraordinary, unasked-for bump up in its limit.

To understand the problems with intellectual debt despite its boon, it helps first to consider a sibling: engineering's phenomenon of technical debt.

In the summer of 2012, the Royal Bank of Scotland applied a routine patch to the software it used to process transactions. It went poorly. Millions of customers could not withdraw their money, make payments, or check their balances.[3] One man was held in jail over a weekend because he couldn't make bail.[4] A couple was told to leave their newly-purchased home when their closing payment wasn't recorded.[5] A family reported that a hospital threatened to remove life support from their gravely ill daughter after a charity's transfer of thousands of dollars failed to materialize.[6] The problem persisted for days as the company tried to figure out what had gone wrong, reconstruct corrupted data, and replay transactions in the right order.

RBS had fallen victim to technical debt. Technical debt arises when systems are tweaked hastily, catering to an immediate need to save money or implement a new feature, while increasing long-term complexity. Anyone who has added a device every so often to a home entertainment system can attest to the way in which a series of seemingly sensible short-term improvements can produce an impenetrable rat's nest of cables. When something stops working, this technical debt often needs to be paid down as an aggravating lump sum – likely by tearing the components out and rewiring them in a more coherent manner.

Banks are particularly susceptible to technical debt because they computerized early and invested heavily in mainframe systems that were, and are, costly and risky to replace. Their core systems still process trillions of dollars using software written in COBOL, a programming language from the 1960s that's no longer taught in most universities.[7] Those systems are now so intertwined with Web extensions, iPhone apps, and systems from other banks, that figuring out how they work all over again, much less eliminating them, is daunting. Consulting firms like Accenture have charged firms like the Commonwealth Bank of Australia hundreds of millions to dollars to make a clean break.[8]

Two crashes of Boeing's new 737 Max 8 jets resulted in the worldwide grounding of its Max fleet. Analysis so far points to a problem of technical debt: The company raced to offer a more efficient jet by substituting in more powerful engines, while avoiding a comprehensive redesign in order to fit the Max into the original 737 genus.[9] That helped speed up production in a number of ways, including bypassing costly recertifications. But the new engines had a tendency to push the aircraft's nose up, possibly causing it to stall. The quick patch was to alter the aircraft's software to automatically push the nose down if it were too far up. Pilots were then expected to know what to do if the software itself acted wrongly for any reason, such as receiving

[3] M Hickman, 'NatWest and RBS Customers May Receive Compensation as 'Computer Glitch' Drags into Sixth Day' *Independent* (26 June 2012) www.telegraph.co.uk/finance/personalfinance/bank-accounts/9352573/NatWest-customers-still-unable-to-see-bank-balances-on-sixth-day-of-glitch.html.

[4] 'RBS Computer Problems Kept Man in Prison' (*BBC News*, 26 June 2012) www.bbc.com/news/uk-18589280.

[5] L Bachelor, 'NatWest Problems Stop Non-Customers Moving into New Home' *The Guardian* (22 June 2012) www.theguardian.com/money/2012/jun/22/natwest-problems-stop-non-customers-home?newsfeed=true.

[6] J Hall, 'NatWest Computer Glitch: Payment to Keep Cancer Girl on Life Support Blocked' *The Telegraph* (25 June 2012) www.telegraph.co.uk/finance/personalfinance/bank-accounts/9352532/NatWest-computer-glitch-payment-to-keep-cancer-girl-on-life-support-blocked.html.

[7] A Irrera, 'Banks Scramble to Fix Old Systems as IT 'Cowboys' Ride into Sunset' *Reuters* (11 April 2017) www.reuters.com/article/us-usa-banks-cobol/banks-scramble-to-fix-old-systems-as-it-cowboys-ride-into-sunset-idUSKBN17C0D8.

[8] Ibid.

[9] N Rivero 'A String of Missteps May Have Made the Boeing 737 Max Crash-Prone' (Quartz, 18 March 2019) https://qz.com/1575509/what-went-wrong-with-the-boeing-737-max-8/.

the wrong information about nose position from the plane's sensors. A small change occasioned another small change which in turn forced another awkward change, pushing an existing system into unpredictable behavior. While the needed overall redesign would have been costly and time consuming, and would have had its own kinks to work out, here the alternative of piling on debt contributed to catastrophe.

Enter a renaissance in long-sleepy areas of AI based on machine learning techniques. Like the complex systems of banks and aircraft makers, these techniques bear a quiet, compounding price that may not seem concerning at first, but will trouble us later. Machine learning has made remarkable strides thanks to theoretical breakthroughs, zippy new hardware, and unprecedented data availability. The distinct promise of machine learning lies in suggesting answers to fuzzy, open-ended questions by identifying patterns and making predictions. It can do this through, say, 'supervised learning', by training on a bunch of data associated with already-categorized conclusions. Provide enough labeled pictures of cats and non-cats, and an AI can soon distinguish cats from everything else. Provide enough telemetry about weather conditions over time, along with what notable weather events transpired, and an AI might predict tornadoes and blizzards. And with enough medical data and information about health outcomes, an AI can predict, better than the best physicians can, whether someone newly entering a doctor's office with pulmonary hypertension will live to see another year.[10]

Researchers have pointed out thorny problems of technical debt afflicting AI systems that make it seem comparatively easy to find a retiree to decipher a bank system's COBOL.[11] They describe how machine learning models become embedded in larger ones and can then be forgotten, even as their original training data goes stale and their accuracy declines.

But machine learning doesn't merely implicate technical debt. There are some promising approaches to building machine learning systems that, in fact, can offer some explanations[12] – sometimes at the cost of accuracy – but they are the rare exceptions. Otherwise, machine learning is fundamentally patterned like drug discovery, and it thus incurs intellectual debt. It stands to produce answers that work, without offering any underlying theory. While machine learning systems can surpass humans at pattern recognition and predictions, they generally cannot explain their answers in human-comprehensible terms. They are statistical correlation engines – they traffic in byzantine patterns with predictive utility, not neat articulations of relationships between cause and effect. Marrying power and inscrutability, they embody *Arthur C. Clarke's* observation that any sufficiently advanced technology is indistinguishable from magic.[13]

But here there is no *David Copperfield* or *Ricky Jay* who knows the secret behind the trick. No one does. Machine learning at its best gives us answers as succinct and impenetrable as those of a Magic 8-Ball – except they appear to be consistently right. When we accept those answers without independently trying to ascertain the theories that might animate them, we accrue intellectual debt.

[10] TJW Dawes and others, 'Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study' (2017) 283(2) *Radiology* https://pubmed.ncbi.nlm.nih.gov/28092203/.

[11] D Sculley and others, 'Hidden Technical Debt in Machine Learning Systems' (2018) 2 *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems* https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

[12] C Rudin, 'New Algorithms for Interpretable Machine Learning' (2014) KDD'14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining www.bu.edu/hic/2018/12/04/air-rudin/.

[13] AC Clarke, 'Hazards of Prophecy: The Failure of Imagination' in AC Clarke, *Profiles of the Future: An Enquiry into the Limits of the Possible* (1962).

Why is unpaid intellectual debt worrisome? There are at least three reasons, in increasing difficulty. First, when we don't know how something works, it becomes hard to predict how well it will adjust to unusual situations. To be sure, if a system can be trained on a broad enough range of situations, nothing need ever be unusual to it. But malefactors can confront even these supposedly robust systems with specially-crafted inputs so rare that they'd never be encountered in the normal course of events. Those inputs – commonly referred to as 'adversarial examples' – can look normal to the human eye, while utterly confusing a trained AI.

For example, computers used to be very bad at recognizing what was in photos. That made categorization of billions of online images for a search engine like Google Images inaccurate. Fifteen years ago the brilliant computer scientist *Luis von Ahn* solved the problem by finding a way for people, instead of computers, to sort the photos for free. He did this by making the 'ESP game'.[14] People were offered an online game in which they were shown images and asked to guess what other people might say was in them. When they were right, they earned points. They couldn't cash the points in for anything, but thousands of people played the game anyway. And when they did, their successful guesses became the basis for labeling images. Google bought *Luis's* game, and the field of human computation – employing human minds as computing power – took off.[15]

Today, Google's 'Inception' architecture – a specially-configured 'neural network' machine learning system – has become so good at image recognition that *Luis's* game is no longer needed to get people to label photos. We know how Inception was built.[16] But even its builders don't know how it gets a given image right. Inception produces answers, but not the kinds of explanations that the players of *Luis*'s game could offer if they were asked. Inception correctly identifies, say, cats. But it can't provide an explanation for what distinguishes a picture of a cat from anything else. And in the absence of a theory of cathood, it turns out that Inception can be tricked by images that any human would still immediately see as one of a cat.

MIT undergraduates were able to digitally alter the pixels of a standard cat photo to leave it visibly unchanged – and yet fool Google's state-of-the-art image detection engine into determining with 'hundred percent confidence' that it was looking at a picture of guacamole.[17] They then went a step further and painted a 3D-printed turtle in a way that looks entirely turtle-like to a human – and causes Google to classify it at every angle as a rifle.[18]

A system that had a discernible theory of whiskers and ears for cats, or muzzles for rifles, would be harder to fool – or at least would only be foolable along the lines that humans could be. But systems without theory have any number of unknown gaps in their accuracy. This is not just a quirk of Google's state-of-the-art image recognizer. In the realm of healthcare, systems trained to classify skin lesions as benign or malignant can be similarly tricked into flipping their previously-accurate judgments with an arbitrary amount of misplaced confidence,[19] and the prospect of

[14] L Von Ahn and L Dabbish, 'Labeling Images with a Computer Game' (2004) CHI'04 Proceedings of the 2004 Conference on Human Factors in Computing Systems 319.

[15] See J Zitthrain, 'Ubiquitous Human Computing' (2008) Oxford Legal Studies Research Paper No. 32 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1140445.

[16] C Szegedy and others 'Rethinking the Inception Architecture for Computer Vision' (2016) 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2818.

[17] A Ilyas and others, 'Black-box Adversarial Attacks with Limited Queries and Information' (*labsix*, 23 April 2018) www.labsix.org/limited-information-adversarial-examples/.

[18] A Athalye and others, 'Fooling Neural Networks in the Physical World with 3D Adversarial Objects' (*labsix*, 31 October 2017) www.labsix.org/physical-objects-that-fool-neural-nets/.

[19] SG Finlayson and others, 'Adversarial Attacks against Medical Deep Learning Systems' (2019) arXiv:1804.05296v3 https://arxiv.org/pdf/1804.05296.pdf.

https://doi.org/10.1017/9781009207898.014 Published online by Cambridge University Press

triggering insurance reimbursements for such inaccurate findings could inspire the real world use of these techniques.[20]

The consistent accuracy of a machine learning system does not defend it against these kinds of attacks; rather, it may serve only to lull us into the chicken's sense that the kindly farmer comes every day with more feed – and will keep doing so. Charmed by its ready-to-hand predictive power, we will embed machine learning – like the asbestos of yesteryear – into larger systems, and forget about it. But it will remain susceptible to hijacking with no easy process for continuing to validate the answers it is producing, especially as we stand down the work of the human judges it will ultimately replace. Intellectual debt entails a trade-off for vulnerability that is easy to drift into just the way that technical debt does.

There is a second reason to worry as AI's intellectual debt piles up: the coming pervasiveness of machine learning models. Taken in isolation, oracular answers can generate consistently helpful results. But these systems won't stay in isolation. As AI systems gather and ingest the world's data, they'll produce data of their own – much of which will be taken up by still other AI systems. The New York Subway system has its own old-fashioned technical debt, as trains run through tunnels and switches whose original installers and maintainers have long moved on. How much more complicated would it be if that system's activities became synchronized with the train departures at Grand Central Terminal, and then new 'smart city' traffic lights throughout the five boroughs?

Even simple interactions can lead to trouble. In 2011, biologist *Michael Eisen* found from one of his students that an unremarkable used book – *The Making of a Fly: The Genetics of Animal Design* – was being offered for sale on Amazon by the lowest-priced seller for just over $1.7 million, plus $3.99 shipping.[21] The next cheapest copy weighed in at $2.1 million. The respective sellers were well established; each had thousands of positive reviews. When *Eisen* visited the page the next day, the prices had gone up yet further. As each day brought new increases from the sellers, *Eisen* performed some simple math: Seller A's price was consistently 99.83% that of Seller B. And Seller B's price was, each day, adjusted to be 127.059% that of Seller A.

*Eisen* figured that Seller A had a copy of the book and, true to the principles of Economics 101, was seeking to offer the lowest price of all sellers by slightly undercutting the next cheapest price. He then surmised that Seller B did not have a copy of the book, so priced it higher – and was then waiting to see if anyone bought the more expensive copy anyway. If so, Seller B could always get it from Seller A and direct delivery of the book to the lazy buyer, pocketing a handsome profit without having to actually personally package and mail anything.

Each seller's strategy is rational – and while algorithmic, surely involved no sophisticated machine learning at all. Even those straightforward strategies collided to produce manifestly irrational results. The interaction of thousands of machine learning systems in the wild promises to be much more unpredictable.

The financial markets provide an obvious breeding ground for this type of problem – and one in which cutting-edge machine learning is already being deployed today. In 2010, a 'flash crash' driven by algorithmic trading wiped more than $1 trillion from the major US indices – for thirty-six minutes. Last fall, JPMorgan analyst *Marko Kolanovic* shared a short analysis within a 168-page market report that suggested it could readily happen again, as more investing becomes

[20] SG Finlayson and others, 'Adversarial Attacks on Medical Machine Learning' 363 *Science* 1287.

[21] M Eisen, 'Amazon's $23,698,655.93 Book about Flies' it is NOT junk (22 April 2011) www.michaeleisen.org/blog/?p=358.

passive rather than active, and simply tied to indices.[22] Unlike technical debt, whose borrowing is typically attributable to a particular entity that is stewarding a system, intellectual debt can accumulate in the interstices where systems bump into each other without formally interconnecting.

A third, and most profound, issue with intellectual debt is the prospect that it represents a larger movement from basic science towards applied technology, one that threatens to either transform academia's investigative rigors or bypass them entirely.[23] Unlike, say, particle accelerators, the tools of machine learning are as readily taken up by private industry as by universities. Indeed, the kind and volume of data that will produce useful predictions is more likely to be in Google and Facebook's possession than at the MIT computer science department or Media Lab. Industry may be perfectly satisfied with answers that lack theory. But when those answers aren't themselves well publicized, much less the AI tools that produce them, intellectual debt will build in societal pockets far away from the academics who would be most interested in backfilling the theory. And an obsession only with answers – represented by a shift in public funding[24] of research to orient around them – can in turn steer even pure academics away from paying off the intellectual debt they might find in the world, and instead towards borrowing more.

One researcher in the abstruse but significant field of 'protein folding' recently wrote an essay exploring his ambivalence about what it means to be a scientist after a machine learning model was able to, well, fold proteins in ways that only humans had previously been able to achieve.[25] He told one publication: 'We've had this tendency as a field to be very obsessed with data collection. The papers that end up in the most prestigious venues tend to be the ones that collect very large data sets. There's far less prestige associated with conceptual papers or papers that provide some new analytical insight.'[26]

It would be the consummate pedant who refused to take a life-saving drug simply because no one knew how it worked. At any given moment an intellectual loan can genuinely be worth taking out. But as more and more drugs with unknown mechanisms of action proliferate – none of them found in the wild – the number of tests to uncover untoward interactions must scale exponentially. In practice, these interactions are simply found once new drugs find their way to the market and bad things start happening, which partially accounts for the continuing cycle of introduction-and-abandonment of drugs. The proliferation of machine learning models and their fruits makes that problem escape the boundaries of one field.

So, what should we do? First, we need to know our exposure. As machine learning and its detached answers rightfully blossom, we should invest in a collective intellectual debt balance sheet. Debt is not only often tolerable, but often valuable – it leverages what we can do. Just as a little technical debt in a software system can help adapt it to new uses without having to

[22] T Heath, 'The Warning from JPMorgan about Flash Crashes Ahead' *The Washington Post* (5 September 2018) www.washingtonpost.com/business/economy/the-warning-from-jpmorgan-about-flash-crashes-ahead/2018/09/05/25b1f90a-b148-11e8-a20b-5f4f84429666_story.html.

[23] K Birchard and J Lewington 'Dispute Over the Future of Basic Research in Canada' *The New York Times* (16 February 2014) www.nytimes.com/2014/02/17/world/americas/dispute-over-the-future-of-basic-research-in-canada.html.

[24] T Caulfield 'Should Scientists Have to Always Show the Commercial Benefits of Their Research?' (*Policy Options*, 1 December 2012) https://policyoptions.irpp.org/magazines/talking-science/caulfield/.

[25] M AlQuraishi 'AlphaFold @ CASP13: "What Just Happened?"' *Some Thoughts on a Mysterious* Universe, (9 December 2018) https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/#comment-26005.

[26] S Samuel, 'How One Scientist Coped When AI Beat Him at His Life's Work' (Vox, 15 February 2019) www.vox.com/future-perfect/2019/2/15/18226493/deepmind-alphafold-artificial-intelligence-protein-folding.

continually rebuild it, a measure of considered intellectual debt can give us a *Promethean* knowledge boost, and then signpost a research agenda to discover the theory that could follow.

For that, we need the signposts. We must keep track of just where we've plugged in the answers of an alien system, rather than tossing crumpled IOUs into a file cabinet that could come due without our noticing. Not all debt is created equal. When the stakes are low, such as the use of machine learning to produce new pizza recipes,[27] it may make sense to shut up and enjoy the pizza, never fretting about the theory behind what makes peanut butter and banana toppings work so well together on a pie. But when the stakes are higher, such as the use of AI to make health predictions and recommendations, we walk on untested ice when we crib the answers to a test rather than ever learning the underlying material. That it is near-irresistible to use the answers makes pursuing an accompanying theory all the more important.

To achieve a balance sheet for intellectual debt, we must look at current practices around trade secrets and other intellectual property. Just as our patent system requires public disclosure of a novel technique in exchange for protection against its copying by others, or the city building department requires the public availability of renovation plans for private buildings, we should explore academic mirroring and escrow of otherwise-hidden data sets and algorithms that achieve a certain measure of public use. That gives us a hope for building a map of debt – and a rapid way to set a research agenda to pay off debt that appears to have become particularly precarious.

Most important, we should not deceive ourselves into thinking that answers alone are all that matters: Indeed, without theory, they may not be meaningful answers at all. As associational and predictive engines spread and inhale ever more data, the risk of spurious correlations itself skyrockets. Consider one brilliant amateur's running list of very tight associations found,[28] not because of any genuine association, but because with enough data, meaningless, evanescent patterns will emerge. The list includes almost perfect correlations between the divorce rate in Maine and the per capita consumption of margarine, and between US spending on science, space, and technology and suicides by hanging, strangulation, and suffocation. At just the time when statisticians and scientists are moving to de-mechanize the use of statistical correlations,[29] acknowledging that the production of correlations alone has led us astray, machine learning is experiencing that success of the former asbestos industry relies on the basis of exactly those kinds of correlations.

Traditional debt shifts control, from borrower to lender, and from future to past, as later decisions are constrained by earlier bargains. Answers without theory – intellectual debt – also will shift control in subtle ways. Networked AI is moving decisions previously left by necessity to, say, a vehicle's driver into the hands of those tasked with designing autonomous vehicles – hence the ongoing hand-wringing around ethical trolley problems.[30] Society, not the driver, can now directly decide whom a car that has lost its brakes should most put at risk, including its passengers. And the past can now decide for the future: Cars can be programmed well ahead of time with decisions to be actualized later.

[27] 'Episode 2: Human-AI Collaborated Pizza' *How to Generate (Almost) Anything* (30 August 2018) https://howtogeneratealmostanything.com/food/2018/08/30/episode2.html.

[28] T Vigen, *Spurious Correlations* (2015).

[29] RL Wasserstein, AL Schirm, and NA Lazar, 'Moving to a World Beyond "p < 0.05"' (2019) 73(S1) *The American Statistician* www.tandfonline.com/doi/pdf/10.1080/00031305.2019.1583913?needAccess=true.

[30] G Marcus 'Moral Machines' *The New Yorker* (24 November 2012) www.newyorker.com/news/news-desk/moral-machines.

A world of knowledge without understanding becomes, to those of us living in it, a world without discernible cause and effect, and thus a world where we might become dependent on our own digital concierges to tell us what to do and when. It's a world where home insurance rates could rise or fall by the hour or the minute as new risks are accurately predicted for a given neighborhood or home. The only way to make sense of that world might be to employ our own AIs to try to best position us for success with renter's insurance AIs ('today's a good day to stay home'); hiring AIs ('consider wearing blue'); or admissions AIs ('volunteer at an animal shelter instead of a homeless shelter'), each taking and processing inputs in inscrutable ways.

When we have a theory, we get advanced warning of trouble when the theory stops working well. We are called to come up with a new theory. Without the theory, we lose the autonomy that comes from knowing what we don't know.

Philosopher *David Weinberger* has raised the fascinating prospect that machine learning could help us tap into natural phenomena that themselves don't avail themselves of any theory to begin with.[31] It's possible that there are complex but – with enough computing power – predictable relationships in the universe that simply cannot be boiled down to an elegant formula like Newton's account of gravity taught in high schools around the world, or Einstein's famed insight about matter, energy, and the speed of light. But we are soon to beat nature to that complex punch: with AI, in the name of progress, we will build phenomena that can only be predicted, while never understood, by other AI.

That is, we will build models dependent on, and in turn creating, underlying logic so far beyond our grasp that they defy meaningful discussion and intervention. In a particularly fitting twist, the surgical procedure of electrical deep brain stimulation has advanced through trial-and-error – and is now considered for the augmentation of human thinking, 'cosmetic neurosurgery'.[32]

Much of the timely criticism of AI has rightly focused on ways in which it can go wrong: it can create or replicate bias; it can make mistakes; it can be put to evil ends. Alongside those worries belongs another one: what happens when AI gets it right, becoming an Oracle to which we cannot help but to return and to whom we therefore become bonded.

[31] D Weinberger, 'Optimization over Explanation' (*Berkman Klein Center*, 28 January 2018) https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d.

[32] N Lipsman and AM Lozano, 'Cosmetic Neurosurgery, Ethics, and Enhancement' (2015) 2 *The Lancet Psychiatry* 585.