Irresolvable Contradictions in Algorithmic Thought

Leonardo Impett

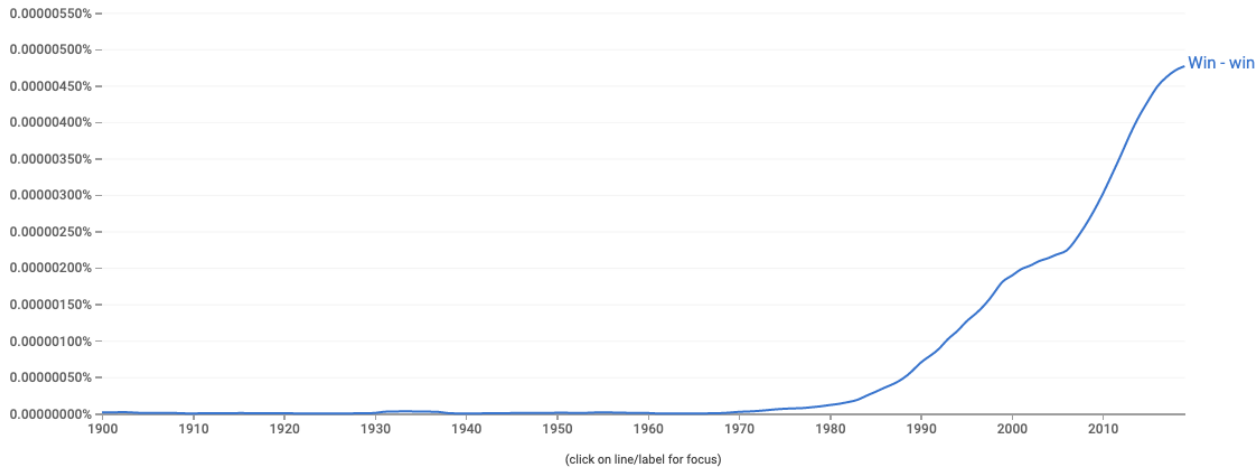

Figure 1. The relative frequency of the term 'win-win' in the Google Books English-language corpus, from the Google Books Ngrams Viewer

'A man cannot have his cake and eat his cake'

– Thomas Howard, the 3rd Duke of Norfolk, to Thomas Cromwell, 14 March 1538 [1]

The 'win-win' situation is perhaps the most characteristic rhetorical device of late capitalism. You'll find it frequently in the white papers of Neoliberal think-tanks and company annual reports, but a recent form of particular prevalence combines moral absolution with economic growth. 'Electric cars emit no carbon, and they're a lot faster from 0–60.' 'Offshoring doesn't just save us money, it invests in developing economies.' 'We can tackle bias and unfairness in AI by making our workforce more diverse, which will also make us more profitable.'[2]

These happy victories, even if genuine, are often pointers to insufficiency. After all, capitalism's ability to radically self-invent in the interest of its own profits has never been in doubt. What is in doubt is its ability to solve questions of public good *against* its own commercial interests. Take the precedents of tobacco and public health, or oil and climate change: wherever common interests pose an existential threat to an industry's survival, that industry has historically responded through the tactics of confusion and delay.

So, when Microsoft lists 'inclusiveness' and 'fairness' as two of the six guiding principles for Responsible AI – indeed, when Microsoft even considers the idea of having a central policy on Responsible AI – we may well feel the same kind of scepticism provoked by, say, Philip Morris' 2018 anti-smoking advertising.

To feel this, however, would be to misunderstand the situation on the ground. Corporate AI labs have become serious nodes for critical thought around the role of AI in society. This year, a paper on fairness in AI from Microsoft Research won the Best Paper award at the prestigious CHI Conference on Human Factors in Computing Systems [3] and a large amount of internationally relevant research is produced by other major corporate tech players. My own first encounter with critical approaches to AI was as part of a paid internship at Microsoft Research's Cairo office, where – as an Engineering undergraduate – my main project was a Bourdieuian reading of a commonly used computer vision dataset. Last year, Amazon announced a $20m funding track with the US National Science Foundation on *Fairness in Artificial Intelligence*.

Of course – despite their well-publicised efforts towards fairer AI – Amazon, Google, Microsoft, Oracle and IBM all bid for a $10 billion cloud computing contract with the US Department of Defense in 2018–19. In October 2018, Google dropped out of the bid, apparently because it 'couldn't be assured that it would align with our AI Principles'. A progressive capitalism working against its own profits? Not quite –

Google's decision came after seven months of traditional organised worker pressure, including demonstrations and high-profile resignations. For Google, Microsoft and Amazon, the political situation is one of internal dissonance rather than calculated hypocrisy: in Hegelian-Marxist terms, we might call it a collateral contradiction (Nebenwiderspruch). This contradiction on the corporate level interplays with an underlying contradiction at the algorithmic level, particularly in relation to deep-learning techniques.

The technical borders of 'deep learning' are not perfectly delineated, but we understand it to refer to a subset of machine-learning algorithms called neural networks. These neural networks are generally divided into separate layers (where the output of one layer is itself the input to the next, and so on). We call a neural network 'deep' [4] when it has a relatively large number of such layers. Before deep learning, machine-learning techniques generally had to rely on hand-crafted features of the data. Rather than being fed with individual pixel values, early machine-learning algorithms for images were fed a spreadsheet of secondary information about an image: brightness, average colours, gradients and silhouettes. The 'depth' (i.e. large number of neural layers) of deep neural networks allows them, instead, to work on unadulterated raw data: on pixel values themselves.[5]

To illustrate the dangerous complications that deep networks introduce, let's consider a case-study: 'smart' CCTV systems that alert security when an 'anomalous' event occurs. The industry for home 'smart' surveillance technology, such as Amazon's CCTV-enabled Ring, is swiftly reaching the $10bn mark [6] and 'anomaly detection' has become an important application domain for research in deep learning.[7]

Let's assume that we train our deep CCTV algorithm on a dataset of real security camera footage, taken at random from across the United States (where most AI companies are based). We mark any event that eventually led to an observed person's incarceration an 'anomaly', an example from which to learn. If our dataset is unbiased in the orthodox statistical sense (i.e. if it contains a representative sample of US arrests) we would find our smart CCTV dataset contained black people involved in such 'anomalous' events at five times the frequency of the white population.[8] If we then trained a perfect deep-learning model – ie one that was able to perfectly reproduce the decision-making processes implicit in its training data – we would have built a CCTV system that flags black people as incidents to be investigated five times more frequently than white people. Even worse, if our classifier was not perfectly accurate (we all – human or machine – have some probability of error), the 'false positive' rate for black people would – through the associative logic of Bayesian reasoning – also be at around 500% of the rate for the white population. In fact, this would be roughly consistent with the rate at which black people are disproportionately stopped without just cause.[9]

This active prejudice would be present in a deep CCTV system that was never explicitly programmed to take account of race as a variable. In the world of deep learning, anything implicit in pixel-values is fair game: prejudice is present in the deep-learning system, because it was present in the data. Many point, therefore, to the possibility of creating less biased datasets – utopic data. This may be possible for the dualistic white-black distinction drawn in the thought-experiment above, but how could we possibly create perfectly fair training data that balances the complex network of intersecting prejudices (sexuality, gender, social class, income, nationality, disability) hinted at in those raw pixel-values? The problem is even greater for so-called 'online machine learning' (often involving 'reinforcement learning'): systems that continuously self-improve based on examples they face in the real world. In these systems, which make up a significant portion of commercially applied AI (e.g. search engines), we cannot introduce utopic data, since live information about how the system interacts with the (dystopic) real world is integral to how it works and learns.

As Louise Amoore writes, 'the features that some would like to excise from the algorithm – bias, assumptions, weights – are routes into opening up their politics'. Indeed, recent research into algorithmic bias allows us to think through this quagmire with considerably more precision and nuance, with implications far beyond computer science. In 2016, Jon Kleinberg, Sendhil Mullainathan and Manish

Raghavan showed [10] that in any real-world society, [11] it is exceedingly difficult to make a 'fair' algorithm. More specifically, it is *impossible* to classify data in a way that meets several well-established definitions of fairness at the same time (lack of active discrimination; equal false positive rates; equal false negative rates). What's more, the result of a 'fair' classification (through whichever definition of fairness) is not the most accurate classification. There exists, in other words, a structural contradiction between the classificatorial logics of non-discrimination and of accuracy (and therefore profit).[12]

We have been talking about algorithms, but what makes the Kleinberg proof so powerful is that it is based only on the eventual decisions taken: it holds true, no matter who makes the classification-decisions (deep-learning algorithm or human, reactionary or progressive). The collateral contradictions (Nebenwidersprüche) in the corporate behaviour of Big Tech, or in the discriminatory logics of deep learning, have their heart in the fundamental contradiction (Hauptwiderspruch) that Kleinberg's proof describes. It forces us to confront head-on the axiom (dating back at least to Weber) underlying the 'win-win' solution: that a fair solution is also an efficient solution. Efficiency relies on prejudice (and enhances it), and if we are to take anti-discrimination seriously, we must sacrifice accuracy, efficiency, profit, growth. We can no longer have our cake and eat it.

---

[1] Letters and Papers, Foreign and Domestic, Henry VIII, Volume 13, Part 1, January–July 1538. Originally published by Her Majesty's Stationery Office (London, 1892), pp. 176–92. This research was partly funded by UKRI's Arts Humanities Research Council program 'Towards a National Collection' under grant AH/V015478/1.

[2] See, for example, Karsten Strauss, 'More Evidence That Company Diversity Leads To Better Profits', *Forbes*, 25 January 2018; Stephen Turban, Dan Wu and Letian Zhang, 'Research: When Gender Diversity Makes Firms More Productive', *Harvard Business Review* (February 11, 2019); Jessica Alsford, 'The data show it: diverse companies do better', *Financial Times (*September 30, 2019).

[3] Michael A. Madaio, et al., 'Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI.' Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

[4] The terminology originated in convolutional neural networks for image recognition, in which the 'height' and 'width' of the neural fields were determined by the resolution of the images; and the 'depth' was therefore the number of neural layers in the model.

[5] Of course, such data is rarely 'raw' – in the case of mobile-phone images, it might have passed through automatic smartphone image enhancement techniques (to increase the contrast or colour saturation), not to mention the mediation of a human agent choosing to point a camera in a particular direction at a particular moment. But it is 'raw' at a technical level – all the information that is implicit in the pixel-values is now up for grabs.

[6] T. J. McCue, 'Home Security Cameras Market To Surpass $9.7 Billion By 2023'*, Forbes* (January 31, 2019).

[7] See, for example, Kwang-Eun Ko and Kwee-Bo Sim, 'Deep convolutional framework for abnormal behavior detection in a smart surveillance system', *Engineering Applications of Artificial Intelligence,* 67 (2018), pp. 226–34.

[8] NAACP Criminal Justice Fact Sheet, 2020, https://www.naacp.org/criminal-justice-fact-sheet/

[9] NAACP Criminal Justice Fact Sheet, 2020, https://www.naacp.org/criminal-justice-fact-sheet/

[10] Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores', 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2017).

[11] Specifically, any society in which different groups (whatever distinguishes them: gender, race, sexuality etc) do not have identical patterns of behaviour.

[12] Only two edge-cases provide statistical coincidences where the accurate and the fair coincide: 'perfect predictions' (i.e. where our algorithm makes no errors), and 'equal base rates' (i.e. where

different groups have statistically identical behaviours).

Leonardo Impett

Leonardo Impett is Assistant Professor of Computer Science at Durham University. He works in the digital humanities, at the intersection of computer vision and art history. He was previously Scientist at the Bibliotheca Hertziana – Max Planck Institute for Art History, Digital Humanities Fellow at Villa I Tatti – the Harvard University Center for Italian Renaissance Studies, and PhD Candidate at the École Polytechnique Fédérale de Lausanne. In trying to bring 'Distant Reading' to art history and visual studies, his current research focuses on unveiling the implicit image-theories of computer vision, and constructing new computer-vision systems based on early modern philosophies of vision. He is an Associate of Cambridge University Digital Humanities, an Associate Fellow of the Zurich Center for Digital Visual Studies, and an Associate Research at the Orpheus Institute for Artistic Research in Music.