spheres
JOURNAL FOR DIGITAL CULTURES

www.spheres-journal.org
ISSN 2363-8621
#5 *Spectres of AI*

PABLO R. VELASCO

ARTIFICIAL INTELLIGIBILITY AND PROXY ERROR – A COMMENT ON HOW A MACHINE LEARNS AND FAILS

A representative moment of Artificial Intelligence (AI) capturing the social imaginary took place in March 2016, when Google's AlphaGo computer program beat professional Go player Lee Sedol. Ten years before, IBM's Deep Blue computer defeated chess grandmaster Garry Kasparov. It is worth to revisit two major insights that a decade of 'intelligent' machines left. First, an image search of both terms – "ibm deepblue" and "google alphago" – would reveal photos of both Kasparov's and Sedol's struggling matches, but is also significantly telling that the Deep Blue query will also return a squared black box. After all, Deep Blue was primarily an advanced piece of hardware, while AlphaGo takes the stage as software. Both consist, of course, of a coupling of logical instructions and computing power, but the machinery in the case of AlphaGo is shown as less relevant, less present.[1] Second, Deep Blue's advanced hardware was needed to run its Minimax algorithm, a non-probabilistic method for minimizing loosing scenarios: for each move made, it examines possible reactions from the opponent in future turns, as far away as the computing power allows.[2] As complex as this algorithm is, it is relatively easy to understand. As for AlphaGo, this particular issue differs noticeably. An article calling for the demystification of AI in the Scientific American journal[3] quotes Alan Winfeld, professor of robot ethics at the University of West England, on the neural networks that make AI systems like AlphaGo

1 Today, any chess software running in a smartphone is likely more powerful than Deep Blue's hardware and software combination.

2 An important difference with more modern AI approaches is that this method looks for possible scenarios, but without allocating probability.

3 Ariel Bleicher, "Demystifying the Black Box That Is AI", *Scientific American*, August 9, 2017. Available at: https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/ [accessed October 30, 2019].

work: "It's very difficult to find out why [a neural net] made a particular decision […] We still can't explain it". There is no need for an apocalyptic position regarding an inescapable black box here: it is possible to review every parameter in the neural network behind AlphaGo's resolution; in this sense it is a box that can be opened. However, the article continues, the 'meaning' of the decision is not exactly intelligible, as it is encoded in billions of connections.

There is no straightforward way to understand the dichotomy presented by the new black box, at the same time subject to scrutiny and ambiguous. That is partly why popular representations of AI, be they either integrated or apocalyptic, are surrounded by a mist of inscrutability. Among other recent critical approaches to AI,[4] Pasquinelli's work develops a much-needed pathway to demystify AI or, perhaps, locate it in its respective mythology by signalling its intrinsic logical limits. Not only are these approaches relevant to identify the biases that come as part of the statistical procedures behind AI 'learning', but also to locate the discussion within social assemblages and political structures (thus, the *corporate* AI). A grammar of the techniques (classification and regression), their aggregated elements (training data, learning algorithm, and model application), along with their own biases, responds to the question of what it exactly means to fail among this paradigm of rationality.

Each aggregated level comes with and replicates some sort of bias: social inequalities, compiled data, and algorithmic approximation techniques, all pre-empt future classifications. This is not unique to machine learning, historically, many classifications have left out entire populations due to 'unfitness', disdain, mishap, etc. This is the case for the early IQ tests, who were designed with a white population in mind, for example.[5] However, it is crucial to identify which particular biases are involved in new classification techniques, such as machine learning, and detach from an idea of neutral AI development. A growing set of literature and institutions are paying attention to inherent biases of AI. An extended report that followed the AI Now Institute 2017

[4] See Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, London, Penguin, 2017; Kate Crawford, "Anatomy of an AI System", *Anatomy of an AI System*, 2018. Available at: http://www.anatomyof.ai [accessed October 30, 2019]; Adrian Mackenzie, *Machine Learners: Archaeology of a Data Practice*, Cambridge, MA/London, The MIT Press, 2017; Anja Bechmann and Geoffrey C. Bowker, "Unsupervised by Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media", *Big Data & Society*, 6 (1), 2019, pp. 1–11. Available at: https://doi.org/10.1177/2053951718819569 [accessed October 30, 2019]; Alex Campolo et al., "AI Now 2017 Report", *AI Now Institute at New York University*, 2017.

[5] Craig L. Frisby and Betty Henry, "Science, Politics, and Best Practice: 35 Years After Larry P.", *Contemporary School Psychology*, 20 (1), 2016, pp. 46–62. Available at: https://doi.org/10.1007/s40688-015-0069-3 [accessed October 30, 2019].

symposium offers extensive recommendations for policy makers, companies and universities[6] on a diversity of areas: labour and automation, bias and inclusion, rights and liberties, and ethics and government. Some of these recommendations have been incorporated in European policy[7] to minimize biased uses of AI. Recommendations for ethical reflection and the exploration for social improvement when designing systems,[8] and archaeological insights of the practices associated with machine learning development[9] gradually debunk a mystified idea of AI as an unapproachable, neutral, one-size-fits-all technology. Increasingly, AI's failure as bias is being addressed by designers, researchers, policy makers, and other stakeholders.

However, there is another understanding of error that, very much like the black-boxed nature of the decision-making algorithms, eludes clarity and signals towards a different paradigm of rationality to be considered. Unlike the modern idea of error as either a mistake, path to knowledge, mode of discovery,[10] or even as errant mode of being,[11] the error within the black box of algorithms such as AlphaGo can only be represented in terms of approximation. Unlike the modern paradigm of rationality, where "the idea of truth is defined by error", the multiplicity of connections in many machine learning algorithms is framed as a statistical approximation, where the idea of error is identified only as "missing the mark" within a model. Error becomes anything that does not come near as what is expected, and is engulfed as part of a procedure: another parameter to be adjusted within the model. It is indeed a paradigm of rationality that detaches itself from a methodology of errors, except perhaps not due to lack of awareness but because it does not follow the modern idea of error (the enlightened error that is outside of the truth). The predominance of understanding truth and error in terms of approximation – besides carrying the pre-emption dangers of an illusion of causation accurately identified by

---

6 Campolo et al., "AI Now 2017 Report".

7 Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'", *AI Magazine*, 38 (3), 2017, pp. 50–57, here: p. 50. Available at: https://doi.org/10.1609/aimag.v38i3.2741 [accessed October 30, 2019].

8 Josh Cowls et al., "Designing AI for Social Good: Seven Essential Factors", 2019. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3388669 [accessed October 30, 2019].

9 Mackenzie, *Machine Learners*.

10 David Bates, "The Epistemology of Error in Late Enlightenment France", *Eighteenth-Century Studies*, 29 (3), 1996, pp. 307–327. Available at: https://doi.org/10.1353/ecs.1996.0016 [accessed October 30, 2019].

11 Martin Heidegger, *The Essence of Truth: On Plato's Cave Allegory and Theaetetus*, translated by Ted Sadler, London, Bloomsbury Academic, 2013; Stephen Watson and Christopher Fynsk, "On the Errancy of Dasein", *Diacritics*, 19 (3/4), 1989, pp. 49–61, here: p. 49. Available at: https://doi.org/10.2307/465388 [accessed October 30, 2019].

Pasquinelli – implies that, unlike previous paradigms of knowledge, AI does not only fail to integrate a methodology of error, but to produce an ontology that incorporates truth and error in a meaningful way.

The use of, for example, Deep Neural Networks (DNN) includes the outsourcing of the process of pattern recognition to the technique. Not only in the sense that techniques do the computational labour, but also regarding the 'visibility' of the decision-making process that enacts successful techniques. In other words, the techniques are not only used as tools (e.g. as sorting algorithms), but left to produce their own 'cognitive' processes to come with a desired output. In some cases, the reasoning behind such processes is overall or partially understood, but in some (such as DNN), the scientific actors do not entirely know how to make the rationale of the process intelligible. Widely used AI techniques add by design a new layer of uncertainty that appends to new scientific discoveries, that is, such techniques *fail* by design to describe what a system 'believes'.

This becomes more evident in new endeavours working on Understandable or Explicable AI, an emergent field looking for ways to deal with its black-boxed logic. Lipton provides an important distinction between two notions of interpretation of machine learning algorithms: one is concerned with post-hoc interpretations, trying to make sense of predictions without elucidating how the models work, while the other attempts to directly interpret the models.[12] This second categorization of interpretability may look at the algorithm, the parameters, how the solution to a problem is sought, and the general complexity of the model (e.g. if it can be thoroughly examined by humans). Fully understanding the model is what critically separates our relation with Deep Blue and AlphaGo: while the decision rules of the former are comprehensible, the generative algorithms that produce those decision rules in the later are not.

Explicable AI, thus, does not provide a one-to-one explanation. On the contrary, explanations are closer to the *art* of translation – indeed, Pasquinelli has referred elsewhere to the linear logic of neural networks as a "combinatory art"[13]. Take Nvidia, for example, a company known for the production of processing cards but also for being actively involved in the development of AI, who recently offered a visual representation of their UAV AI cognitive system. Interestingly, the way

---

[12] Zachary C. Lipton, "The Mythos of Model Interpretability," *ArXiv:1606.03490 [Cs, Stat]*, June 10, 2016. Available at: http://arxiv.org/abs/1606.03490 [accessed October 30, 2019].

[13] Matteo Pasquinelli, "Basic Structure of a Neural Network", 2017. Available at: https://www.academia.edu/33205589/Basic_structure_of_a_neural_network [accessed October 30, 2019].

the AI scientist at Nvidia provided a description is by re-interpreting what the algorithm 'sees' as important. This consisted in developing a second method to determine in a visual and human-readable form what the network 'thinks':

> "Once PilotNet was up and running, we wanted to know more about how it makes decisions. So we developed a method for determining what the network thinks is important when it looks at an image […] To understand what PilotNet cares about most when it gets new information from a car camera, we created a visualization map"[14].

Intelligibility in this scenario requires a second layer: it demands to build a system to understand what was previously understood by the network. Many approaches to make understandable the decision-making processes of neural networks use a proxy model. Observation is mediated by a second system, complicating the scientific apparatus. The reasoning is adapted from one mode of processing to a second mode of accessing the process. The system is read in different ways, not only by different accounts but also by different reasoning systems interacting with each other.[15] Interpretability of the system's logic, and interpretability of its errors, is proxy-based within the current AI paradigm.

To think of truth and error in terms of approximation and surrogates, does affect cultural and social arrangements. It is not only, as rightly argued by Pasquinelli, that the developers' ideologies do not acknowledge the social impact of their schemes, but that the idea of error is also transmuted within an ideology of improvement. What is wrong becomes what can be optimised, and failure is subsumed to an idea of progress. This, too, is normalised as a culture of constant improvement: optimising work, the body, the mind, free time. The 'limits' of AI show a pervasive feature of contemporary cultures and normalise an idea of progress attuned to a capital mode or production. Power as 'the normalisation of code' is expressed through the procedural technique of AI via the reproduction of biases, and the superimposition of classification to causation, as appropriately suggested by Pasquinelli, but also as a paradigm of rationality

---

[14] Danny Shapiro, "How NVIDIA's Neural Net Makes Decisions", *The Official NVIDIA Blog*, April 27, 2017. Available at: https://blogs.nvidia.com/blog/2017/04/27/how-nvidias-neural-net-makes-decisions/ [accessed October 30, 2019].

[15] See, for example, the Local Interpretable Model-agnostic Explanations (LIME) model, which approximates neural network models to "local interpretable" models, such as visual artifacts. Cp. Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier", *ArXiv:1602.04938 [Cs, Stat]*, February 16, 2016. Available at: http://arxiv.org/abs/1602.04938 [accessed October 30, 2019].

underpinned by approximation instead of meaning. An artificial intelligibility of neither a clear identification of processes and their anomalies acts as a proxy, a representation of error tuned to statistical models of truth and error that downplay the need for complex and more nurtured paradigms of rationality.