spheres
JOURNAL FOR DIGITAL CULTURES

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

MATTEO PASQUINELLI

HOW A MACHINE LEARNS AND FAILS – A GRAMMAR OF ERROR FOR ARTIFICIAL INTELLIGENCE

"Once the characteristic numbers are established for most concepts, mankind will then possess a new instrument which will enhance the capabilities of the mind to a far greater extent than optical instruments strengthen the eyes, and will supersede the microscope and telescope to the same extent that reason is superior to eyesight."[1] — Gottfried Wilhelm Leibniz.

"The Enlightenment was […] not about consensus, it was not about systematic unity, and it was not about the deployment of instrumental reason: what was developed in the Enlightenment was a modern idea of truth defined by error, a modern idea of knowledge defined by failure, conflict, and risk, but also hope."[2] — David Bates.

"There is no intelligence in Artificial Intelligence, nor does it really learn, even though it's technical name is machine learning, it is simply mathematical minimisation."[3] —Dan McQuillan.

"When you're fundraising, it's Artificial Intelligence. When you're hiring, it's Machine Learning. When you're implementing, it's logistic regression."[4] — Joe Davidson.

---

1 Gottfried Wilhelm Leibniz, "Preface to the General Science", 1677.
2 David W. Bates, *Enlightenment Aberrations: Error and Revolution in France*, Ithaca, NY, Cornell University Press, 2002, p. 18.
3 Dan McQuillan, "Manifesto on Algorithmic Humanitarianism", presented at the symposium Reimagining Digital Humanitarianism, Goldsmiths, University of London, February 16, 2018.
4 Joe Davison, "No, Machine Learning is not just glorified Statistics", *Medium*, June 27, 2018. Available at: towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3 [accessed March 21, 2019].

What does it mean for intelligence and, in particular, for Artificial Intelligence to fail, to make a mistake, to break a rule? Reflecting upon a previous epoch of modern rationality, the epistemologist David Bates has argued that the novelty of the Enlightenment, as a quest for knowledge, was a new methodology of error rather than dogmatic instrumental reason.[5] In contrast, the project of AI (that is to say, almost always, corporate AI), regardless and maybe due to its dreams of superhuman cognition, falls short in recognising and discussing the limits, approximations, biases, errors, fallacies, and vulnerabilities that are native to its paradigm. A paradigm of rationality that fails at providing a methodology of error is bound to end up, presumably, to become a caricature for puppetry fairs, as it is the case with the flaunted idea of AGI (Artificial General Intelligence).[6]

Machine learning is technically based on formulas for error correction, but the nature, scale and implications of error is rarely discussed in the developer community. Machine learning developers possess and keep on expanding a vast armamentarium of error correction tricks; however, they are committed to a restless 'code optimisation' without acknowledging the social impact of their logical approximations. For the complexity of the mathematics involved, the public debate on AI is unable to consider the logical limitations in AI and remains polarised between *integrated* and *apocalyptic* positions, between technophilia and technophobia.[7] The integrated position follows in the footsteps of Ray Kurzweil on his happy voyage towards Singularity believing that mathematics will solve all problems, and that mass automation will develop free from disruptions for the social order. In the specular apocalyptic position, misunderstanding the black box effect in machine learning, authors such as Nick Bostrom among others warn of a forthcoming dark age of reason in which blinded machines run amok.[8] This last position shares regions with conspiracy theory sentiments for which AI systems cannot be studied, known, and controlled. Even the apocalyptic position remains at the level of speculation ('*what if AI…*') and fails to clarify machine learning's inner logic ('*what is AI?*').

---

5 "It was [in the Enlightenment], perhaps for the first time in modern thought, that error assumed a significant role not just in the definition of knowledge but in the very search for knowledge itself." David W. Bates, *Enlightenment Aberrations: Error and Revolution in France*, Ithaca, NY, Cornell University Press, 2002, p. ix.

6 A reference to the 2016 robot Sophia that was built by Hanson Robotics. Ben Goertzel, histrionic patron of the so-called Artificial General Intelligence paradigm, supervised the project.

7 Cp. Umberto Eco, *Apocalypse Postponed*, Bloomington, IN, Indiana University Press, 2000.

8 Cp. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies,* Oxford, UK, Oxford University Press, 2014.

Luckily, a critical view of AI is slowly emerging. Thanks to popular books such as Cathy O'Neil's *Weapons of Math Destruction*, among others, it is becoming clear that the problem of AI has nothing to do with intelligence *per se* but with the manner in which it is applied to the governance of society and labour via statistical models – ones that should be transparent and exposed to public scrutiny.[9] As Yarden Katz has remarked, AI is just a marketing operation used to rebrand what was known a decade ago as large-scale data analytics and data centre business.[10] Digging into the core elements of algorithmic biases, Kate Crawford has stressed the broad ethical implications of machine learning classification and taxonomies, reminding that "machine learning is the biggest experiment of classification in human history."[11] Kate Crawford and Vladan Joler's essay "Anatomy of an AI system" is another example of incisive investigation of the black box of AI, in which they deconstruct the Amazon Echo device by remapping each of its components onto the global ecology and economy. The times seem ripe for a radical critique of machine intelligence: Dan McQuillan, for example, advocates the rise of a counter-culture that positions itself against the opaque normative apparatus of machine learning.[12]

Generally speaking, one can study AI either as a technical construction or as a social construction. However, the discussion about AI's limits may be inaccurate if technical limits are divorced from social limits, and vice versa. Deleuze and Guattari's observations of the clock can be applied to AI in a useful way: the clock can be viewed as mechanical gear that projects universal time, or as an abstract discipline that controls collective time.[13] These two perspectives are of course imbricated and stimulate each other. It is, however, the social assemblage that tells the truth about the technical one and makes it historically possible and powerful. To paraphrase what Guattari once said of machines in general, machine intelligence is, eventually, constituted of "hyper-developed and hyper-concentrated forms of

---

9 Cp. Cathy O'Neil, *Weapons of Math Destruction*, New York, Broadway Books, 2016. See also: Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, NYU Press, 2018; and Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York, St. Martin's Press, 2018.

10 Cp. Yarden Katz, "Manufacturing an Artificial Intelligence Revolution", SSRN, November 2017. Available at: http://dx.doi.org/10.2139/ssrn.3078224 [accessed March 21, 2019].

11 Kate Crawford, "The Trouble with Bias", keynote lecture held at NIPS, 2017.

12 Cp. Daniel McQuillan, "People's Councils for Ethical Machine Learning", *Social Media and Society,* 4 (2), 2018. Available at: https://doi.org/10.1177%2F2056305118768303 [accessed March 21, 2019].

13 "The same machine can be both technical and social, but only when viewed from different perspectives: for example, the clock as a technical machine for measuring uniform time, and as a social machine for reproducing canonic hours and for assuring order in the city". Gilles Deleuze and Felix Guattari, *Anti-Oedipus,* Minneapolis, University of Minnesota Press, 1983, p. 141.

certain aspects of human subjectivity."[14]

Working at the intersection of the humanities and computer science, this text aims to sketch out a general grammar of machine learning, and to systematically provide an overview of its limits, approximations, biases, errors, fallacies, and vulnerabilities. The conventional term Artificial Intelligence is retained in this text to indicate the public reception and spectacularization of machine learning and the business of data analytics (Big Data). Technically speaking, it would be more accurate to call Artificial Intelligence machine learning or computational statistics but these terms would have zero marketing appeal for companies, universities and the art market. Given the degree of myth-making and social bias around its mathematical constructs, Artificial Intelligence has indeed inaugurated the age of *statistical science fiction.*

Introducing the Nooscope: A General Diagram of Machine Learning

The godfather of convolutional neural networks, Yann LeCun, argues that current AI systems are not sophisticated versions of cognition but of perception.[15] In the late 1950s, machine learning emerged as a form of visual **pattern recognition** that was then extended to the analysis of non-visual data. In the case of self-driving cars, the patterns to recognize are the most common visual features of a road scenario and, in the case of automatic translation, the patterns are the most common sequences of words across two languages. What machine learning calculates, however, is not an exact pattern but the **statistical distribution of a pattern**. Just scraping the surface of AI marketing, one finds a complex statistical construct to examine. How are these statistical models constructed? How accurate and reliable are they? What is the relation between statistical models and human intelligence? In fact, it would do good to reformulate the naive question 'Can a machine think?' into the theoretically sounder question 'Can a statistical model think?'

Artificial Intelligence is not 'intelligent' at all. It would be more precise to frame AI as an instrument of knowledge or logical magnification that *perceives* patterns that are beyond the reach of the human mind. Leibniz, addressing this modality of AI, makes use of the

---

14 Felix Guattari, *Schizoanalytic Cartographies*, London, Continuum, 2013, p. 2.

15 Yann LeCun, "Learning World Models: the Next Step towards AI", keynote lecture, International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, July 2018.

telescope and microscope as metaphors for his *calculus ratiocinator*.[16] In a similar manner, a machine learning system can be compared to a **nooscope**, a device that maps and perceives complex patterns through vast spaces of data (what the digital humanities term **distant reading**).[17] Nevertheless, each instrument of measurement and perception comes with inbuilt, contingent aberrations. In the same way that the lenses of microscopes and telescopes are never perfectly curvilinear and smooth, the *logical lenses* of AI systems have their own faults and aberrations. To study the impact of AI is to study the degree by which information flows are diffracted, distorted and lost by AI. To understand the nature of such information loss, one must study the algorithmic anatomy of the statistical models that underlie machine learning.

In mathematical terms, machine learning is used to predict an output value y given an input value x. Algorithms draw a function that relates x to y by learning from past data in which both x and y are known: $y = f(x)$. Constructing such function, the algorithm will be able to predict y based on future configurations of x. For example, given pictures of animals (x), the algorithm learns their association with the categories 'cat' or 'dog' (y) and then tries to classify new pictures accordingly. In this case, the input number x is a digital image, and the output number y is a percentage related to a semantic label (97 % 'cat', 3 % 'dog'). This is a process of *classification* that is distinguished from *regression* in which the output is a continuous number. An example of the latter would be an algorithm that learns to predict the credit score, output y, for any age of a group of students, input x. Classification and regression are both instances of supervised learning, whereby the algorithm takes data in which the relation between input x and output y is known and tries to guess the output y for future unknown inputs x. It is said that a machine learning algorithm *approximates* the function that maps y to x.

A machine learning system appears to a user or operator as composed of three elements or stages: training data, learning algorithm, and model application.

1. **Training data**: The training dataset contains data to be analysed in order to extract knowledge and 'intelligence', that is, patterns of association among its elements. In supervised learning, the training dataset is composed of two elements: the input x (e.g. raw images, student ages) and the output y (labels that describe those images, credit scores). In unsupervised or self-supervised

[16] See the opening quote by Leibniz.

[17] Cp. Franco Moretti, *Distant Reading*, London, Verso Books, 2013. See also Joseph Vogl, "Becoming Media: Galileo's Telescope", *Grey Room*, 29, 2007, pp. 14–25.

learning, only the input *x* is given from which an unknown pattern *y* must be discovered.

2. **Learning algorithm**: The learning algorithm extracts patterns from the training data by reading the association between the input *x* and the output *y* and constructing a statistical description of this association. The statistical model is the core of the machine learning, repository of the 'intelligence' extracted from the training data. However, it is never 100 % accurate and there is no scientific method to evaluate it: the training process stops when a human operator decides that *an acceptable error rate is reached for a test dataset*.
3. **Model application**: When the statistical model is considered sufficiently trained and 'fits' the training data, it can be applied to different tasks, such as classification and prediction. In classification (or *recognition*), a new value *x* is associated with a label *y*, if *x* falls within the distribution of the statistical model. In prediction (or *generation*), a new value *x* is used to generate and predict its corresponding value *y* by using the same statistical model (pattern generation is, logically, the same as prediction).

The assemblage of these three elements (Data + Algorithm + Model) is proposed as a general diagram of machine learning. Continuing the metaphor of optical media like telescopes and microscopes, it can be said that the information flow that crosses such instrument of knowledge (here denominated *nooscope*) behaves like a light beam that is projected by the training data, diffracted by the algorithm and its statistical model and reflected back to the world with built-in distortion. The following passages describe each individual component focusing in particular on the nature of the statistical model that is found at the core of machine learning.

Training Data, or the Collective Source of Machine Intelligence

Mass digitisation, which started after WWII with the commercialisation of industrial mainframes and reached its peak in the 2000s with global datacentres, laid the groundwork for a regime of *intelligence extractivism*. Machine intelligence is trained on vast datasets that are accumulated in ways neither technically neutral or socially impartial. Neutral data do not exist, as they are dependent on individual labour, personal data and social behaviours that accrue over long periods of time, from extended networks and diverse cultural taxonomies.[18]

[18] Cp. Lisa Gitelman (ed.), *Raw Data is an Oxymoron*, Cambridge, MA, MIT Press, 2013.

Training data are probably the most important factor in the quality of the 'intelligence' that machine learning algorithms extract. The training dataset is usually comprised of input data and ideal output data: raw digital images, for instance, can be associated with labels (how humans commonly categorise those images with their meaning). As outlined in mathematical terms above, machine learning is the calculation of the relation between the initial images (input) and their labels (output) with the purpose to predict the labels (output) of similar future images (input). The carving out, formatting and editing of the training dataset is a laborious and delicate undertaking, which is probably more significant than the technical parameters that control the learning algorithm.[19] In the preparation of the training datasets four stages can be recognised:

1. **Production:** individual labour or phenomena that produces information.
2. **Capture:** the capturing of information by an instrument that turns it into data.
3. **Formatting:** the encoding of information into a specific data format.
4. **Labelling:** the application of categories of a given taxonomy to the dataset.

The most popular training datasets used for machine learning (NMIST, ImageNet, Labelled Faces in the Wild, etc.) originate in corporations, universities, and military agencies of the Global North (although taking a more careful look, one discovers a profound division of labour that innervates into the Global South). Training data can be provided by spontaneous online behaviours (via social media, news coverage, mobile phones geolocation, etc.) or by screen work that is crowdsourced (via Amazon Mechanical Turk, for instance). In both cases, invisible and underrecognized forms of labour are utilized. Personal data, in particular, are buried and disappear into privatised datasets unknowingly and without transparency.[20] This is why such datasets also trigger issues of data sovereignty, privacy and civil rights that political bodies and the law are slowly becoming aware of (see the GDPR data privacy regulation that was passed in May 2018 by the European Parliament).

---

[19] For instance, it took nine years of manual work to label the 14 million images of the training dataset ImageNet, which was sponsored by Google, Amazon, Princeton, and Stanford universities.

[20] See Adam Harvey's project Megapixel (megapixels.cc). Madhumita Murgia, "Who's using your face? The ugly truth about facial recognition", *Financial Times,* April 19, 2019.

THE MODALITIES OF MACHINE LEARNING: TRAINING, CLASSIFICATION AND PREDICTION

When training data are ready to be analysed, they are presented to the learning algorithm, which is chosen out of many options by a human operator considering specific parameters. For instance, convolutional neural networks require specification from a very complex topology and set of hyper-parameters (number of layers, neurons, type of connection, behaviour of each layer and neuron, etc.). Though neural networks initially emerged as a technique for pattern recognition, computer scientists today prefer the more abstract and accurate expression **input-output mapping** in order to avoid the dated comparison to biological systems and visual perception. Nonetheless, the construction of a relation between an input x and output y is still fundamentally the search for a pattern. A primary example of basic pattern recognition is Frank Rosenblatt's Perceptron, which, created in 1957, was the first operative neural network. Given a visual matrix of 20x20 photoreceptors, this machine could learn how to recognise a simple letter. Today, given a much more complex input, such as the video recording of a busy street, the neural network of a self-driving car is asked to control mechanical gears and make ethical decisions when dangerous situations arise, thereby calling for an extremely complex input-output mapping. Regardless of their complexity, from the numerical perspective of machine learning, notions such as image, movement, form, style and decision can be all described as statistical distributions of a pattern. From the point of view of the statistical model, three modalities of operation of machine learning are given: 1) training, 2) classification, and 3) prediction. In more intuitive terms, these can be defined as: pattern abstraction, pattern recognition, and pattern generation.

1. In the **training** modality (*pattern abstraction*), the algorithm 'learns' the association of an input x to an output y (its label, for instance). As already mentioned, the algorithm weaves a statistical distribution of the underlying patterns and extracts them from their background. The statistical model will be considered trained when an acceptable error rate on a test dataset is reached (to date, there is no scientific method to determine when a model is sufficiently trained, that is, when an AI appears to be 'intelligent').
2. In the **classification** modality (*pattern recognition*), new input data x are compared with the statistical model to determine whether they fall within its statistical distribution or not. If so, they are assigned the corresponding output label y. Today object

classifiers exist that can detect all the most common objects on a road scenario and apply labels such as a person, car, truck, bicycle or traffic light in a matter of milliseconds – of course, with a margin of error.

3. In the **prediction** modality (*pattern generation*), new input data *x* are used to predict their output value *y*. In this modality, one may say that the statistical model is run *backwards* to generate new patterns rather than recording them. The expression "art created by AI" actually means that a human operator applies the generative modality of neural networks after training them with a given dataset. For instance, after being trained by the MIDI dataset of a music composer, a neural network can generate a new melody that *resembles* the composer's style. The generative modality is useful as a sort of algorithmic "reality check", as it shows what the model has learnt, i.e. how the model "sees the world".

Three Types of Bias

The information feedback loop between AI and society, that is between machine learning and its training data, is not a virtuous one, but rather is corrupted by technical bias. Any training dataset – regardless of how accurate it may seem – is a statistical sampling and therefore a partial view of the world. Moreover, the degree of information compression of machine learning algorithms affects the original proportions of the training data, which in turn amplify bias. Bias is the most debated and known issue of machine learning for its direct social implications and it is a good way to start illustrating the logical limitations of its statistical models. In machine learning it would be necessary at least to distinguish between world, data and algorithm biases.

World bias is already apparent in society before technological intervention, yet datasets reinforce race, gender and class inequalities, further normalizing the already operable stereotypes. The naturalisation of bias by machine learning, that is, the integration of inequality into an algorithm as apparently "unbiased data", can, of course, all by itself be harmful.[21] In order to pinpoint the categories of bias, Kate Crawford has distinguished between a resource allocation harm (when an algorithm denies mortgages to a minority group, for instance) and a social representation harm (such as denigration, under-representation or unfair determination of race, gender and class).[22]

---

21 Cp. Eubanks, *Automating Inequality*.
22 Cp. Kate Crawford, "The Trouble with Bias".

Data bias, on the other hand, is introduced through the capturing, formatting and labelling of data from the training dataset. The act of capturing and formatting the data itself has the potential to affect the resolution and accuracy of the information, but the most delicate part of the process is data labelling. Universities, corporations and military agencies build training datasets with rough and cheap labour. They often make use of old and conservative **taxonomies** causing a distorted view of world cultures and diversities. These taxonomies often reflect social hierarchies and are an expression of normative power, as Foucault has already elucidated.[23] Today, cultural and scientific taxonomies are embedded in and formalised by machine learning: Their normative power is no longer institutional but computational.

Algorithmic bias (also known as "machine bias", "statistical bias" or "model bias") is the further amplification of world bias and data bias caused by computational errors, information compression and the approximation techniques of machine learning algorithms. Due to their ratios of information compression, machine learning algorithms *diffract* and *distort* world and data biases, causing inequalities to be even more unequal. One way to illustrate this diffraction and amplification is to consider the illusion of anamorphic perspective used in painting and graphic design. Machine learning's view of the world is also *anamorphic:* Even if it respects the shape, or topology, of the world, it distorts its proportions.

The Logical Limits of the Statistical Model

At the core of current AI systems lies a learning algorithm whose purpose is to compute a statistical model of the training data. Computer scientists simply call it "**the model**". The model is the statistical representation of a large and diverse training dataset into one file. Since the time of Rosenblatt's Perceptron, the first operative neural network, the key objective of machine learning has been to store one small statistical model rather than memorising, for example, a thousand pictures of the same object from different angles. The model is calculated using different techniques (e.g. neural networks, Support Vector Machines, Bayesian networks) that always take the form of **statistical inference** and whose output takes the form, accordingly, of a **statistical distribution**. Technically one says that the model learns the statistical distribution of the training data by mapping the correlations (also known as patterns or dependencies) between input and desired output. The statistical model ultimately constructs a

[23] Cp. Michel Foucault, *The Order of Things*, London, Routledge, 2005.

function *f* that, when effective, describes the training data fittingly and predicts the output of a future input.

Let's take a classic example of machine learning: LeNet, developed by Yann LeCun in 1988, is a convolutional neural network for the optical recognition of numbers in postal codes and bank cheques. The training data are provided by the MNIST database, that contains the 60,000 handwritten numbers (collected among two social groups only: US high school students and employees of the Census Bureau). The internal model of LeNet records the statistical association of given pictures of handwritten numbers with their correct label, that is a numeral in this case.[24] After being trained, the LeNet statistical model will recognise future occurrences of handwritten numbers with a margin of error.

A statistical model is said to be successfully trained when it can **generalise** the patterns of the training dataset to new data 'in the wild' by elegantly **fitting** the training data with the lowest margin of **error** possible (there is *always* a margin of error in machine learning). If a model learns the training data too well, it will be able to recognise only exact matches and will overlook patterns with close similarity. In this case, one says that the model is **overfitting**, as it is not able to distinguish patterns from background, that is, it has meticulously learnt everything *including* the noise. On the other hand, the model is **underfitting** when it is not able to formulate patterns from the training data. In overfitting, there is no information compression, whereas in underfitting the model has lost most of the valuable information.[25]

It is common to describe AI as the statistical measure of a correlation between data points. In fact, machine learning *learns* nothing in the proper sense of the word; it just maps an input *x* with an output *y* by drawing a function that *approximately* describes their tendency, and then applies that function to future inputs to predict their outputs. This function is also an approximation in the sense that it guesses the "missing parts" of the data graph: either through **interpolation**, which is the projection and prediction of an output *y* that falls within the known interval of input *x* in the training dataset, or through **extrapolation,** which is the projection and prediction of output *y* beyond the limits of *x*, often with high risks of inaccuracy.

---

[24] Cp. Yann LeCun et al., "Backpropagation applied to handwritten zip code recognition", *Neural Computation*, 1 (4), 1989, pp. 541–551.

[25] A third case maybe be given when a model learns a wrong pattern association. If apophenia is the human tendency to perceive meaningful patterns in random data, underfitting is a sort of machine apophenia. Machine apophenia happens if a statistical model sees a pattern that is not there, that is, if it reads noise as similar to an existing pattern.

Machine learning is incredibly efficient *qua* algorithm for analysing data and approximating a mathematical function that describes them. Computer scientists are, in fact, more at ease with the definition of AI as a technique of **information compression** than the popular conception of it as a manifestation of superhuman cognition.[26] Since ancient times, algorithms have been procedures of an economic nature, designed to achieve a result in the shortest number of steps consuming the least amount of resources, such space, time, energy, etc. The current arms race between AI companies is still about finding the fastest algorithms to compute statistical models. Information compression, therefore, measures the ratio of profit in these companies, but also the ratio of **information loss** – and that loss often means a loss of the world cultural diversity.

The analogy of optical media illuminates the features of AI better than the analogy of the human brain. Leaving aside the fact, for the time being, that the first operative neural network, the Perceptron, was a *vision machine*,[27] there are epistemic similarities between machine learning and optical media such as, using Leibniz's suggestions, the microscope and the telescope.[28] Machine learning, like these devices, presents problems in both **information resolution** and **information diffraction**, with statistical models performing a corrective role similar to that of lenses in optical media. In terms of **information obfuscation**, a well-known issue of machine learning is probably the **black box effect,** present in large neural networks (Deep Learning). "Black box" is a popular term used to describe how information compression erases a lot of apparently useless information, resulting in a condition of obfuscation that is irreversible.[29] This occurs as each layer of neurons discharge most of the data received from the previous one, in the process forgetting some links in the chain of "reasoning". Outside computer science, "black box" has become a generic metaphor to indicate the seeming complexity of AI systems as they can appear inscrutable and opaque, if not alien and out of control. Projects such as Explainable Artificial Intelligence, Interpretable Deep Learning and

---

[26] Computer scientists would argue that AI truly belongs to a subfield of *signal processing*, that is, *data compression*.

[27] Cp. Paul Virilio's book *The Vision Machine*, London/Bloomington, British Film Institute/Indiana University Press, 1994.

[28] As already mentioned, machine learning is a sort of *statistical cinema*, projecting the new genre of *statistical science fiction*.

[29] There are also issues of error propagation in which some characteristics of the GPU hardware can generate an error chain that reaches the higher layers of feature abstraction. See Li, Guanpeng et al., "Understanding error propagation in deep learning neural network (DNN) accelerators and applications", Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ACM, 2017.

Heatmapping among other have demonstrated, however, that it is possible to break into the "black box" and to make its obscure chain of computation interpretable to the users.[30]

Due to the degree of information compression and information loss that occurs in its statistical models, machine learning necessitates a reduction of the labels and categories that are initially present in the training datasets. In a technique called **dimensionality reduction**, for instance, categories that show *low variance* (i.e. whose values fluctuate only a little) are aggregated and eliminated to reduce calculation costs. Dimensionality reduction, then, leads to something that can be called **category reduction**, which is the shrinking of cultural taxonomies as well. Eventually, the effect of machine learning on world diversity is **normalisation**, that is, an equalisation of anomalies to an average norm. The technical term regression actually refers to the phenomenon of regression towards the mean that Francis Galton observed measuring the heights of people. Neural networks for facial recognition, for instance, show a tendency to favour images of people with light skin colour. The **regression towards the mean** is, then, not just a mathematical technique of machine learning but one with clear social consequences and political implications

Approximation Techniques and the Perils of Correlation

As Dan McQuillian aptly puts it: "There is no intelligence in Artificial Intelligence, nor does it really learn, even though it's technical name is machine learning, it is simply mathematical minimisation."[31] It is important to remember that the 'intelligence' of machine learning is not driven by the application of exact formulas of mathematical analysis but by algorithms for **approximation**, that is, by heuristic procedures. The shape of the correlation function between input x and output y is calculated algorithmically, step by step, through tiresome mechanical processes of gradual adjustment. This is the same procedure used in **differential geometry**, or calculus, in which small squared blocks are used to approximate an irregular area in place of drawing an exact curvilinear shape. Neural networks are said to be among the most efficient algorithms for learning because these differential methods of

[30] Nevertheless, the full interpretability and explicability of machine learning statistical models remains a myth, too. Cp. Zacharay C. Lipton, "The Mythos of Model Interpretability", arXiv preprint, January 10, 2016. Available at: https://arxiv.org/abs/1606.03490 [accessed March 10, 2019].

[31] Dan McQuillan, "Manifesto on Algorithmic Humanitarianism", presented at the symposium Reimagining Digital Humanitarianism, Goldsmiths, University of London, February 16, 2018.

approximation allow to *guess* any function given enough layers of neurons and computing time (as proven by the so-called Universal Approximation Theorem). When one says that "neural networks can solve any problem", it means that they can *approximate* the shape of any curve (any non-linear function) in a multi-dimensional space of data.[32] Brute-force gradual approximation of a function is the core feature of today's AI, and only from this perspective can one understand its potentialities and limitations.

Another problematic of machine learning is how **statistical correlation** between two elements is used to explain the **logical causation** from the one to the other. In the grammar of AI errors, this is not a mistake attributed to the machine but human fallacy. It is commonly understood that *correlation does not imply causation*, meaning that a statistical correlation alone is not sufficient to demonstrate causation. Such a logical fallacy easily becomes a political one. The illusion of causation can be used, for instance, to endorse predictive policing algorithms. When machine learning is applied to society in this way, predictive correlations are transformed into a political apparatus of **preemption**. Dan McQuillan remarks: "The predictive nature of machine learning promotes preemption, i.e. action that attempts to anticipate or prevent the predicted outcome."[33] Preemption, as the automation of decision-making, contributes to an exclusion of collective participation from social and political institutions. Machine learning may even support arbitrary and nonsensical correlations (e.g. between one's daily consumption of cheese, ethnicity and credit score a statistical correlation can always be found). This is what's called **algorithmic apophenia**, the illusory consolidation of correlations or causal relations that do not exist in the material world, but only in the mind of AI.[34]

The Unprediction of the New

Another logical limit found at the core of machine learning is the inability to predict and recognise a new **unique anomaly**, that is, an

---

[32] Nota bene: In these passages, the dividing line between input and output datapoints has been described as a curve. Actually, machine learning calculates such differential approximations in n-dimensional spaces by drawing, then, hyperplanes (rather than a curve on a two-dimensional matrix).

[33] Dan McQuillan, "People's Councils for Ethical Machine Learning", *Social Media and Society*, 4 (2), 2018, pp. 1–10, here: 3, emphasis added. Available at: https://doi.org/10.1177%2F2056305118768303 [accessed March 21, 2019].

[34] See also "Illusory Correlation", *Wikipedia,* last edited March 11, 2019. Available at: http://en.wikipedia.org/wiki/Illusory_correlation [accessed March 21, 2019]. On apophenia see also Matteo Pasquinelli, "Anomaly Detection: The Mathematization of the Abnormal in the Metadata Society", paper presented at transmediale, 2015. Available at: www.academia.edu/10369819 [accessed March 21, 2019].

anomaly that appears only once, such as a new metaphor in poetry, a new joke made in everyday language or a mysterious object in the middle of the road. AI systems with speech recognition algorithms run into trouble when confronted by local dialects, for example. Worse, social minorities often fall outside the radar of AI logistics and are excluded (for example, people speaking with a Scottish accent to Amazon Alexa or black communities bypassed by Amazon delivery).[35] The *undetection of the new* (something that is unexpected, i.e. has never before "been seen" by a machine and therefore not classified in a known category) is a particularly hazardous problem for self-driving cars, which have already caused fatalities because of this. **Adversarial attacks** exploit such blind spots in machine learning, using uncanny patterns that obstruct the machine's visual reading of the environment: these patterns are sometimes designed by a human mind knowing that an AI 'mind' has never seen them.

In machine learning, the problem of the **prediction of the new** is logically related to the problem of the **generation of the new**. Interestingly, the logical definition of a security issue also describes the logical limit of creativity in machine learning. The trite question "Can AI make art?" should be reformulated in technical terms: Can AI create works that are not imitations of the past? Is AI able to extrapolate beyond the stylistic boundaries of the training data? The answer is: not really. The 'creativity' of machine learning is limited to the **detection of the old** styles from the training data and the subsequent random improvisation along such styles. In other words, machine learning can explore and improvise only within the borders of the categories that are set by the training data. The artworks of the Obvious Collective (*nomen est omen*), a collaborative project that creates paintings using AI, provide visual evidence of these limitations. The style of their portraiture is highly normalised and aesthetically predictable.[36] It would therefore be more accurate to term AI art as *statistical art*.

In terms of natural language processing, one can question whether AI is able to invent new **metaphors** in a consistent and non-random way. In a pre-machine learning age, when asked if a metaphor could be invented by an algorithm, Umberto Eco replied:

> "No algorithm exists for the metaphor, nor can a metaphor be produced by means of a computer's precise instructions,

---

[35] Cp. David Ingold and Spencer Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?", *Bloomberg*, April 21, 2016. Available at: www.bloomberg.com/graphics/2016-amazon-same-day [accessed March 21, 2019].

[36] Cp. James Vincent, "Christie's sells its first AI portrait for $432,500, beating estimates of $10,000", *The Verge*, October 25, 2018. Available at: www.theverge.com/2018/10/25/18023266 [accessed March 21, 2019].

> no matter what the volume of organized information to be fed in."[37]

Any new metaphor is the breaking of a rule and the invention of new one, Eco argued. Can an algorithm be programmed to break the rules (patterns) of its training data in a creative way? Machine learning will never be able to detect or generate Rimbaud's famous line "I is another" after running a statistical analysis of a million newspapers. Machine learning never invents codes and worlds, rather draws **vector spaces** that reproduce statistical frequencies of old data. In computational statistics, a new metaphor is a *new vector* with no similarities of frequency with *old vectors*—something that would easily disappear in the following computational passage. Besides, a metaphor is not the statistical *correlation* of two meanings but the construction of a *new world model* in which this new expression would acquire a logical sense (a *causation*) that it did not have in the old world model. A new metaphor is the invention of a constituent paradigm. Often metaphors are banal, but sometimes they can be brilliant and *open*, like when they leave space for endless interpretation, a process key to the humanities and not only. One wonders who dreams to mechanise hermeneutics, the art of interpretation and aesthetic judgement – processes that are to remain unbound. Although, the art of interpretation can be enriched and extended, of course, by new instruments of logical magnification and pattern exploration.

Conclusion

Any **anomaly** (also a social and political one) is the invention of a new code or rule. On the other hand, power is often based on the normalisation of codes and rules, which seek to minimise the occurrence of the anomalous. Machine learning is no exception when it is applied to the measure and governance of society. See, for instance, the experiment of word embedding by Bolukbasi et al., which used Word2vec as a pre-trained statistical model to analyse Google News posts as training data. When the algorithm was asked to resolve the equation "man is to computer programmer as woman is to *x*", it replied problematically with x = 'homemaker', showing the effect of AI in reinforcing stereotypes.[38] Social and cultural diversities easily disappear in machine learning, as algorithms cannot express semantic depth

---

[37] Umberto Eco, *Semiotics and the Philosophy of Language*, Bloomington, Indiana University Press, 1986, p. 127.

[38] Cp. Tolga Bolukbasi et al., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", *arXiv.org*, July 21, 2016. Available at: arxiv.org/abs/1607.06520 [accessed March 21, 2019].

unless becoming slow and inefficient.[39] AI is representing a more and more standardised world, in which traditional institutional and social norms are translated and further amplified into new statistical and computational norms.[40]

This essay attempted to review the limitations that affect AI as a mathematical and cultural technique, stressing the role of error in the definition of intelligence in general. It constructed a tentative index of limits, approximations, biases, errors, fallacies and vulnerabilities of machine learning. It described machine learning as composed by three parts: training dataset, statistical algorithm and model application (as classification or prediction). It then distinguished three type of bias: world, data and algorithmic bias. It argued that the logical **limits** of statistical models produce or amplify **bias** (that is often already present in the training datasets) and causes **errors** in classification and prediction. However, it is not a machine issue, but a political **fallacy**, when a statistical correlation between numbers within a dataset is received and accepted as causation among real entities in the world. The degree of information compression by the statistical models used in machine learning causes **information loss** also with respect to the granularity of categories and taxonomies, resulting into social and cultural diversity loss. The ultimate limit of AI models is found in the inability to detect and predict a **unique anomaly**, such as a metaphor in natural language. For the same reason, AI systems are also vulnerable to **adversarial attacks** that can be launched by an external operator aware of the weak regions of a statistical model. In the final analysis, the main effect of machine learning on society as a whole is cultural and social **normalisation**. Corporate AI but extends the normative power of former knowledge institutions into the new computational apparatuses. The distorted normativity of AI proceeds from the logical limitations of statistical modelling – a technique that is worshipped, embarrassingly, as animistic totem of superhuman cognition.

---

[39] The impact of AI on society is already registered in everyday behaviours, when people adjust to the algorithm rather than the other way around. It is getting common, for instance, to adjust pronunciation and neutralise intonation to be sure that the voice recognition software of a call centre or smartphone picks up words correctly. This auto-corrective behaviour is an unconscious integration and absorption of the biases of machine learning by society itself.

[40] Cp. Matteo Pasquinelli, "Arcana Mathematica Imperii: The Evolution of Western Computational Norms", in Maria Hlavajova et al. (eds.), *Former West*, Cambridge, MA, MIT Press, 2017, pp. 281–293.