phg foundation
making science work for health

Black box medicine and transparency

Machine learning landscape

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish and Alison Hall

Acknowledgements

The *Black Box Medicine and Transparency* project was funded by the Wellcome Trust as a part of their 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z]. We thank the Wellcome Trust for their support.

The series of reports is informed and underpinned by a series of roundtables and interviews. These roundtables and interviews are detailed in the Report of Roundtables and Interviews. Further, highlights from both are seeded throughout all reports, being found in 'A Salient Feature' boxes.

Disclaimer

URLs in this report were correct as of February 2020

This report is available from www.phgfoundation.org

Published by PHG Foundation 2 Worts Causeway, Cambridge, CB1 8RN, UK

+44 (0)1223 761900

February 2020

Correspondence to: intelligence@phgfoundation.org

How to reference this report:

Ordish J, Hall A. *Black Box Medicine and Transparency: Machine Learning Landscape*. PHG Foundation 2020.

Contents

1. Machine Learning Landscape

Machine learning promises to change the practice of healthcare and transform medical research. Delivering on such promises requires a robust yet flexible policy and a regulatory framework to support the introduction of new technologies. This report provides an introduction to the topic of machine learning, a working definition for 'machine learning' and 'artificial intelligence', consideration of the breadth of applications for machine learning in medical research as well as healthcare, and an overview of the emerging policy landscape.

2. What is machine learning?

This section distinguishes artificial intelligence from machine learning, defines machine learning, and highlights notable differences between data modelling and algorithmic modelling cultures.

a. Artificial intelligence or machine learning?

Machine learning is one methodological approach to artificial intelligence.[1] *Artificial intelligence* (AI) can be defined as 'the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.'[2] As Lipton (2018) notes, early work on AI included a broad set of approaches, machine learning being only one of this set.[3] For instance, *rule-based expert systems* that attempt to systemise knowledge as conditional if-then rules were popular in the 1980s, being largely eclipsed by machine learning in recent years.[4]

We prefer the term 'machine learning' when describing many of the technologies detailed in this report for the following reasons. First, it is the most accurate term to describe the set of techniques we discuss. That is, most of the tools, systems, or devices we discuss belong to the subset of machine learning techniques. Second, there has been a trend to prefer 'machine learning' over 'AI' to distance current research from grandiose claims associated with early research and so the label 'AI.' Consequently, machine learning typically denotes methods that only have task-specific intelligence and lack the broad powers of cognition feared when 'AI' is mentioned.[5] Following this, we use the term machine learning, making clear where the method belongs to AI in general instead.

A Salient Feature | Roundtable 2

Some participants thought that we should do away with the terms 'AI' and 'machine learning' in some circumstances. One thought was that replacing these terms with 'complex modelling' or an equivalent term might avoid public aversion surrounding AI or machine learning and make clear that these methods are contiguous with statistical modelling.

b. Machine learning

Machine learning as a programming paradigm differs from classical programming in that machine learning systems are trained rather than explicitly programmed.[6] Classical programming combines rules and data to provide answers. Machine learning combines data and answers to provide the rules (see Figure 1 below). Machine learning models are trained with many examples (data) relevant to the task, the algorithm finding structure in these examples to provide rules to automate the task.

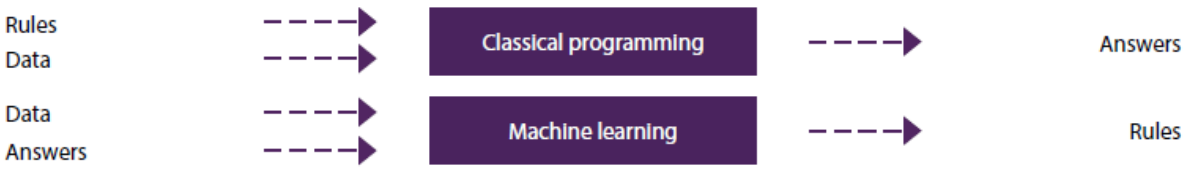


Figure 1: Chollet 2017

Machine learning consists of these three ingredients: features that 'define a language in which we describe the relevant objects, be they emails or complex organic molecules'; tasks being the abstract representation of problems we wish to solve with these objects; and models being the product of applying the machine learning algorithm to the training data.[7]

We distinguish between the following:

> The *machine learning algorithm* (or untrained model) concerns 'how the algorithm learns a model from the data and what kind of relationships it can learn.'[8]
>
> The *machine learning model* (or trained model) is 'produced as the output of a machine learning algorithm applied to training data.'[9] Sometimes also called 'the trained model,' the model here has been trained according to the machine learning algorithm on a training set of data.
>
> The *machine learning system* relates to the device which encompasses the machine learning model. The wider system might include the user interface, supporting architecture, visualisation of the model, as well as any physical device in which the software is embedded.

The term 'machine learning' describes a diverse set of methods to detect and predict patterns - there is no one machine learning technique. It is common to divide the field into three paradigms of machine learning:

I. 'Supervised' (or 'predictive') learning uses training data consisting of labelled sets of input-output pairs. Following these pairs, the model will then learn the features of the input data associated with the labelled outputs. For example, to construct an email spam filter, sample emails (inputs) known to be or not be spam will be labelled as such (output) to constitute a model.

II. 'Unsupervised' (or 'descriptive') learning approaches attempt to find patterns of interest in the data. Unlike supervised learning, the relationship between the inputs and outputs is unknown. Many unsupervised machine learning models are directed toward finding structure in a dataset, often a necessary step to solve a supervised machine learning problem.

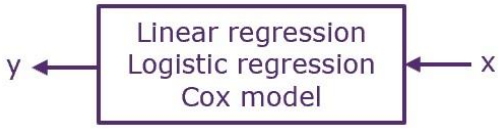
III. 'Reinforcement learning' tells us 'how to act or behave when given occasional reward or punishment signals.'[10] In this way, an 'agent' receives information about its environment and learns to pick actions that maximise some reward. Reinforcement learning has applications across a diverse set of fields, for instance, self-driving, robotics, resource management, and education.

Is machine learning different from traditional modelling techniques?

c. Two cultures

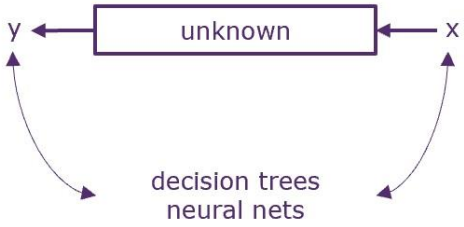
Broadly, machine learning can be conceptualised as a form of statistical modelling. Breiman (2001) distinguishes between two different approaches to statistical modelling:

I. *Data modelling approaches* start by 'assuming a stochastic data model for the inside of the black box.'[11] Following this approach, the black box is filled in by estimating parameters from data and for prediction (see Figure 2 below). Typically, these models are validated by goodness-of-fit tests.



a. Figure 2: Breiman 2001

II. *Algorithmic modelling approaches* 'consider the inside of the box complex and unknown.'[12] These approaches seek to predict response variables based on input variables (see Figure 3 below). Typically, these models are validated by their predictive accuracy.



b. Figure 3: Breiman 2001

As noted by Renkl and Molnar (2019), the algorithmic modelling method often produces black box models as they provide no direct explanation for their predictions.[13] Machine learning typically constitutes an algorithmic modelling approach. Consequently, where black box problem arises, so does the problem of interpretability.

Section 2 key messages:

- **Machine learning is a subset of artificial intelligence, machine learning models being trained with many examples (data) relevant to the task, the algorithm finding structure in these examples to provide rules to automate the task.**
- **Machine learning often counts as an 'algorithmic modelling approach.' This approach assumes a black box that is complex and unknown, predicting input variables to output variables without explaining what happens in between.**

3. Where is machine learning being used?

There are many articles outlining the variety of uses for machine learning in healthcare and research.[14] The following two sections provide an indicative look, first at machine learning for medical research and, second, machine learning for healthcare - both tables emphasise the breadth of machine learning applications in these areas. Prior to these tables, it is important to note the following:

First, some machine learning applications are already in use, are close to implementation, or are more speculative in nature, being uncertain to ever make it to market or be put into service.

Second, the field of machine learning for healthcare and research is currently undergoing a process of systemisation. This is evidenced by journals adopting new standards for publication of articles concerning machine learning,[15] a multitude of articles being published on how to read publications on machine learning with a sceptical eye,[16] and the notable gulf in producing evidence linking an interesting machine learning model to evidence of clinical utility.[17]

Third, the various applications listed vary in how automated they are, that is, each different system automates tasks or decisions in healthcare/research to varying extents. Regardless, it is fair to say that most near-term uses are assistive only - they support, augment, and enhance human healthcare professionals or researchers.

A Salient Feature | Roundtable 2

Participants in Roundtable 2 noted that near term applications for machine learning will be assistive for three main reasons. First, it represents the state of, and limitations of, current technology. Second, combining machine learning with human healthcare professionals and researchers combines the strengths of both and mitigates against differing weaknesses. Third, liability concerns may influence the decision to shift the intended use of machine learning systems to be assistive only, so devices are considered clinical decision support, the primary liability falling to the healthcare professionals interpreting the model.

Fourth, some machine learning systems will qualify as medical devices, being subject to comparatively stringent requirements to evidence their safety and to ensure that they meet their intended use. Some machine learning models, while qualifying as medical devices, will count as health institution exempt devices, being exempt from many of the requirements to evidence safety and effectiveness.[18] Other machine learning systems will not qualify as medical devices at all - perhaps because they are for research use only, or because they are for operational/administrative use only and so lack a specifically medical purpose,[19] or because these devices constitute lifestyle or wellbeing devices, thereby also lacking a medical purpose.[20] For a fuller treatment of machine learning as a medical device, see our *Algorithms as a medical device* report.[21]

a. Machine learning for medical research

Machine learning promises to change numerous diverse parts of medical research and its practice. Table 1 below outlines the breadth of applications for machine learning in this sector.

Table 1: indicative uses of machine learning for medical research

General area of impact	Challenge the tool addresses	Example of a tool	Description of tool
Literature aggregation	The task of compiling systematic reviews. An estimated 2.5 million English language scientific articles are published each year and rising	Project Transform (with EPPI-Centre at University College London), Cochrane[22]	Machine learning to assist with searching, study eligibility assessment, data extraction, and evidence synthesis
Hypothesis generation and targeting A	Linking millions of single-nucleotide polymorphisms to individual traits, typically complex diseases in genome-wide association studies	COMBI[23]	Support vector machine model to narrow a subset of possible SNPs and perform hypothesis testing
Hypothesis generation and targeting B	Finding novel antibiotic candidates	Stokes et al (2020) model[24]	Deep learning models to predict antibiotics based on structure
Knowledge discovery	Disparate sources of knowledge resulting in challenges in knowledge discovery	WuXi NextCODE[25]	Domain-specific AI algorithms for biological understanding, drug discovery and optimal clinical trial design
Understanding fundamental biological processes A[26]	Predicting gene targets of microRNAs	Zurada (1994) model[27]	Deep recurrent neural networks to predict gene targets for microRNAs

Understanding fundamental biological processes B	Predicting the structure and function of proteins	Wang et al (2016)[28]	Machine learning to predict the structure and function of proteins
Epidemiological research - pollution	Finding patterns, extracting information, and making predictions using large epidemiological datasets	Bellinger et al (2017)[29]	Machine learning to source apportionment and forecast air pollution in large datasets
Drug Discovery/development	Tailored drug discovery	BenevolentAI	Drug discovery platform that draws on mined and inferred biomedical data
Cohort selection	Identifying cohorts for clinical trials and studies	Chen et al (2013) using active learning to identify three disease cohorts: rheumatoid arthritis, colorectal cancer, and venous thromboembolism[30]	Active learning to extract phenotypic information from electronic health records
Drug repurposing	Developing effective treatments for rare diseases can be difficult as getting novel drugs to market is an expensive exercise and rare disease populations are, by definition, small	Healx's HealNet[31]	Machine learning to draw on a number of datasets: clinical trials, disease symptoms, drugs targets, multi-omic data, and chemical structures to identify new uses for existing drugs. That is, old drugs, new tricks

As demonstrated, machine learning has a variety of applications for medical research. Medical research inevitably blends into healthcare and public health, as this research underpins the evidence base for current practice but might also have therapeutic objectives itself as research often aims to examine potential diagnoses or treatments.[32]

b. Machine learning for healthcare and public health

Machine learning promises to change the practice of healthcare, having a breadth of applications across patient pathways and the direct to consumer market. Table 2 below outlines the breadth of applications for machine learning in this sector.

Table 2: Indicative uses for machine learning in healthcare and public health

General area of impact	Challenge the tool addresses	Example of tool	Description of tool
Infectious disease tracking	Forecasting zoonotic disease, distinguishing reservoirs from nonreservoirs	Han et al (2015) model[33]	A machine learning model that combines 86 predictor variables to predict novel zoonotic reservoirs and geographic regions where emerging pathogens are most likely to arise
Screening – breast cancer	Breast cancer screening from mammography	McKinney et al (2019) paper[34]	A deep learning model for identifying breast cancer in screening mammograms
Early detection	Distinguishing between stable mild cognitive impairment and progressive mild cognitive impairment (dementia)	Giorgio et al (2020) paper[35]	A machine learning model to predict whether an individual with mild cognitive impairment will decline or remain stable
Medical imaging	The time consuming task of manually delineating radiological images	InnerEye[36]	Automatic delineation of healthy anatomy from tumours
Stratification	Predicting patient risk for complex diseases	Ho et al (2019) review[37]	Machine learning algorithms to improve polygenic risk scores
Phenotyping	Difficulty of recognising facial features related to rare diseases	DeepGestalt[38]	Deep learning driven facial analysis software for rare disease phenotyping and phenotype driven genetic variant prioritisation via smartphone
Appointment scheduling	Non-attendance of hospital appointments	University College London Hospitals and University College London model[39]	A machine learning model to predict which patients will fail to attend appointments
Triaging	Triaging in emergency departments	Levin et al (2017) model[40]	Random forest model including features relating to vital signs, chief complaint, and active medical history that predicts the need for critical care,

			emergency procedures, and inpatient hospitalisation
Diagnosis -Radiology	Diagnosis of difficult to diagnose wrist fractures	OsteoDetect[41]	AI analysis of wrist radiographs to highlight regions of distal radius fractures
Diagnosis - Ophthalmology	Detecting retinal disease	De Fauw et al (2018) model[42]	Deep learning to analyse three-dimensional optical coherence tomography scans to make accurate referrals for retinal disease
Prognosis	Predicting survival in cancer patients beyond 120 days post-palliative chemotherapy	Ng et al (2012) model[43]	A neural network including features relating to patient attributes tumour attributes, treatment attributes, clinical attributes, and laboratory attributes to predict survival of cancer patients beyond 120 days after palliative chemotherapy
Alert systems	Detection of drug to drug interactions, tailoring medication doses in response to age and renal insufficiency	Kuperman et al (2007) review[44]	Machine learning to detect contraindications and tailor dosage
Treatment A	Therapeutic decision support	Cambridge Cancer Genomics[45]	AI platform intended to support oncologists to provide personalised cancer therapy
Treatment B	Differentiating between malignant and benign tissue in breast surgery	iKnife[46]	Analysing the vaporisation of tissue using Rapid Evaporative Ionisation Mass Spectrometry to differentiate malignant and benign tissue while in surgery
Management of conditions	Predicting patient glucose levels	Plis et al (2014) model[47]	Support vector regression to predict blood glucose levels
Patient facing tools	Counselling patients for genetic services	Clear Genetics[48] OptraHealth[49]	AI chatbot/digital assistant for conversing with patients about genetics AI chatbot that can be queried via virtual assistant tools such as Amazon Alexa and Microsoft Cortona

In addition to the variety of direct uses for machine learning in healthcare and public health, machine learning also has applications supporting health systems via operational and administrative support. For example, automated scheduling systems to manage staff rotas.

As demonstrated, machine learning has diverse applications in the healthcare sector. However, the success of its implementation depends upon the technology being supported by policy. It is to the development of that policy that we now turn.

Section 3 key messages:

- **Machine learning has wide application across medical research and healthcare. Notably, some applications are already in use, some are close to implementation, others are more speculative. Further, it is also important to note that machine learning is also undergoing systemisation - reporting standards are being improved.**
- **Machine learning for medical research has a number of applications. Medical research often blends into healthcare, research underpinning the delivery of care but also often constituting care, healthcare and including therapeutic intent as well.**
- **Machine learning for healthcare has broad potential application, much of the patient pathway and direct to consumer market, being subject to at least speculative machine learning involvement.**

4. Policy context

As technical aspects of AI and machine learning gain pace, and ML applications are implemented across many different sectors, there has been a proliferation in the policy documentation generated on this topic. This documentation varies in territorial scope, sectoral specificity, and the time of publication. A comprehensive review of the policy landscape is outside the scope of this report, but this section highlights some of the key resources which are relevant to health care and medical research within the UK.

This documentation emanates from a variety of different sources. In this section, we will focus on guidance from non-legislative sources, such as policy think tanks, and professional bodies, sometimes in collaboration with regulators. This policy guidance can be classified in a number of different ways: in terms of its jurisdictional scope, the extent of its focus on AI and machine learning, and, lastly, its sectoral scope (i.e. relating to health). It can also be classified according to the extent to which it has statutory weight and is likely to be used by regulators, or will be used to guide, but not mandate, best practice.

In the UK, there are several key sources of advice and guidance on AI development. Important stakeholders include the Information Commissioner's Office (ICO) which constitutes the statutory authority for upholding information rights, including the General Data Protection Regulation, UK Data Protection Act 2018 and law enforcement processing. Within the health sector, NHSX has responsibility for establishing a framework for developing AI in the health and care system. Other organisations, including the National Institute for Health and Care Excellence have developed evidence standards, complementing guidance from ICO and NHSX to ensure that new technologies are clinically effective and offer economic value.

a. Jurisdiction: International, supranational and national resources

There is increasing recognition that determining appropriate ethical and regulatory oversight of AI applications is a universal challenge, which can best be met by consistent and harmonised approaches. Although there are some examples of global policy initiatives mostly concerning autonomous or intelligent systems,[50] the requirement for developers to take account of European legislation and regulation suggests that European level guidance might be more authoritative in practice. One of the key elements of this project concerns the legal and regulatory requirements for transparency and explanation, and legislation which applies at European level. Therefore, the policy context includes guidance on interpretation of relevant European laws.

b. EU High-level expert group on artificial intelligence

This independent group, convened in June 2018 had the mandate of creating a framework for developing Trustworthy AI through two deliverables, AI Ethics Guidelines[51] and Policy and Investment Recommendations[52]. In draft Guidance, published in April 2019, it opined that Trustworthy AI has 3 components (1) lawful (applicable laws) (2) ethical (3) robust (from a technical and social perspective). Although the Guidelines cover second and third elements only and offer guidance in varying levels of abstraction, the authority for these Guidelines stem

from fundamental rights enshrined in the Charter of Fundamental Rights of the European Union and in relevant international human rights law.

Chapter 1 cites explicability as an ethical principle that should underpin the development, deployment and use of AI systems. Where explicability is impossible, for example, for some 'black box' algorithms, the guidance states that there may be a requirement for other explicability measures (such as traceability, auditability and transparent communication on system capabilities). The guidance notes that the degree to which explicability is needed 'is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.'[53]

Chapter II sets out seven requirements that AI systems should meet. These build on the order of principles and rights in the EU Charter. Transparency – listed 4th – incorporating requirements for traceability, explainability, and communication, is one of the seven identified requirements but it is not acknowledged that meeting some of the other requirements (e.g. for technical robustness and safety, or accountability) may also be predicated on transparency.[54] This chapter also highlights the need for clear and proactive communication with stakeholders about the use of an AI system and its capabilities and limitations, and the requirement to provide human interaction as an alternative to using an AI system, where needed to ensure compliance with fundamental rights.

Chapter III suggests that these ethical and practical requirements need to be tailored in accordance with specific AI applications, and suggests both technical and non-technical measures.[55]

The High-Level Expert Group on AI have also developed a set of policy and investment recommendations building on the Trustworthy AI framework.[56] This incorporates 33 recommendations that 'can guide Trustworthy AI towards sustainability, growth and competitiveness, as well as inclusion – while empowering, benefiting and protecting human beings.' It is notable that these recommendations distinguish between business to consumer (B2C) (1/3 value), business to business (B2B) (2/3 value) and public to citizen (P2C) scenarios where the trust of individuals is described as 'an even more crucial prerequisite.'[57] A revised version of the assessment list is currently being developed. Thus whilst transparency underpins much of this guidance, it is moderated by context, specifically application and user.

c. National policy guidance

NHSX is a virtual body incorporating teams from the Department of Health and Social Care, NHS England and NHS Improvement, with responsibility for establishing a framework for developing AI in the health and care system in the UK. Building on the NHS Long Term Plan,[58] which included applications of machine learning from incident data to improve patient safety alerts, and in mental health to predict suicide, core sectoral guidance is enshrined in a Code of conduct for data-driven health and care technology[59] which supports the development of data-driven technologies in a 'safe, ethical, evidenced and transparent way' through the application of principles. The requirement for transparency forms part of a number of the principles in this code, for example:

- Principle 4: be fair, transparent and accountable about what data is being used
- Principle 6: be transparent about the limitations of the data used
- Principle 7: show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and provision.[60]

The application of these principles is elaborated in a recent report which advocates that the 'right to an explanation' should be situated within a wider context:[61]

> 'Ethical and behavioural principles are necessary but not sufficient to ensure the design and practical implementation of responsible AI. The ultimate aim is to build transparency and trust.'

A key step is to explain the algorithm to those taking actions based on its outputs and those on the 'receiving end' of the decision making process. Dimensions include both content (e.g. the extent of automation/human intervention; what is meant by the term – meaningful explanation) and process (e.g. the need to coordinate with patient representative groups to develop meaningful language), and the use of trusted third parties (disease specific charities) to act as advocates for patient groups.

This guidance from NHSX is aimed at the entire lifespan of an AI tool, from inception to implementation and post-marketing surveillance. This guidance is supplemented by publications from other regulatory bodies such as NICE [62] which sets out the evidence standards for digital health technologies intended for use within the UK health and care system. Aimed at technology developers and commissioners of digital health technologies, this guidance adduces functional classifications for digital technologies through a checklist, and within each tier, provides two levels of evidence – a minimum evidence standard and best practice standard.

Despite this emerging guidance, and a proliferation of tools to support developers across the innovation pathway including the development of an online workbook of the Code in the form of a self-assurance portal[63] a substantial minority of developers either lack insight about the need for explainability, or acknowledging the need, are uncertain what the requirements are, or how best to satisfy them. For example, the NHSX report cites a state of the nation survey in which 19% of developers were uncertain if they had incorporated the explainability of the system into its design, and a further 9% stated that they had not done so.[64] This survey suggests that there is continuing uncertainty about the nature of explanation and its status in the development process.

Section 4 key messages:

- **There has been a proliferation of policy on AI. This can be classified in terms of its jurisdictional scope, its focus on AI and machine learning, and sectoral scope: consequently it varies in specificity and statutory weight.**
- **In the UK, key sources of advice and guidance include the Information Commissioner's Office, which is the UK statutory authority for upholding information rights, and within the health sector, NHSX, which has responsibility for establishing a framework for developing AI in the health and care sector.**
- **There is increasing recognition that determining appropriate ethical and regulatory oversight of AI is a universal challenge which is best met by consistent and harmonised approaches.**
- **Important sources of guidance include the AI Ethics Guidelines developed by the EU High-Level Expert Group on artificial intelligence which recognise explicability as an ethical principle underpinning the development, deployment and use of AI systems.**
- **The UK NHSX's Code of conduct for data-driven health and care technology also highlights the importance of transparency as a principle underpinning the development process. However both sets of guidance do not provide more granular information about the form, content and timing of an explanation.**

5. A call for interpretability

This report has defined machine learning, outlined the variety of applications for the technology in healthcare as well as research, and has digested an array of policy that aims to support the technology's uptake. This last section outlines two cases that illustrate the importance of interpretability in machine learning, the topic of the next Interpretable Machine Learning report.

Why might interpretability of machine learning models be important in healthcare and research? In regards to healthcare, the following two examples might be instructive.

Caruana et al (2015) describe a series of models to predict the probability of death for patients with pneumonia.[65] The group found that neural networks produced the most accurate models. However, when the group trained in parallel a less accurate but interpretable rule-based model, the group found that this model learned the following rule: 'HasAsthma(x) LowerRisk(x).' Consequently, it was shown that a confounding variable influenced the neural networks, the models correctly identifying that those with asthma were less likely to die, but only because as a group they were more likely to receive treatment.

Zech et al (2015) trained a convolutional neural network to screen for pneumonia using x-rays.[66] Subsequent manual image review noticed that the model was able to differentiate between those x-rays taken by portable scanner (identified by the word 'portable' and inversion of colour in the x-ray) and those by static scanner, the model finding this distinction significant, portable scanners being used in the emergency department but not for inpatient units. Consequently, when the model found the word 'portable' significant it introduced a potentially confounding factor into the screening process.

Following Caruana (2015) and Zech (2015), it is clear that interpretability (or the lack thereof) has the potential to impact upon the safety and effectiveness of machine learning systems. We examine the interpretability of machine learning in the Interpretable Machine Learning report and provide a framework for developers and product managers to think through interpretability with respect to their model in the Interpretability by Design Framework.

Section 5 key messages:

- **Caruana (2015) notes an example where a confounding factor was found to underpin a model predicting risk of death through pneumonia. Zech (2015) outlines an example where a confounding factor was found to dictate the outcome of a convolutional neural network to screen for pneumonia in x-rays.**
- **The Caruana and Zech examples demonstrate the potential importance, (although not universally required) of interpretability in ensuring models are safe and meet their intended purpose.**

6. The landscape of machine learning

The landscape of machine learning for healthcare and medical research includes an astonishingly broad range of applications. These applications present an opportunity to change the practice of healthcare and medical research. However, their successful implementation into service is contingent upon a number of factors, for instance, the introduction of machine learning requires robust, agile policy to facilitate uptake. Notably, the breadth of machine learning in these sectors also has to be safe, usable, and ultimately effective. Machine learning is commonly thought to face barriers in establishing its safety, usability, and effectiveness because the technique used for this purpose can often be a black box. The next Interpretable Machine Learning report considers interpretability of machine learning models and the methods to render otherwise uninterpretable models interpretable.

References

[1] Lipton ZC. *From AI to ML to AI: On Swirling Nomenclature & Slurried Thought*. Available from: http://approximatelycorrect.com/2018/06/05/ai-ml-ai-swirling-nomenclature-slurried-thought/ [Accessed 24 February 2020].

[2] High P. *Carnegie Mellon Dean of Computer Science on the Future of AI*. Available from: https://www.forbes.com/sites/peterhigh/2017/10/30/carnegie-mellon-dean-of-computer-science-on-the-future-of-ai/#78648b612197 [Accessed 24 February 2020].

[3] Lipton ZC. *From AI to ML to AI: On Swirling Nomenclature & Slurried Thought*. Available from: http://approximatelycorrect.com/2018/06/05/ai-ml-ai-swirling-nomenclature-slurried-thought/ [Accessed 24 February 2020].

[4] Buchanan BG, Duda RO. Principles of Rule-Based Expert Systems. *Advances in Computers*. 1983; 22: 163-216.

[5] Korf RE. Does Deep-Blue use AI? *ICGA Journal*. 1997; 20(4): 243-245.

Lipton ZC. *From AI to ML to AI: On Swirling Nomenclature & Slurried Thought*. Available from: http://approximatelycorrect.com/2018/06/05/ai-ml-ai-swirling-nomenclature-slurried-thought/ [Accessed 24 February 2020].

Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press; 2014.

[6] Chollet F. *Deep Learning with Python*. Version 6. New York: Manning Publications; 2017: 2-3.

[7] Flach P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge: Cambridge University Press; 2012: 13.

[8] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[9] Flach P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge: Cambridge University Press; 2012: 13.

[10] Murphy K. *Machine Learning: A Probabilistic Perspective*. Massachusetts: MIT Press; 2012: 2.

[11] Breiman L. Statistical Modeling: The Two Cultures. *Statistical Science*. 2001; 16(3): 199.

[12] Ibid.

[13] Renkl E, Molnar C. Introduction. In: Molnar C, Casalicchio G, Konig G, et al (eds.) *Limitation of Interpretable Machine Learning*. 2019. Available from: https://compstat-lmu.github.io/iml_methods_limitations/introduction.html [Accessed 24 February 2020].

[14] For instance: Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nature Biomedical Engineering*. 2018; 2(10): 719-731.

[15] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *The Lancet*. 2019; 393(10181): 1577-1579.

[16] For instance: Chen PHC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*. 2019; 322(18): 1806-1816.

[17] Chen PHC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*. 2019; 322(18): 1806-1816.

[18] Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices [2017] OJ L117/1, art 5(5)

Regulation (EU) 2017/746 of the European Parliament and of the Council on *in vitro* diagnostic medical devices [2017] OJ L117/176, art 5(5).

[19] C-219/11 *Brain Products v BioSemi* [2012] ECR-I 742, para 17.

[20] Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices [2017] OJ L117/1, recital 19.

Regulation (EU) 2017/746 of the European Parliament and of the Council on *in vitro* diagnostic medical devices [2017] OJ L117/176, recital 17.

[21] Ordish J, Murfet H, Hall A. *Algorithms as medical devices*. PHG Foundation, 2019.

[22] Takeda K. *Text Mining to Improve the Health of Millions of Citizens*. Available from: https://docs.microsoft.com/en-gb/archive/blogs/machinelearning/text-mining-to-improve-the-health-of-millions-of-citizens [Accessed 22 Feburary 2020].

[23] Meith B, Kloft M, Rodriguez JA, et al. Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Scientific Reports*. 2016; 6: 36671.

[24] Stokes JM, Yang K, Swanson K, et al. A Deep Learning Approach to Antibiotic Discovery. *Cell*. 2020; 180(4): 688-702.

[25] Wuxi NextCODE. *Wuxi NextCODE*. Available from: https://www.wuxinextcode.com/ [Accessed 23 February 2020].

[26] Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *Journal of the Royal Society*. 2018; 15(141).

[27] Zurada J. End Effector Target Position Learning Using Feedforward with Error Back-Propagation and Recurrent Neural Networks. *Proceedings of 1994 IEEE International Conference on Neural Networks*. 1994; 4: 2633-2638.

[28] Wang S, Peng J, Ma J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports*. 2016; 6(1): 1-11.

[29] Bellinger C, Jabbar MS, Zaiane O. A Systematic Review of Data Mining and Machine Learning for Air Pollution Epidemiology. *BMC Public Health*. 2017; 17(1): 907.

[30] Chen Y, Carroll RJ, Hinz ER, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*. 2013; 20(2): 253-259.

[31] HealX. *Fragile X Syndrome: Drug Repurposing Summary Report*. HealX. 2017.

[32] Kass NE, Faden RR, Goodman SN, et al. The research-treatment distinction: a problematic approach for determining which activities should have ethical oversight. *Hastings Center Report*. 2013; 1.

[33] Han BA, Schmidt JP, Bowden SE, et al. Rodent reservoirs of future zoonotic diseases. *Proceeding of the National Academy of Sciences*. 2015; 112(22): 7039-7044.

[34] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020; 577(7788): 88-94.

[35] Giorgio J, Landau S, Jagust W, et al. Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease. *Preprint*. 2020.

[36] Microsoft. *Project InnerEye – Medical Imaging AI to Empower Clinicians*. Available from: https://www.microsoft.com/en-us/research/project/medical-image-analysis/ [Accessed 22 February 2020].

[37] Ho DS, Schierdling W, Wake M, et al. Machine learning SNP Based Prediction for Precision Medicine. *Frontiers in Genetics*. 2019; 10.

[38] Gurovich Y, Hanani Y, Bar O, et al. DeepGestalt – Identifying Rare Genetic Syndromes Using Deep Learning. *arXiv*. 2018.

[39] University College London Hospitals. *UCLH developers AI to identify patients likely to skip appointments*. Available from: https://www.uclh.nhs.uk/News/Pages/UCLHdevelopsAItoidentifypatientslikelytoskipappointments.aspx [Accessed 22 February 2020].

[40] Levin S, Toerper M, Hamrock E, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of Emergency Medicine*. 2018; 71(5): 565-574.

[41] FDA. *FDA permits marketing of artificial intelligence algorithm for aiding providers in detectin*g wrist fractures. Available from: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures [Accessed 25 February 2020].

[42] De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*. 2018; 24(9): 1342-1350.

[43] Ng T, Chew L, Yap CW, et al. A clinical decision support tool to predict survival in cancer patients beyond 120 days after palliative chemotherapy. *Journal of Palliative Medicine*. 2012; 15(8): 863-869.

[44] Kuperman GJ, Bobb A, Payne TH, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. *Journal of the American Medical Informatics Association*. 2007; 14(1): 29-40.

[45] Cambridge Cancer Genomics. *OncOS*. Available from: https://www.ccg.ai/oncos [Accessed 22 February 2020].

[46] St John E, Balog J, McKenzie JS, et al. Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. *Breast Cancer Research*. 2017; 19(1): 59-73.

[47] Pils K, Bunescu R, Marling C, et al. A machine learning approach to predicting blood glucose levels for diabetes management. *Workshops at the 28th AAAI Conference on Artificial Intelligence*. 2014.

[48] ClearGenetics. *ClearGenetics*. Available from: https://www.cleargenetics.com/ [Accessed 22 February 2020].

[49] OptraHEALTH. *GeneFax*. Available from: https://www.optrahealth.com/ [Accessed 22 February 2020].

[50] Institute of Electrical and Electronics Engineers. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Version 2. 2017: 158-161.

[51] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019.

[52] High-Level Expert Group on Artificial Intelligence. *Policy and Investment Recommendations for Trustworthy AI*. 2019.

[53] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019: 12.

[54] Ibid, 18.

[55] Ibid, 24.

[56] High-Level Expert Group on Artificial Intelligence. *Policy and Investment Recommendations for Trustworthy AI*. 2019.

[57] Bughin J, Seong J, Manyika J, et al. *Notes from the AI Frontier: Modeling the impact of AI on the world economy discussion paper*. McKinsey Global Institute. 2018: 7.

[58] NHS. *The NHS Long Term Plan*. 2019.

[59] Department of Health and Social Care. *Code of conduct for data-driven health and care technology*. Available from: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology [Accessed 22 February 2020].

[60] Ibid.

[61] Joshi I, Morley J (eds.). *Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care*. London: NHSX. 2019: 28.

[62] National Institute for Health and Care Excellence (NICE). *Evidence Standards Framework for Digital Health Technologies*. 2019. Available from: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf [Accessed 22 February 2020].

[63] Joshi I, Morley J (eds.). *Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care*. London: NHSX. 2019: 36.

[64] Ibid, 23.

[65] Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015: 1721-1730.

[66] Zech JR, Badgeley MA, Liu M, et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PloS Medicine*. 2015: 1-15.

The Black box medicine and transparency report was funded by the Wellcome Trust as part of the 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z].

We thank the Wellcome Trust for their support.

The PHG Foundation is a non-profit think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare and deliver improvements in health for patients and citizens.

For more information contact:

intelligence@phgfoundation.org

UNIVERSITY OF CAMBRIDGE

phg foundation

making science work for health