phg foundation
making science work for health

A right to explanation?

Black box medicine

Discussion paper

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish and Alison Hall

Acknowledgements

This work was supported by the Wellcome Trust, grant number: 213623/Z/18/Z

URLs in this paper were correct as of November 2019

This discussion paper can be downloaded from:
www.phgfoundation.org

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK
+44 (0)1223 761900

May 2019

Correspondence to:
intelligence@phgfoundation.org

How to reference this publication
A right to explanation?
PHG Foundation (2019)

A right to explanation?

From research that underpins scientific discovery to how we diagnose and ultimately treat patients, machine learning is set to transform healthcare[1]. Machine learning's implementation into practice will, in part, depend upon how this technology is perceived by potential users and patients.

Machine learning models are built upon training and test data. The data processing which underpins the development of machine learning applications and their continued use is therefore key. This data may count as personal data (and sensitive personal data) and be regulated by the General Data Protection Regulation (GDPR). One of the most contentious elements of the GDPR is the right to explanation. The very existence of this right, its interpretation, and how it might be satisfied is contested. This paper outlines the right to explanation and other mechanisms the GDPR provides that might require explanation of machine learning models and their outputs.

Summary

- Machine learning for healthcare is a promising technology but some models may be black boxes - their workings may be opaque
- The GDPR contains a specific right to explanation under Article 22. However, only a subset of machine learning for healthcare will trigger this narrow right
- It is unclear how the right to explanation and transparency requirements will apply to machine learning, key questions include: when does the right apply, what has to be explained, and what kind of explanation would suffice?

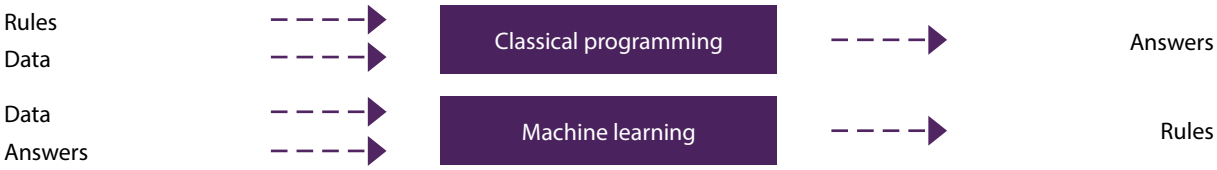
Machine learning for healthcare

Machine learning has many potential applications in healthcare, the table below details three near-implementation applications .

Challenge the tool addresses	Example of a tool	Solution the tool provides
Manual interpretation of radiological images is time consuming	Microsoft Research's Inner Eye	Machine learning for automatic delineation of healthy anatomy from tumours
Diagnosing wrist fractures in a timely manner is difficult	OsteoDetect	AI analysis of wrist radiographs to highlight regions of distal radius fractures
An estimated 1.5 to 3 million people in the UK who attended emergency departments 'could have had their needs addressed in other parts of the urgent care system[2]	Babylon Health's Babylon Check	Automated triage system to route patients to the appropriate service

Machine learning and black boxes

Classical programming combines rules and data to provide answers. Machine learning combines data and answers to provide the rules (see diagram below). Machine learning systems are trained with many examples (data) relevant to the task, the system finding structure in these examples to provide rules to automate the task[3].



A potential disadvantage of using these tools is that many machine learning models may be black boxes, that is, models 'whose internal workings are either unknown to the observer or known but uninterpretable to humans'[4]. In short, it may be difficult to explain why a machine learning model generated a certain output. However, not all machine learning models are human uninterpretable - some techniques are visualizable and so susceptible to human interpretation. Moreover, there are methods to make an otherwise opaque machine learning model somewhat transparent by creating a model-agnostic explanation that approximates the relationship between inputs and outputs, illustrating the model's internal workings. Further, rather than explaining the model as a whole, example-based explanations may be used to explain particular decisions of the model. However, this raises the question: why explain?

Why explain?

A legal obligation to provide an explanation is only one reason to ensure a machine learning model is human interpretable. In the context of healthcare, it might be necessary to explain the workings and contextualise the outputs of a machine learning model for it to be regarded as a viable product and be trusted by clinicians and patients. There may also be an ethical imperative to explain models, especially if models are used for an individual's diagnosis or treatment or for maintaining accountability.

These reasons aside, various sources of law may generate an obligation to explain otherwise human uninterpretable models. Chiefly, medical negligence, medical device law, administrative law, and human rights instruments may individually or collectively generate a duty to explain. In this paper we focus on obligations found in the GDPR.

Duties to explain under the GDPR

The GDPR provides data subjects with at least two potential routes to open black boxes, namely:

I. the right to explanation under Article 22(1); and

II. the general principle of transparency spread across the Regulation but rooted in Article 5(1)(a).

We examine both the right to explanation and the general principle of transparent processing in turn.

Structure of the right to explanation

> Article 22(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

This right to explanation is a composite right found across the GDPR. Article 22(1) contains a general prohibition against automated processing. However, most elements referencing explanation are found elsewhere in the rights to information and access, specifically Articles 13(2)(f), 14(2)(g), 15(1)(h) and supporting interpretative aids (recitals).

It is these provisions triggered by Article 22(1) that contain reference to giving 'meaningful information about the logic involved' and the consequences of data processing.

Will my device trigger Article 22(1)?

Not all machine learning for healthcare will be caught by the right to explanation (narrowly interpreted) in Article 22(1). To trigger Article 22(1), the processing of data in question must be:

I. based solely on automated processing; and

II. produce legal effects concerning or similarly significantly affects the data subject.

Working Party 29 Guidelines on Automated individual decision-making elaborate on each of these elements:

'Based solely on automated processing' means there is no human involvement in the decision process. However, this human involvement cannot be 'fabricated' and must be more than a token gesture - the human must have actual authority and influence over the decision.

'Legal effects' means the decision affects the data subject's legal status, legal rights, or rights under contract.

'Similarly significant affects' means that the decision must have similar significance to legal effects, being sufficiently important to be 'worthy of attention.' Recital 71 gives some examples: 'e-recruiting practices without human involvement' and 'automatic refusal of online credit applications.'

While the above guidance is vague, it is clear that only a subset of near-use machine learning for healthcare will either be solely automated and also produce legal/similarly significant effect (see table below). Only machine learning devices in Category A will engage the narrow right to explanation in Article 22(1).

Triggering Article 22(1)	Based solely on automated processing	NOT based solely on automated processing
Produces legal effects or similarly significant affects	(Article 22(1) triggered) A	B
Does NOT produce legal effects or similarly significant affects	C	D

Transparency apart from Article 22

Article 22 is not the only mechanism under the GDPR that might generate a duty to explain machine learning models and their outputs. The general principle that personal data be 'processed lawfully, fairly and in a transparent manner' underpins the rights to information, access, and other GDPR rights[5]. In short, general principles of transparency and the need to meet other GDPR rights may necessitate explanation, even if this requirement is less onerous than that found in Article 22.

What does transparency in general require? The rights to information and access, accompanied by their recitals will likely require some explanation of machine models as a whole. More controversially, these general transparency requirements may require explanation of a specific decisions and processing of machine learning models. While it is unclear what a duty to explain under the general principle of transparency might require, it is clear that explanation of specific decisions would be a more demanding requirement.

What does a duty to explain require?

The proper interpretation of the right to explanation and its relation to the broader principle of transparency is highly contentious. These interpretative debates have real consequences for what the GDPR will require in terms of explanation of machine learning. Broadly, there are those that emphasise the human rights pedigree of the GDPR, noting that the purpose of the right to explanation is to vindicate more general rights to transparency[6]. These commentators typically think that the right to explanation can require explanation of systems as a whole but as well as individual decisions.

On the other hand, there are those that base their interpretation on the gradual evolution of the GDPR from the Data Protection Directive, drawing a sharp distinction between the interpretative recitals and the legally effective articles of the GDPR[7]. These commentators typically think that the right to explanation does not require explanation of individual decisions and prefer to call the right a 'right to be informed' instead.

These interpretative disputes over the right to explanation and transparency have deep implications for determining what the duty to explain requires. Major uncertainties include:

When is explanation required? Is explanation required before the data is processed and/or after processing?

What is to be explained? Must data controllers explain the model and how it functions as a whole and/or must they provide an explanation of individual decisions post-processing?

What kind of explanation is required? Might counterfactual explanations (that describe the nearest possible world where the result sought was obtained) suffice[8]?

The GDPR's right to explanation and transparency requirements were implemented to protect data subjects and foster good data protection practice. However, the interpretation of these requirements and how they apply to machine learning is in a state of chronic uncertainty. This uncertainty threatens to undermine the goals of the Regulation and acts as a barrier to the development and implementation of machine learning for healthcare. Further guidance clarifying the above questions is urgently needed.

References

1. Ordish J, Hall A. Algorithms as Medical Devices. PHG Foundation. 2019.

2. NHS England: Next Steps on the NHS Five Year Forward View. 2017.

3. Chollet F. Deep Learning with Python. Manning Publications; 2017: 2-3

4. Guidotti R, Monreale A, Ruggieri S, et al. A Survey of Methods for Explaining Black Box Models. ACM Computer Surveys. 2019; 51(5): 1-42

5. Article 5(1)(a), Articles 13-15. Regulation (EU) 2016/679 GDPR

6. Selbst AD, Powles J. Meaningful Information and the Right to Explanation. International Data Privacy Law. 2017; 7(4): 233-242

7. Wachter S, Mittelstadt B, Floridi L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law. 2017; 7(2): 76-99

8. Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology. 2017; 31(2): 1-44

The PHG Foundation examine the right to explanation and related questions in our Wellcome Trust funded project Black Box Medicine and Transparency.

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
+44 (0) 1223 761900

@phgfoundation
www.phgfoundation.org