On Promoting Patient Rights and Algorithmic Accountability in Automated Decision-Making Systems

Ashe Magalhaes[a,1] **and Connor Hounslow**[b,1]

[a]University of Edinburgh Informatics Department; [b]University of Edinburgh School of Social and Political Sciences

This manuscript was compiled on May 14, 2019

As automated decision-making systems proliferate into social protection systems like healthcare, it is critical that a measure of algorithmic accountability be included to safeguard patient rights. This paper examines the DeepMind-NHS Service Agreement as a case study to explore the interpretability-explainability gap that exists in DeepMind's recent AI research and why such a gap may be threatening to patients. This paper advocates for two legislative steps to better ensure due process: (a) disclosing if a decision is automated and (b) to external auditing mechanisms.

Algorithmic Accountability | Explainability | GDPR | DeepMind | NHS

Algorithmic accountability incorporates a broader debate on fairness and transparency by calling for organizations to provide justification for decisions made by automated systems (1). As decision-making increasingly transfers from humans to computers, accountability is necessary to clarify contributing factors to the decision (2) so that a clear attribution of responsibility and standards exists for the subjects that believe the decision is erroneous or unfair. The problem of algorithmic accountability stems from the integration of black-box, or opaque, deep learning classifiers within automated decision-making systems (ADMS) (3). Such classifiers are troubling given the discriminations and biases that permeate society and their potential to be reflected and reinforced in algorithms (4).

This is particularly relevant as ADMS are integrated within social protection systems, such as the health care system. We examine the Service Agreements between Google DeepMind and the National Health Service (herein NHS) and note the interpretability-explainability gap that exists in DeepMind's recent Artificial Intelligence (AI) research, raising concerns around the protections offered to patients.

1. The Proliferation of Automated Decision-Making Systems

ADMS are a form of statistical risk assessment that have the potential to streamline bureaucratic procedures and improve services. As an example, ADMS are used in the financial sector to aggregate customer data (e.g. tenure with bank, number of accounts, demographic variables) in order to automate lending decisions (5, 6). The rise of ADMS can be seen in their increasing deployment across immigration, criminal justice, and healthcare (7–9). As it relates to health and social care, ADMS are promising given their ability to reduce suffering through early detection of disease (10), reduced error rate in diagnosis (11), and personalization in treatment (12).

While deep learning models offer the promise of novel predictive capabilities, they are notoriously unexplainable (13), meaning that it is not clear why a given output was produced. Although existent human bureaucratic processes could be perceived as unexplainable, experts believe AI applied to patient data must be held to a higher, more transparent, standard (14).

2. Defining Explainability

In machine learning, the first step towards explainability is interpretability, loosely defined as comprehending *what* a model did or might have done as distinct from *how* it did it. The difference between interpretability and explainability can be better understood through the analogy of a chemistry experiment: interpretability is the observation of a difference in color or smell during a chemical reaction while explainability is the understanding of the molecular interactions that produced the observed output.

In order to develop a more robust understanding of explainability, the United States Defense Advanced Research Projects Agency developed a framework for measuring explanations. An ADMS' explanation effectiveness was measured by a user's satisfaction with the explanation, the ability of the user to intervene in the the process at some point, a clarification of the ADMS' mental model, the user trust within the system, and the utility of explanation (15). This framework provides an initial technical understanding of what explanation can and should amount to within social protection systems.

3. The DeepMind-NHS Case Study

The DeepMind-NHS collaboration can be used as a case study to support the need for stronger safeguards for patients as data-subjects, given the current state of explainability in AI. DeepMind is a leading AI research company and as a subsidiary of Alphabet Inc., benefits from the computing resources of one of the world's most valuable conglomerates. We operate under the assumption that if their current (2019) body of publications does not reflect an ability to integrate explainability with deep learning classifiers, it is reasonable to expect that this is the case for the AI field more widely.

In 2015, DeepMind partnered with The Royal Free London NHS Foundation Trust to build a smartphone app, Streams, to help clinicians manage acute kidney injury [16]. This led to the transfer of an estimated 1.6 million patients' sensitive medical data to DeepMind, which has since been found to be in

[1]Ashe Magalhaes and Connor Hounslow contributed equally to this work.

[2]To whom correspondence should be addressed: ashe.magalhaesgmail.com and chounslow11@gmail.com

violation of data protection legislation by the UK's Information Commission Office [17]. While DeepMind has acknowledged faults in their data-handling process, work on Streams and AI research on patient data continues under DeepMind Health, which reports directly to Alphabet Inc. [18].

4. Examples of the Interpretability-Explainability Gap

This section examines DeepMind's three most recent publications, which do not mention an application of their research to healthcare. This survey is meant to illustrate the gap between interpretability and explainability in deep learning models, which may prove problematic if applied to patient data.

A. Agents that Infer Representations From Artificial Constructs. DeepMind's research on unsupervised speech representation learning includes mapping discrete representations to phonemes, discrete components of speech sound [19]. While this mapping adds a form of interpretability to learned representations, this may be problematic for assessing agent performance because phonemes are somewhat arbitrary categories that humans have imposed on speech signals rather than quantitative acoustical physical waves. More broadly, this metric of interpretability asks algorithms to infer representations which are artificial constructs, which linguists do not agree on, making full model explainability challenging.

B. Agents that Infer Causal Structure. DeepMind's research includes the first demonstration of model-free reinforcement learning which generates causal reasoning, measured by an agent's ability to perform tasks dependent on causal structure [20]. In healthcare, causality is likely a combination of genetic factors, environmental factors, and lifestyle choices, making the isolation of causal structure difficult [21]. Therefore, task assessments on such tasks as applied to healthcare may be interpretable but not entirely explainable, as with the chemistry example explained in section II.

C. Agents that Pose New Objectives. Finally, open-ended learning algorithms can create agents that exhibit unknown or unexpected behavior, producing a population of improved agents in settings such as Chess or Go [22]. This adaptive approach to posing new objectives which an agent maximizes may be promising for producing diverse populations that simulate human expert decisions [23] but may not be explainable or could prove problematic if harmful objectives are maximized.

5. A Patient's Rights in the UK

Given the interpretability-explainability gap that exists in DeepMind's recently published AI research, the use of ADMS in healthcare challenges the existing set of standards–rights and responsibilities–that defines the relationship between doctors and patients within the NHS. The agreements between the NHS trust and Google DeepMind operate within an existing system of human rights within health and social care systems; this section defines the current landscape of patient's rights in the UK.

As it relates to patients' data, the primary legislative texts are the European Union's General Data Protection Regulation and the Data Protection Act (2018). The GDPR establishes a data subject's 'right to be informed' about the logic involved (Articles 13-15) and the 'right not to be subject to automated decision-making' (Article 22). This does not amount to the 'right to explanation' where an individual is able to explain how the ADMS arrived at some conclusion, i.e. a post *hoc* explanation (16).

The standard for a patient's right to due process within social protection systems is stated by the International Labour Organization in the Convention Concerning Minimum Standards of Social Security, 1952 (No.101). Art 70 states that 'Every claimant shall have a right of appeal in case of refusal of the benefit or complain to its quality or quantity.'

Within the UK, the standard of due process for a doctor-patient relationship is set out in the NHS Constitution, established by the Health Act 2010, and the Human Rights Act 1998. There is a duty of care on behalf of the NHS professionals and the Trust itself (17). There is a related right 'to be given information about the test and treatment options available to you, what they involve, and their risks and benefits.' A corresponding duty for doctors is communicating information around the treatment options, their risks and effects. As it relates to the decision of treatment, there is a duty on the side of the doctors to be involved in deciding their health care. This is set out in the NHS Constitution. The right to autonomy is further recognized in the Human Rights Act (Art.8) and in case law, where Lord Donaldson stated in *Re T (Adult) [1992] 4 All ER 649* (18):

> An adult patient who ... suffers from no mental incapacity has an absolute right to choose whether to consent to medical treatment ... This right of choice is not limited to decisions which others might regard as sensible. It exists notwithstanding the reasons for making the choice are rational, irrational, unknown or even non-existent.

6. The Potential of ADMS to Violate Patient Rights

The introduction of ADMS into healthcare is problematic because they have the potential to violate the human rights standards enumerated above. In each of the three cases below, ADMS fails to meet a reasonable standard of information that allows the patient to know and understand the treatment options before him or her. In turn, the patient is relegated to the sidelines and their autonomy is negated.

1. A patient should understand the validity of what the classifier is learning. If ADMS learns contentious artificial constructs, like phonemes in the case of speech, it is reasonable to request experiments with varied learned representations in order to interpret how the output decision may change the risk from treatment. If the experiments demonstrate a high variability in the output, for example if the decision for the patient to take a medication with painful side effects fluctuates, she should have recourse for challenging the treatment option.

2. Similarly, the patient should have access to any structures (causal or otherwise) that form the standard for AI agent task assessment. If a patient wishes to contest a decision around the causality of her brief smoking habit as a determinant in her lung cancer, she can point to the causal structure the agent learned as controversial given the presence of other more relevant genetic and lifestyle

facts.

3. And, a patient in critical condition should have access to the series of objectives that were learned in an automated decision that schedules her hospital care. In the likely case that this decision considers hospital staff, medical resources, and the condition of existing patients, it is possible the algorithm is optimizing for global utility over the well-being of a singular patient, resulting in that patient receiving care that is optimal for the whole ecosystem but not her individually. Ultimately, this should be explained to the patient so that she has the possibility of appealing the automated decision or switching her care provider.

7. Safeguards for Data-Subjects

The standard of explainability as set out in GDPR Art 22 (1) does not match the technical capabilities of explanation and (2) does not provide a clear and meaningful way for a patient to challenge a health and social care decision made by the algorithm. In order to improve the due process rights, a patient's care should meet two standards or rights surrounding care.

First, a patient should have a right to know the extent to which an algorithm is being used in their healthcare treatment. This right would reasonably allow a patient to choose not to have a have an algorithm be used in their healthcare process.

Second, the patient should also have a right to request an audit of the algorithm. That is, a right to review the supply chain of the algorithm. An automated decision will likely not be replicable without the algorithms, data, and chosen hyperparameters which comprise its 'supply chain'. An ADMS' supply chain is the "training data, test data, models, application program interfaces (APIs) and other infrastructural components" that serve as necessary components for any responsible form of auditing [24]. Without an explanation of this supply chain, the data-subject or a technical consultant, cannot challenge potential scientific flaws in deep learning classifiers. While it may not be reasonable to force a data controller to open-source all three to the general public because of intellectual property concerns, a third-party auditor should have access in order to create some type of oversight, at least until mechanisms for explainability catch up.

The Services agreement between Royal Free and Google DeepMind provides the potential for securing due process on behalf of the patients through auditing. Section 8.1 of the agreement states that: '...DeepMind shall use reasonable endeavours to develop and provide the Trust with a service to allow the Trust to obtain an accessible audit history in relation to the Data.'

In establishing these standards, patients would have clearer grounds for challenging the use of an algorithm in the provision of health and social care. These standards furthermore cement basic rights to health and autonomy found within the International Covenant on Economic, Social and Cultural Rights, and with the Human Rights Act 1998.

8. Conclusion

As ADMS proliferate, it is important that we monitor the technical gap that exists between interpretability and explainability in order to gauge the power we give automated decisions in affecting citizens, especially vulnerable populations such as medical patients. Without standards of due process that support a patient's rights within the health and social care system, we as a society risk giving algorithms more decision-making power. The existing duties established within the UK illustrate what general human right standards must be met within health and social care. However, further clarification of patient rights in relation to ADMS is required in light of the interpretability-explainability gap. These rights do not amount to a right to explainability established in existing data regulation such as the GDPR. The necessary standards that should be put into place include (a) disclosing if a decision is automated and (b) allowing for external auditing mechanisms.

Notes

The legal component of this paper is a layperson's account as both authors do not have a formal educational background in law. The authors' analysis is based on research into the areas which have been discussed in the literature and through our own reflection on the human rights issues that arise from the technical abilities.

ACKNOWLEDGMENTS. The authors would like to thank Dr. Mark Sprevak and Dr. Nehal Bhuta for providing feedback on the many drafts leading to submission.

1. Binns R (2017) Algorithmic accountability and public reason. *Philosophy & technology* pp. 1–14.
2. Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31(4):611–627.
3. Pasquale F (2015) *The black box society*. (Harvard University Press).
4. Chander A (2016) The racist algorithm. *Mich. L. Rev.* 115:1023.
5. Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.
6. Vieira A, Sehgal A (2018) How banks can better serve their customers through artificial techniques in *Digital Marketplaces Unleashed*. (Springer), pp. 311–326.
7. Molnar P, Gill L (2019) Bots at the gate: A human rights analysis of automated decision making in canada's immigration and refugee system.
8. Simmons R (2018) Big data, machine judges, and the legitimacy of the criminal justice system. *UCDL Rev.* 52:1067.
9. Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions in *Conference on Fairness, Accountability and Transparency*. pp. 134–148.
10. Amato F, et al. (2013) Artificial neural networks in medical diagnosis.
11. Esteva A, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115.
12. Dilsizian SE, Siegel EL (2014) Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16(1):441.
13. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73:1–15.
14. Holzinger A, Dehmer M, Jurisica I (2014) Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics* 15(6):I1.
15. Gunning D (2017) Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.
16. Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7(2):76–99.
17. Herring J (2016) *Medical Law and Ethics*. (Oxford University Press).
18. Service NH (2013) The nhs constitution: the nhs belongs to us all.