How Can the Law Address the Effects of Algorithmic Bias in the Healthcare Context?

Zoya Yousef

ABSTRACT

This paper examines how UK ‘hard law’ can be adapted to regulate algorithmic bias in the healthcare context. The causes of algorithmic bias are explored, setting the foundation to understand how the law can address this issue. Elements of the tort of negligence, the Equality Act 2010, and the Medical Devices Regulations 2002 are analysed, revealing the limitations of these frameworks in their application to algorithmic bias. Following this, recommendations for how the law can ensure that algorithms do not perpetuate existing biases and discriminate against patients are made. This paper acknowledges that addressing algorithmic bias will involve a combination of hard and soft law measures. However, urgent systemic change (data sharing and workplace diversity) is also needed to enable the law to address the effects of algorithmic bias most effectively in the healthcare context.

* LLB (LSE) ’23, MPhil (University of Cambridge) ’24. Interested in the intersection of AI, law, innovation, and medicine.
INTRODUCTION

The UK healthcare system is increasingly turning to algorithms to solve complex medical problems with a greater degree of accuracy and at a faster pace.\(^1\) Medical algorithms are being used in diagnostics,\(^2\) treatment recommendations,\(^3\) and resource allocation. However, this has raised concerns about the risk that medical algorithms pose in perpetuating existing biases and discriminating against patients on a larger scale than humans.\(^4\) Addressing this concern is essential so that the benefits of medical algorithms can be realised without compromising patient safety.

A formal definition of an algorithm does not exist, and its core characteristics remain a point of contention within computer science and mathematics literature.\(^5\) Introductory textbooks on algorithms have traditionally defined it as ‘a set of steps to accomplish a task’.\(^6\) This can encompass many different decision-making processes — from complex optimisation models to the trivialities of baking recipes — so the scope is narrowed for the purposes of this analysis. This paper focuses on examining medical algorithms specifically, which are defined as computer-based models that assist healthcare decisions or analyse

---

medical information.7 Machine learning (‘ML’) algorithms are an advanced set of algorithms that self-learn their parameters through a process of trial-and-error training based on large datasets.8 The use of classical algorithms, as well as those that deploy ML techniques, will be the main focus of this paper since ML is increasingly being adopted in the medical context because of its accuracy.9

The concept of ‘black box algorithms’ is of relevance to the discussion of the use of Artificial Intelligence (‘AI’) in medicine. Black box models use advanced ML techniques to make predictions, but the reason for the algorithm’s output is inscrutable or untraceable because the methodologies of these models are extremely complex and often defy human understanding.10 One such example is a ML algorithm that is used to detect skin cancer from images of skin lesions. While this model might be very accurate in identifying skin cancer, it is not clear what combination of factors the algorithm extracts from the images, and how much weight is afforded to each in making its decision. This contrasts with other models, such as algorithms based on linear and logistic regression, where humans can refer to the model's more transparent parameters to interpret what weights the model uses to make recommendations. Black box algorithms will be a recurring point of consideration throughout this paper as the opaqueness of these models poses a challenge to the traditional understanding of informed and transparent decision-making exercised in medical law.

At the end of March 2023, the UK Government published a white paper outlining its plans to regulate AI. The white paper aims to avoid introducing ‘heavy-handed and rigid’11 legislation, which could stifle innovation. Instead, it

11 Department for Science, Innovation & Technology, ‘A pro-innovation approach to AI regulation’ (CP, 815) 2
<https://assets.publishing.service.gov.uk/media/64cb71a547915a00142a91c4/a-pro-
seeks instead to propose a more flexible approach by empowering existing regulators (such as the Medicines and Healthcare products Regulatory Agency) to tailor context-specific rules in alignment with specific AI principles. The UK’s resistance to introducing new AI legislation stands in marked contrast with the EU, which is in its later stages of passing the EU AI Act — the first legal framework to regulate AI. There is great value in the UK’s light-touch and pro-innovation approach to the regulation of algorithms in the medical context. However, this paper argues that it should be supplemented by clear amendments to existing ex-ante and ex-post laws to ensure that action can be taken to remedy and prevent harm caused by algorithmic bias in healthcare. Currently, it is not clear how existing hard law provisions can protect patients against algorithmic bias, as the novelty of this technology does not fit within the established legal and medical ethics frameworks. AI’s multifaceted capabilities will mean that a ‘one-size-fits-all’ approach to legislation could overregulate and underregulate in different medical applications. For example, the use of AI in disease diagnosis will have very different implications to its use in hospital administration. Furthermore, AI’s inherent unpredictability makes it challenging for current laws to respond to its unknown risks and potential uses. Finally, the majority of medical algorithms and data sets are developed by private actors and protected by a web of intellectual property laws which prevent regulators from anticipating its dangers.

---

12 ibid.
As a result, the public lacks trust that they will be protected from the injuries caused by algorithmic bias. There is also a concern that organisations that develop and use algorithms will escape liability for the harms caused by bias. The lack of legally binding measures in the UK’s approach also prevents bias-mitigating approaches from being consistently integrated into the design of medical algorithms which is vital to ensure patient safety and harness innovation in healthcare. To date, the regulatory literature focuses on proposals and improvements to soft laws governing AI. These proposals alone are insufficient to address the novelties of algorithmic bias with effective force. Thus, a gap remains in research to explore how the UK’s hard law instruments can adequately regulate algorithmic biases in the healthcare context.

It is not within the scope of this paper to design a new hard legal framework to regulate algorithmic bias in the healthcare sector, nor would this be desirable since the UK has ruled out the possibility of a new AI regulator. Instead, this paper seeks to explore the problems that existing hard law provisions will face in their response to algorithmic bias and suggest recommendations on how the law might be adapted to protect patients and improve medical care. It is also cautioned that hard laws can only do so much to address algorithmic bias,

---

17 Thomas P, Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le, ‘Trust and medical AI: the challenges we face and the expertise needed to overcome them’ 28 Journal of the American Medical Informatics Association 891.
21 Department for Science, Innovation & Technology (n 11).
and simultaneous systemic change is necessary to enable the law to address the effects of algorithmic bias in its entirety. To date, there has been little dialogue between the legal, AI, and medical communities which has resulted in a lack of practical and informed normative standards to address the unique challenges presented by algorithmic bias in the healthcare setting. Therefore, this paper aims to propose recommendations that are technically feasible, reflective of bioethical principles, and fit sensibly within the boundaries of the law.

In the first section, the paper sets out a definition of algorithmic bias and explains the difference (and relationship) between algorithmic bias and data bias. The author will then begin to analyse how existing legal measures will be challenged by algorithmic bias and offer recommendations to ensure that the law does not exacerbate healthcare inequalities. In the second section, the ex-post legal protections granted by the common law of negligence and the Equality Act 2010 will be discussed. Within negligence law, the challenges posed by applying the reasonable person standard of care and foreseeability criterion will be considered. This leads the author to suggest that the standard of care should be amended to the level of a ‘reasonable algorithm’ where medical algorithms have caused injury to a patient as a result of bias. Where black box systems might enable defendants to avoid liability, the defendant must be able to provide evidence that they have maximised the extent to which the algorithm may be ‘causally interpretable’. This would be necessary for harm to be deemed unforeseeable. In relation to the Equality Act 2010, the challenges involved in establishing and proving discrimination will be discussed. It will be proposed that the ‘error rate parity’ measure of algorithmic fairness should be adopted in practice as a minimum standard of statistical evidence for discrimination.

The third section looks at ex-ante legal protections and considers how the Medical Device Regulations 2002 fail to regulate general-purpose AI. To address this, the author proposes an amendment to the statutory wording from ‘intended purpose’ to a ‘reasonably foreseeable purpose’. This would ensure that biases that emerge from the medical use of general-purpose AI are addressed and mitigated. In the final section, the paper acknowledges the limits of hard law provisions in relation to the specific challenges of algorithmic bias and the value of preventative measures. This leads the author to discuss urgent systemic changes (including greater data sharing and AI workforce diversity) which are required
alongside hard and soft law measures for a rounded approach to address the effects of algorithmic bias within healthcare.

ALGORITHMIC BIAS AND ITS CAUSES

There are many definitions of algorithmic bias.\(^{22}\) However, for the purposes of this article, algorithmic bias will be broadly defined as involving ‘systematic errors and erroneous assumptions in computer-based models that generate unfair outputs’.\(^{23}\) To tackle the question of how the law can address the algorithmic bias that arises in healthcare, it is important to distinguish between two interrelated causes of algorithmic bias that arise at different stages of the development process: algorithmic bias and data bias.\(^{24}\) Disagreement over the causes of algorithmic bias has been prevalent in academic debate.\(^{25}\) While there have been some emerging lines of argument that algorithmic bias does not exist (only data bias),\(^{26}\) this view is too simplistic. As Suresh and Guttag point out, focusing only on data as a driver of algorithmic bias treats data as an independent variable


dissociated from the process that produced it.\textsuperscript{27} Thus, a closer and more holistic inspection of the algorithmic development process is necessary to reveal how deep-rooted systemic norms and human choices within the value chain of medical algorithms can cause harmful biases.

This paper will first discuss algorithmic bias. This encompasses \textit{ML} bias where the algorithm self-learns biases over time, and \textit{coding} bias in the way that the algorithm is trained and designed. The second driver of algorithmic bias is data-driven bias, which is caused by (a) non-representative datasets and (b) biased samples. These two strands of algorithmic bias reveal that algorithms cannot make their own bias \textit{per se} — even though the media often propagates this misconception.\textsuperscript{28} Instead, algorithmic bias stems from how algorithms are created and trained on \textit{human} data. This may lead one to consider that algorithms are not any more biased than humans. However, algorithms are peculiar in their inherent potential to extend existing human biases beyond isolated instances of human judgement within specific contexts and geographies. For example, if a single doctor were biased in their medical practice, their decisions would impact only a few unfortunate patients. On the other hand, a diagnostic algorithm powered by biased medical data could adversely impact a much larger pool of patients in hospitals around the world. Furthermore, it might be easier for humans to interrogate biases in their decision making by assessing the reasoning used to make judgements. Algorithms will simply make decisions based on training materials without an attempt to even discern the reliability of the data. Therefore,

while algorithms only replicate human biases, the scale and speed with which these biases can be translated without oversight or interrogation presents a worrying possibility that must be prevented.

As such, a legal framework should be developed to ensure that patients are safeguarded against the undue amplification of human biases, as well as to help mitigate the potential of new biases evolving in the algorithmic development process. Unlike human decision-making, some algorithms can be subjected to comprehensive scrutiny, which makes it possible to trace and rectify biases in programming and training data. There have even been some attempts to reduce the opacity of black box algorithms through ‘explainable AI’ techniques designed to prompt algorithms to explain their decisions.²⁹ Therefore, evolving technological developments could work alongside legal developments to establish better transparency standards to mitigate existing forms of biases that have been overlooked in human decision-making processes.³⁰ Reflecting a better understanding of where biases emerge in the development of medical algorithms, a hard law framework can target the relevant causes and lead to a more equitable healthcare system.³¹

Algorithmic Bias

Machine Learning Bias

ML bias arises when an algorithm is expected to communicate one thing (the ideal target) but instead communicates another using proxy features (the actual target).³² These features are used by an algorithm as a surrogate to predict its goal, forming a divergence between the ideal target and the actual target. This results in algorithmic bias.

³¹ ibid.
An example of ML bias can be illustrated by Obermeyer et al’s study of an algorithm that was used across US hospitals to predict a patient’s risk of future cardiomyopathy (heart disease). The ideal target of this algorithm would be patients with the highest risk of developing cardiomyopathy based on accurate risk factors and biomarkers. However, the algorithm based its prediction on the patient’s cost of medical care. The algorithm assumed that there was a meaningful correlation between high healthcare costs and future cardiomyopathy risk. However, the algorithm did not account for inequitable access to healthcare, which means that while some patients may incur fewer health costs than others, they nevertheless require the same care. The wedge between what the algorithm was meant to predict, and what it was actually predicting, led to black patients being deprioritised for treatment despite being equally (if not more) in need than white patients. Remedying this algorithmic bias would have increased the percentage of black patients receiving necessary additional care from 17.7% to 46.5%. The health implications of the bias are significant: early intervention and identification of cardiomyopathy can reduce mortality rates and disease progression significantly. Since the algorithm used proxy features to distort the prediction basis of the ideal target, the outcome disproportionately prioritised white patients at the expense of other patient groups. Fortunately, the ML bias present in this algorithm could be detected through a simple regression analysis which showed that by taking one black and one white patient with the same health risk, the algorithm would be more likely to recommend extra care for the white patient than the black patient. However, interrogation is not always that simple. Black box systems can create complex inferential links and base their decisions

35 Obermeyer et al (n 32) 450.
37 Emily Bembeneck, Rebecca Nissan, and Ziad Obermeyer, ‘To Stop Algorithmic Bias, We First Have to Define It’ (Brookings, 21 October 2021) para 13 <https://www.brookings.edu/articles/to-stop-algorithmic-bias-we-first-have-to-define-it/> accessed 3 February 2024.
on hidden proxy features that humans might not be able to understand or have the resources to investigate.  

*Coding Bias*

How a developer defines an algorithm’s features, problems, and success will also contribute to the presence of bias within models. Algorithms are often designed to tackle what their developers consider to be the most pertinent problem, in the manner they deem most appropriate. This can be based on the developer’s preconceptions of the correct answer without considering the perspective of the intended patients using the algorithm.

The capacity to undertake health-related AI projects is concentrated within high-income countries, such as the US, the UK, and China, inadvertently skewing healthcare demands in favour of these populations. For example, sickle cell disease (common among patients of African descent) is allocated thirty times less research than cystic fibrosis (concentrated amongst patients from a white ethnicity) despite it affecting much fewer people in England. The sociodemographic composition of those who work in algorithmic research and development is deeply unrepresentative and does not reflect patient populations: 75% of global professionals with AI skills are male and less than 1 or 2% of employees in technical roles at Google are black. While this data is not specific

---

40 Davide Cirillo and others, ‘Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare’ (2020) 81 Nature Journal 3
<https://www.nature.com/articles/s41746-020-0288-5> accessed 5 February 2024.
43 Sue Duke, ‘Will AI make the gender gap in the workplace harder to close?’ (World Economic Forum, 21 December 2022) para 4
44 Nico Grant, ‘Very Lonely.’ The Unsettling Hum of Silicon Valley’s Failure to Hire More Black Workers’ (Bloomberg, 8 June 2018)
to employees developing medical algorithms, the largest AI developments being made are within research institutions of unrepresentative technology companies — such as Google’s DeepMind and Google Health. Therefore, it is likely that many algorithms, by design, do not account for the influences of genetics and sociodemographic factors on differences in health profiles among underrepresented individuals.\textsuperscript{45} As an illustration, a developer creating an algorithm to detect cardiovascular disease risk may base the target variable on chest pain. While this is a strong signifier for men and women generally, it ignores symptoms that are also primarily present in women: hypertension, unusual fatigue, and shortness of breath.\textsuperscript{46} Failure to account for these differences creates biased predictions of cardiovascular risk for women and discriminatory health outcomes.\textsuperscript{47} Identifying coding biases in algorithms is challenging since they may reflect deeply rooted societal prejudices that may be subtly guised in everyday life (such as microaggressions, stereotypes or implicit biases) but amplified in the aggregation of large datasets by algorithms.

**Data-Driven Bias**

### Sample Bias

Sample bias occurs where data is derived from systems that are biased in terms of systemic access and treatment of marginalised groups. This means that data samples used to train algorithms reflect existing societal biases within healthcare which are replicated in the outputs of algorithms. Sample bias can exacerbate health disparities for certain populations since those who are already

\textsuperscript{45} David Cirillo, ‘Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare’ (2020) 3 npj digital medicine 1
poorly served by healthcare systems are further misdiagnosed or mistreated at a
greater scale and frequency.

Epic System Inc is a large private electronic health record company in
the US, and its clinical prediction tool, ‘the Deterioration Index’ reflects how
sample bias influences algorithms. Epic developed a predictive ‘no-show’ model
through which hospitals could double book patients deemed a high ‘no-show’ risk
to reduce the waste of resources and maximise clinical bookings. If both patients
showed up for their appointments, they would each receive a shorter
appointment. The algorithm used information from patients’ electronic health
records, including ethnicity and socioeconomic class to generate predictive ‘no-
show’ values. In doing so, the model failed to account for the barriers many ethnic
minority patients experience in accessing medical care. Therefore, Epic’s
algorithm was more likely to double book ethnic minority patients without
considering how the lack of cultural inclusivity, location, communication
practices, and discriminatory attitudes inherent in the healthcare system prevent
these patients from accessing medical care. Double booking ethnic minority
patients further perpetuated the existing biases that restricted patient access and
widened the pre-existing racial inequalities in healthcare.

It may be argued that a possible solution to this issue would be to require
developers to use data sets that do not reflect social inequalities. However, such
data sets are not readily accessible or do not exist. Data from the National
Institute for Health and Care Research has revealed that areas in the UK with the
highest burden of disease are also the areas which have the lowest number of

48 Sara Murray, Robert Wachter, and Russel Cucina, ‘Discrimination by Artificial
Intelligence in a Commercial Electronic Health Record – a Case Study’ (Health Affairs, 31
Jan 2020) <https://www.healthaffairs.org/content/forefront/discrimination-artificial-
intelligence-commercial-electronic-health-record-case-study> accessed 9 February 2023;
Vishal Khetpal and Nishant Shah, ‘How a largely untested AI algorithm crept into
hundreds of hospitals’ (Fast Company, 28th May 2018)
<https://www.fastcompany.com/90641343/epic-deterioration-index-algorithm-
49 Ala Szczepura, ‘Access to health care for ethnic minority populations’ (2005) 81
Postgrad Med Journal 144.
The Effects of Algorithmic Bias in the Healthcare Context

Vol. IX

patients participating in clinical research. Furthermore, datasets with valuable medical information are often shielded by strong intellectual property rights which privilege access to the corporations that can afford to pay and navigate the complex restrictions on data usage.

Non-Representative Data

Non-representative training data comprises overly homogenous datasets that do not adequately reflect the intended or likely patient population that will be treated using the algorithm. The prevalence and incidence of diseases and their risk factors often vary significantly by population group, especially where conditions are linked to a genetic predisposition. Therefore, if datasets do not adequately account for populations at particular risk, prediction models that are used to train algorithms may have a lower sensitivity for these populations and systematically under-detect the target condition.

MIMIC-III is a large publicly available database used for the development of algorithms in adult critical care. Yet, the data from MIMIC-III is significantly biased towards white and socio-economically advantaged populations. This means that an algorithm trained on this data will better optimise successful outputs for well-represented individuals, as opposed to less represented populations, which include marginalised and vulnerable patient groups. A prominent example of this includes dermatology algorithms used to detect melanoma. Daneshjou et al’s study revealed that three of the most popular dermatology algorithms performed substantially less accurate classifications of

52 Henderson, Flood, and Scassa (n 24) 481.
55 Henderson, Flood, and Scassa (n 24) 482.
lesions on dark skin tones.\textsuperscript{56} The datasets on which these algorithms were trained included just 5-10\% of skin lesion samples of black patients which exhibit different characteristics to those of white patients.\textsuperscript{57} This has significant effects on black patients who have the highest mortality rate from melanoma, with an estimated five-year survival rate of only 67\%, compared to 92\% for white patients.\textsuperscript{58} Consequently, unrepresentative training data can cause misdiagnoses, hindering early intervention and treatment and causing preventable deaths for patient populations.

The underrepresentation of certain patient groups in health data can be attributed to the data’s source. A large source of health data is from clinical trials which often exclude the elderly,\textsuperscript{59} patients with comorbidities, and pregnant women.\textsuperscript{60} A recent review paper analysed 224 random controlled heart failure trials between 2000-2020 and found that women made up only 28\% of participants, an unsatisfactory result given that heart disease is the second leading cause of death for women in the UK.\textsuperscript{61} The lack of representation is partly due to an attempt from companies to limit the extra costs and expenses that accompany more diverse clinical trial participants.\textsuperscript{62} The impact of the menstrual cycle,

\textsuperscript{56} Roxana Daneshjou and others, ‘Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set’ (2022) 8 Science Advances 31.
\textsuperscript{57} ibid.
\textsuperscript{60} Rieke van der Graaf and others, ‘Fair inclusion of pregnant women in clinical trials: an integrated scientific and ethical approach’ (2018) 19 Trials 78.
\textsuperscript{61} Magdalene Au and others, ‘A Systemic Review of Sex-Specific Reporting in Heart Failure Clinical Trials: Trial Flow and Results’ (2022) 1 Journal of American Cardiology Advance.
biological differences, and external factors cause varying responses in drug intervention which requires further analyses and explanation post-trial.⁶³ In addition, the lack of trust that historically mistreated patients feel towards medical research means that further effort is required to recruit from these groups.⁶⁴

Data is also often derived from technological sources such as wearable devices and mobile health applications. However, due to the digital divide, technology is not equally accessible to all patient groups. This is due to a lack of access to network resources, cultural barriers, and communication discrepancies.⁶⁵ Mitchell et al’s study found that older racial and ethnic minorities were the least likely to use health-related technologies.⁶⁶ If algorithms are trained on data derived from wearable technology, the datasets will likely be skewed in favour of educated, white, young adults with an annual household income greater than US $75,000.⁶⁷ The Open Wearables Initiative is used by Novartis, Merck, and Eli Lilly who freely share datasets from wearable health applications to develop, validate, and test new algorithms.⁶⁸ However, given that there is no information about the initiative promoting the diversity of participants, it is likely that these algorithms will be based on data that excludes lower-income groups, the elderly, and racial minorities. Therefore, data-driven bias results from non-representative data sources which cause the accuracy of algorithms to be in favour

of privileged patient groups, and disadvantages those traditionally underrepresented in medical research and technology.

**EX-POST LEGAL PROTECTIONS: NEGLIGENCE AND DISCRIMINATION**

Suffering harm from a biased medical algorithm could ground a claim in the common law of negligence or the Equality Act 2010. This section will be dedicated to the deficiencies of the common law tort of negligence in relation to the standard of care and foreseeability principles. Recommendations to ensure that the law does not unfairly leave parties uncompensated for losses resulting from injury will be discussed. This will be followed by an analysis of the challenges of establishing and proving discrimination under the Equality Act 2010. The paper proposes a new minimum standard of evidence for the law to better recognise instances of algorithmic discrimination.

The author acknowledges that while the proceeding section concerns ex-post legal frameworks, these only constitute reactive ‘wait and see’ protections. Beyond their deterrence effects and ability to compensate, the harms from algorithmic bias are likely to have already occurred to injured parties. In addition to this, they depend on a potential claimant’s willingness to pursue a case. It is likely that those who fall victim to algorithmic bias will also be the same individuals who will have the greatest difficulty to legally challenge discrimination due to a lack of knowledge and resources. Nevertheless, ex-post legal protections are a vital discussion point in determining the role of the law in addressing algorithmic bias within the healthcare context because of the potential strength of protection that these measures can afford. In addition, hard law instruments are an indication of the legislature’s commitment to tackling algorithmic bias. Later sections of this paper will discuss ex-ante legal protections and systemic

---

70 Henderson, Flood, and Scassa (n 24) 490.
changes that are also required for the law to comprehensively protect patients from the effects of algorithmic bias in the healthcare context.

The Common Law Tort of Negligence

Medical algorithms seek to identify patterns in data that are inconceivable to humans to better recommend treatments and diagnose patients. This comes with great risk, as misdiagnosis and mistreatment by a medical algorithm could be fatal. These risks may not be particularly rare either; IBM’s Watson recommended that a cancer patient with severe bleeding be given a drug that caused the bleeding to worsen.\textsuperscript{72} The current use of medical algorithms is moving towards assisting clinicians with decisions, as opposed to autonomously interacting with patient decisions. However, it is not inconceivable that AI could replace practitioners in certain elements of patient care as medical algorithms outperform doctors in terms of diagnostic speed and competency — such as in medical imaging.\textsuperscript{73} Furthermore, there are already many AI-driven symptom checker chatbots that are routinely being used by patients at home to understand their health and medical conditions without visiting hospitals.\textsuperscript{74} Therefore, the exact degree of involvement of doctors, patients, and algorithms in the future healthcare ecosystem remains unclear. Nevertheless, this uncertainty is not a reason to avoid discussion of how the law will adapt to the possibility of medical algorithms that could independently mistreat or misdiagnose patients.

This section will be dedicated to the discussion of how elements in the UK’s negligence framework can protect patients who receive medical mistreatment because of a biased algorithm. To ground a claim in negligence, a claimant must establish a duty of care, a breach of duty, causation, and foreseeability.\textsuperscript{75} The patient must also have experienced harm from the biased

\textsuperscript{74} William Wallace and others, ‘The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review’ (2022) 5 npj Digital Medicine 118.
\textsuperscript{75} Daniele Bryden and Ian Storey, ‘Duty of care and medical negligence’ (2011) 11 Continuing Education in Anaesthesia Critical Care & Pain 124.
medical algorithm in the form of a personal injury. It is possible that patients could also have a negligence claim if they only experience psychiatric injury (and no physical injury) from a biased medical algorithm. Discussion is limited to negligence claims from personal injury as the legal framework concerning psychiatric injury will require a separate and more focused discussion in relation to medical algorithms. Thus, the proceeding section will detail negligence law’s challenges in relation to (i) the inadequacies within the duty of care requirement and, (ii) the subsequent liability loophole caused by the foreseeability criterion.

The Reasonable Person Standard of Care

Under the tort of negligence, the standard of care criterion presents difficulty to claimants who have been exposed to a biased medical algorithm and suffered an injury as a result. Following Bolam, the test for determining the standard of care for medical professionals is based on whether the action conforms to a substantial body of professional opinion. Applying this precedent to the use of algorithms in medical practice, it is necessary to consider who would be the most appropriate comparator for the standard of care requirement under tort law: other similar medical algorithms or human medical professionals.

Where the competencies and decision methods of algorithms begin to surpass human understanding, it will be challenging to find a respectable body of medical opinion to comment on whether a reasonable human medical professional would have made the same decision. There is a large time lag between the viability of a new practice and professional acceptance — some estimates

---

77 Bolam v Friern Hospital Management Committee [1957] 1 WLR 582; Wilsher v Essex Area Health Authority [1988] AC 1074; FB v Princess Alexandra Hospital NHS Trust [2017] EWCA Civ 334.
place this at around seventeen years. Thus, it will be difficult to find a responsible body of medical opinion for the early uses of medical algorithms. It is important to note, however, that the Bolam test does not favour quantity over quality; it could be foreseen that a small group of specialists in the medical field will be required to apply their expertise to new algorithms. For instance, in De Freitas, the court held that eleven doctors specialising in spinal surgery out of more than 1,000 orthopaedic and neurosurgeons in the country constituted a responsible body of opinion. This could offer some possibility of a standard of care for some early applications of medical algorithms, but it could also lead to skewed results if the group of specialists are not diverse, perpetuating existing biases in medical practice or a reluctance to support the use of medical algorithms. Thus, allowing a small group to set the standard of care does not adequately set an appropriate legal threshold or contribute to remedying algorithmic bias (it could instead perpetuate bias).

It might be said that humans can still identify a reasonable standard of care even if algorithms perform the task. For example, studies have evaluated the performance of AI models which take student medical examinations — usually outperforming the scores of doctors. In practice, however, medical algorithms are used in circumstances where their ability, as opposed to accuracy and speed, means that they can identify correlations in data in a way that exceeds human abilities. It would be very difficult to evaluate the reasonableness of an algorithm’s early prediction of the spread of cancer from medical imaging or the detection of dementia from voice recognition. These are tasks that algorithms can easily

---

80 Zoë Slote Morris, Steven Wooding, and Jonathan Grant, ‘The answer is 17 years, what is the question: understanding time lags in translational research’ (2011) 104 Journal of the Royal Society of Medicine 510.
82 Haixia Wang and others, ‘Diversity in people’s reluctance to use medical artificial intelligence: identifying subgroups through latent profile analysis’ (2022) 5 Frontiers in Artificial Intelligence.
perform with higher competency than expert oncologists or neurologists.\textsuperscript{84} Thus, the use of a human standard of care will become increasingly challenging as medical algorithms become more complex and there will likely be a stage where the ability of medical professionals to comment on the responsibility of the algorithm’s output will diminish.\textsuperscript{85}

It will be extremely difficult for the medical community to gain a thorough and quick expertise in the biases of medical algorithms\textsuperscript{86} to satisfy the \textit{Bolam} test. In any case, the outcome will remain subject to the \textit{Bolitho} test.\textsuperscript{87} \textit{Bolitho} requires that the medical opinion is subject to the requirement of logic and the balancing of comparative risks and benefits to reach a defensible conclusion. Therefore, the medical professionals defending an algorithm’s biased recommendation would need to provide an explanation as to why the algorithm was not negligent. The opacity of medical algorithms, especially black box models, will make it difficult for professionals to adequately justify their opinions because the algorithm will extrapolate data in a way that is beyond human understanding. Black box systems can even preclude the original programmers from testifying with certainty as to the algorithm’s rationale which caused the alleged biased output.\textsuperscript{88} Therefore, the standard of care criterion raises difficulties in deciding how to evaluate medical algorithms where they are performing tasks at a higher ability than humans. This will be particularly problematic where there are inadequate peer algorithms to determine the standard to which algorithms should act.

\textbf{The Foreseeability Loophole}

By inserting a layer of inscrutable, unintuitive, and statistically derived code in between a human decisionmaker and the


\textsuperscript{86} Liam McCoy and others, ‘What do medical students actually need to know about artificial intelligence?’ (2020) 86 npj Digital Medicine 3.

\textsuperscript{87} \textit{Bolitho v City and Hackney Health Authority} [1998] AC 232.

consequences of her decisions, AI disrupts our typical understanding of responsibility for choices gone wrong.\textsuperscript{89}

If a claimant can establish that the defendant owed her (i) a duty of care, (ii) which has been breached, and that (iii) the damage was caused by the defendant, the courts will consider foreseeability. Foreseeability links the defendant’s conduct and its risk to the claimant and considers whether the harm caused was too remote to justify the claimant receiving compensation.\textsuperscript{90} It is a basic principle of tort law that ‘a defendant is responsible for and only for such harm as he could reasonably have foreseen and prevented.’\textsuperscript{91}

Medical algorithms are often based on ML software which develops iteratively based on real-world training data. This feature is what provides doctors with such accurate predictions and makes algorithms so useful in the healthcare context. However, this also causes a foreseeability problem as algorithms can operate in ways that are unforeseeable and uncontrollable by the original programmers. This, in turn, gives rise to a liability gap under the traditional tort law framework.\textsuperscript{92} The foreseeability loophole would mean that patients’ claims against biased medical algorithms would fail because the harm caused was of an inherently unforeseeable nature. This article will not go into depth about who should be accountable in the value chain of a biased algorithm because much attention has been devoted to this topic elsewhere.\textsuperscript{93} Furthermore, unless there is a strict liability regime to compensate for harm caused by biased medical algorithms, the foreseeability loophole is fundamental to a claimant’s case. This is true whether the algorithm itself, the medical professional, the hospital, the downstream or upstream developer, or the company will be considered liable or jointly liable. Discussion of the foreseeability loophole will be relevant to the defendant(s) in determining whether they should be responsible for compensating

\textsuperscript{89} Andrew Selbst, ‘Negligence and AI’s Human Users’ (2020) 100 Boston University Law Review 1315.
\textsuperscript{90} The Wagon Mound No 1 [1961] AC 388.
\textsuperscript{91} Tony Honoré and Herbert Lionel Adolphus Hart, Causation in the Law (2nd edition, Oxford University Press 1959) 255.
\textsuperscript{93} Helen Smith, ‘Clinical AI: opacity, accountability, responsibility and liability’ (2020) 36 AI and Society 535.
the harm and what mitigating measures they can adopt to avoid liability. Therefore, the wide responsibility gap must be addressed to ensure that injured patients can make an effective claim in tort law to compensate for harms suffered by biased medical algorithms.

Selbst argues that without interpretable or explainable algorithms, it is impossible to claim that an erroneous judgement from a black box algorithm could have been foreseen ahead of time.\textsuperscript{94} Since the outputs of these algorithms cannot be understood or interrogated, courts cannot predict the likelihood of harm as they would if a claimant was calculating the risk that a cricket ball would be hit beyond a fence,\textsuperscript{95} or that a claimant will suffer chronic fatigue syndrome following a minor car accident.\textsuperscript{96} The opacity of medical algorithms makes it difficult for judges to determine what factors the defendant (even if this is the AI developer) might have taken into consideration to determine whether the type of harm caused by the medical algorithm was reasonably foreseeable or was a negligent decision.

One solution put forward might be to mandate that medical algorithms are explainable. This would allow for information about the reasoning of the output recommendation to be understood by stakeholders affected by the medical algorithm to better assess the remoteness of the harm.\textsuperscript{97} Therefore, consideration can be given to whether the defendant should be held liable for the harm caused by a medical algorithm (for example, if it was based on a very obvious biased correlation) or if it would be unfair to do this since the harm was too remote. However, medical algorithms are relied on for their ability to aggregate large data sets so it would be against their very purpose if their implementation was limited to cases within human comprehension. Therefore, under the tort framework, the liability gap is likely to persist. The unforeseeable nature of medical algorithms

\textsuperscript{94} Andrew Selbst, ‘Negligence and AI’s Human Users’ 100 Boston University Law Review 1315.
\textsuperscript{95} Bolton v Stone [1951] AC 850.
\textsuperscript{96} Page v Smith [1996] AC 155.
\textsuperscript{97} Tyler J Bradshaw and others, ‘Artificial Intelligence Algorithms Need to Be Explainable- or Do They?’ (2023) 64 Journal of Nuclear Medicine 6; Carlo Combi and others, ‘A Manifesto on Explainability for Artificial Intelligence in Medicine’ (2022) 133 Artificial Intelligence in Medicine.
could easily shield defendants from liability and leave patients uncompensated for their injuries.

**Negligence Recommendation: The Reasonable Algorithm Standard of Care**

The great changes come from the interplay of the same forces in the legal world as in the rest of life – custom, innovation, revolution, are the phases of this process and the end-product is not certainty but growth – Justice Cardozo

It has been argued that the current framework relating to the reasonable person standard of care is inadequate. There is an incongruity in comparing the standard of care of a medical algorithm to a human medical professional, given the respective levels of technical abilities and different decision-making processes. To address this issue, this paper proposes that the negligence framework adapts to a ‘reasonable algorithm’ standard of care.98 Instead of a medical algorithm being held to the standard of a human, it will be held to the standard of care based on other hypothetical medical algorithms. This follows the logic in *Bolam* where a separate standard of care for professionals was introduced to reflect the understanding that professionals have specialised skills and knowledge. This, in turn, means that they should be held to a higher standard of care. It is also a necessary improvement to ensure that the safety and accuracy of algorithmic decisions are pushed to exceed those of a reasonable human.99 Establishing a clear reasonable algorithm standard of care in law will also motivate developers to use more representative data sets and interrogate algorithmic processes for the presence of bias. It is necessary that laws are developed to ensure that patients can be compensated for *ex post* harm. At the same time, clarity enables developers to consider legal and fairness principles in the development of medical algorithms to mitigate harms and prevent them from occurring beforehand.

---


The issue with the reasonable algorithm standard of care is that there does not yet exist a comparable standard for reasonableness. In the usual reasonable human standard of care criteria, where there is a lack of equivalent reference points to determine whether a professional acted reasonably, judges compensate by drawing on their own personal perspectives or intuitions.\(^{100}\) The ‘reasonable algorithm’ criterion presents an obstacle, given that humans will lack an understanding of how to predict how other algorithms would act in similar circumstances.\(^{101}\) While there are thousands of companies and start-ups worldwide developing medical algorithms,\(^{102}\) the accumulated knowledge as to how other algorithms might perform in each situation is limited. Even where predictions can be made, it is uncertain whether these predictions can also be labelled as a ‘reasonable’ standard of care. The same biases will also persist in these algorithms since models are usually trained on the same, limited datasets. This makes comparisons for bias redundant.\(^{103}\) Furthermore, where algorithms can be compared to one another, just because another medical algorithm may perform better statistically, statistical success will not be a reliable indicator of whether there has been a breach in a certain case. This overlooks the individualised and context-dependent nature of harm which requires that the court assess the facts of the case to determine whether the defendant’s actions (or omissions) constitute causation for the claimant’s injury.

Nevertheless, the lack of comparator algorithms should not be a reason to resist implementing a reasonable algorithm standard of care. While the AI industry is still in its infancy and existing medical algorithms do not provide an adequate comparator as of yet, it is suggested that audits of algorithms could fulfil the role of a comparator until more appropriate comparators are developed. These can use best practices and industry ideals to determine reasonableness of


\(^{103}\) Jesutofunmi Omiye and others, ‘Large language models propagate race-based medicine’ (20 October 2023) 6 npj Digital Medicine.
medical predictions. Algorithmic audits will allow judges to evaluate how the algorithm may have performed in the same scenario if different inputs, parameters, or patient groups were considered. This could point to whether the algorithm’s decisions were based on errors or biases which would not have been replicated in other medical algorithms. It would also ensure that the standard of care is not based on overgeneralised statistical measures of success, as algorithms would be tested contextually and in the specific circumstances. Many medical algorithmic audits have been proposed which interrogate errors and unexpected results in algorithms before and during real-world deployment. Liu et al proposed requirements for the medical algorithm auditing process, including an exploratory error analysis, subgroup testing, and adversarial testing. The exploratory error analysis investigates the commonality and unexpectedness of the medical bias. The subgroup testing interrogates the possibility of error features that might be patient or task specific, and adversarial testing attempts to understand the prevalence and source of errors in the worst sub-groups. This is only one example of a medical algorithm auditing framework, but it demonstrates the comprehensive evaluation and thoughtful interrogation that is undertaken to assess the standard of the algorithm’s performance. The advantage of this proposed algorithmic audit is that it addresses the unique medical context in which input patient or disease data could produce vastly different outcomes.

The use of medical algorithmic audits would connect to the UK Government’s push for new regulatory sandboxes to be used for new AI models, by facilitating the real-life testing of medical algorithms to determine compliance with uncertain legal and medical standards. Therefore, in the early days of the use of medical algorithms it seemed sensible to move from a human standard of care to the standard of a reasonable algorithm based on medical algorithmic audits. If properly designed and conducted, audits will provide

106 Ibid.
107 Ibid.
objective and quantifiable evidence to help determine the standard of a reasonable algorithm in litigation. Furthermore, with an enhanced focus on developing audits, medical algorithms will be developed with bias mitigation strategies by design to prevent the chances of negligent decisions occurring within healthcare decisions.

Aside from the practical considerations of establishing a reasonable algorithmic standard of care based on audits, additional issues warrant discussion. One of the primary objectives of tort law's negligence framework is to influence and deter potential wrongdoers by outlining clear expectations and guidelines for compliance. Without specific guidance regarding which rules or standards take precedence in particular circumstances, an algorithm is unlikely to be able to utilise the vague notion of reasonableness to adjust its decision-making. It is possible that an algorithm could eventually be trained on enough judicial decisions and medical ethics guidelines to learn what constitutes reasonable decision-making.\footnote{Alex Tamkin and others, ‘Evaluating and Mitigating Discrimination in Language Model Decisions’ (Anthropic, 6 December 2023) <https://www-files.anthropic.com/production/images/Anthropic_DiscriminationEval.pdf?dm=1701894346> accessed 12 December 2023.} This would mean that an algorithm could be programmed to consider the potential consequences of an ‘unreasonableness’ finding as part of the parameters that it weighs before reaching its decision.\footnote{Karni Chagal-Feferkorn, ‘How Can I Tell if My Algorithm Was Reasonable?’ (2021) 27 Michigan Technology Review 213, 227.} For example, if a medical algorithm makes a biased decision based on a lack of representative data about skin cancer in dark-skinned patients, the parameters of imaging findings for dark-skinned patients may be adjusted as a result of the unreasonable finding to ensure that the biased outputs are interrogated.\footnote{ibid.} This process of self-learning is already used in ML medical algorithms. It would, however, take several decades to gather enough data for algorithms to independently understand the judicially developed standard of a reasonable algorithm and adapt to the considerations mandated in medical ethics guidelines and best practices. Therefore, the deterrence of medical algorithms towards making biased decisions would be a slow and gradual process. While this might be useful to implement when more litigation involving medical algorithms becomes more frequent, further guidance for the courts about approaching initial cases is required.
There are also conceptual criticisms for the introduction of a ‘reasonable algorithm’ criterion. The ‘Homunculus Fallacy’ holds that algorithms and humans should be indistinguishable under the law since algorithms act according to how they are programmed to act by humans.\textsuperscript{112} Therefore, creating an independent standard of care for algorithms would be illogical. However, this puts too much weight on the human developers and human data that train the algorithms without acknowledging the autonomous nature in which the information is processed by the algorithm. This argument reveals that analysing an algorithm’s own reasonableness separately from the standard of care of a human remains reasonable. Once deployed, an algorithm can continually refine its parameters through iterative processes, adapting to new health data and evolving circumstances without direct human intervention. This iterative learning process creates a dynamic system where the algorithm can make decisions that may deviate from original human intent. Consequently, the 'reasonable algorithm' criterion seeks to acknowledge and scrutinise this evolving autonomy, recognising that as a medical algorithm processes information and refines its outputs, it operates with a degree of independence from its human creators. Therefore, the law should recognise that an algorithm’s biased recommendations are independent from its human influence which should be reflected in a separate standard of care to ensure that algorithms do not escape liability for biased medical decisions.

Arguably, a determination that an algorithm acted unreasonably under the proposed standard of care would, nevertheless, be meaningless in terms of the goals of tort law; an algorithm itself cannot pay damages nor can it be financially deterred from making biased recommendations in the future.\textsuperscript{113} The question of whether algorithms could bear a legal personality has attracted much debate in the literature,\textsuperscript{114} but considerations of this issue are beyond the scope of this paper. Nevertheless, it is argued that algorithms do not need to bear the consequences of their unreasonableness themselves for the aims of tort law to be met. Rather, a ‘reasonable algorithm’ analysis could be used as a tool for

\textsuperscript{113} Chagal-Feerkorn (n 110) 222.
\textsuperscript{114} Jasper Doomen, ‘The Artificial Intelligence Entity as a Legal Person’ (2023) 32 Information and Communications Technology Law 3.
determining the liability of the manufacturer of the system or a medical practitioner using the system.\textsuperscript{115} This would be similar to when the court examines the behaviour of an employee to determine whether the employer is vicariously liable for their actions.\textsuperscript{116} While the value chain of medical algorithms is extremely complex and comprises many different actors (as explored in the third section),\textsuperscript{117} the main tension exists broadly between doctors and manufacturers. In this instance, medical professionals could be held liable if they feed an algorithm false or misleading data or rely on an algorithmic output when it was obvious that it would cause the patient harm. However, in less clear cases, manufacturers are probably in the best position to absorb algorithmic liability. It is likely that manufacturers will purchase insurance for harms caused by medical algorithms, and if algorithms become very accurate, insurance is likely to be an affordable investment which can be easily offset by licensing profits.\textsuperscript{118} Therefore, a new standard of care based on the ‘reasonable algorithm’ using medical audits would be the most suitable option to determine how the courts should establish whether there has been a breach of the standard of care. This fits within existing research\textsuperscript{119} which is already designing algorithmic audits for medical use cases which can be adapted to fit within the legal framework to enhance medical practice and ensure patient safety.

**Negligence Recommendation: Reasonable Foreseeability and Casual Interpretability**

Cutting-edge applications of algorithms are often black box systems.\textsuperscript{120} This makes it difficult to hold defendants accountable for injuries caused by biases that could not have been foreseen within the opaque decision-making process. To prevent defendants escaping liability for black box algorithms that cause harm,

\textsuperscript{115} Chagal-Feferkorn (n 110) 213.  
\textsuperscript{116} *Woodland v Essex CC* [2013] UKSC 66.  
\textsuperscript{118} Emiliano Marchisio, ‘In support of “no-fault” civil liability rules for artificial intelligence’ (2021) 1 SN Social Sciences 54.  
\textsuperscript{119} Liu and others (n 105) 384.  
it is proposed that for harm to be unforeseeable, the defendant must be able to provide evidence that they have maximised the extent to which the algorithm may be causally interpretable.

The causal interpretation of an algorithm requires an explanation as to why the model’s input yielded a certain output.\textsuperscript{121} For example, if an algorithm is used to predict whether a patient has a certain disease based on their medical history, causal interpretability would require understanding the causal mechanisms that weigh specific medical conditions or risk factors in the diagnostic output. This must be separated from explainability which focuses on how exactly the model works,\textsuperscript{122} as this would unduly disadvantage the developers of cutting-edge algorithms for unforeseeable injuries.\textsuperscript{123} Under the proposed foreseeability requirement, it is only necessary that defendants attempt to demonstrate causal interpretability to determine whether the harm was foreseeable and could have been mitigated in the design process.

The requirement of causal interpretability might involve active and continuous monitoring of the performance and safety of algorithms to address biases, errors, and vulnerabilities in the model. The author suggests that the determination of whether a developer has fulfilled the requirements to determine that a medical algorithm is causally interpretable is decided on a case-by-case basis. This is important because the evidence available relating to the competencies of interpretable algorithms are constantly evolving and depend on different models of algorithms (ML, black box, decision tree). This criterion is intentionally flexible, but the judge could be guided by: (a) whether the developer has provided comprehensive documentation that explains the variables and features of the algorithm; (b) any detailed attempts to simplify the model to enhance interpretability without needing to sacrifice performance; (c) methods shown for


\textsuperscript{122} ibid.

identifying and highlighting features that significantly contribute to the model’s outputs; and (d) input and feedback from medical experts to understand the causes of output decisions. Without such proof that a defendant has attempted to gauge the foreseeability of harm from the causal interpretation of algorithms, the harm caused can be deemed reasonably foreseeable.

The casual interpretability requirement reflects the court’s normative perspective with respect to foreseeability, as it focuses on what the defendant ought to have foreseen (objective) as opposed to whether they did or did not (subjective).\textsuperscript{124} The law should encourage the developers of algorithms to make an active and conscious effort to pre-emptively foresee when biases and errors have influenced decisions and prevent these from materialising into medical decisions.

Requiring that the defendant should provide evidence that they have fulfilled the requirement’s causal interpretation holds the appropriate balance between reducing the risk of harm that an algorithm could exert and stifling innovation. Since complex models, such as black boxes, provide a higher prediction accuracy than more interpretable systems, requiring defendants to always produce causally interpretable algorithms would be against the advancement of cutting-edge medical algorithms. Nevertheless, it is important that the foreseeability requirement at least encourages causal interpretability so that developers can prevent the unintended materialisation of algorithmic biases that could harm patients. Without such a requirement, algorithms could be made overly complex and black box just to evade liability. Therefore, the law can close the foreseeability loophole by requiring developers to attempt to make medical algorithms causally interpretable so that injury caused by a biased algorithm can be better foreseen.

**Equality Act 2010**

A person subject to discrimination from algorithmic bias in the healthcare context may also launch a complaint under the Equality Act 2010. Section 13 grants protection from discrimination on the protected grounds of ‘sex, race, colour, language, religion, political or other opinion, natural or social

\textsuperscript{124} Fardon v Harcourt-Rivington [1932] All ER Rep 81.
The discrimination framework considers both direct and indirect discrimination. Direct discrimination refers to the adverse treatment based \textit{explicitly} on a protected attribute, such as sexual orientation or gender. For example, Amazon’s recruitment algorithm that penalised CVs which included the word ‘women’s’ could be pursued under a direct discrimination claim since it used gender as an explicit proxy feature.\(^{127}\) Indirect discrimination does not require that rules and practices are specifically directed against groups of a protected characteristic, so long as they have the same effect of putting them at a disadvantage.\(^{128}\) The algorithm used in US hospitals which required black patients to be much sicker than white patients to be recommended for the same care would fit under the indirect discrimination category.\(^{129}\) This is because the algorithm did not explicitly use race as a variable within the algorithm, but healthcare expenditure was used as a measure to determine the need of care. This meant that black patients who had not spent as much on healthcare were discriminated against. In the context of algorithmic bias in the medical context, indirect discrimination will be of most use since it is likely that claims will refer to the apparently neutral rules and patterns of an algorithm. This is because developers will attempt to avoid situations where protected characteristics form the sole basis of an algorithm’s recommendation.

To bring a case alleging indirect discrimination under the Equality Act 2010, a claimant must meet several evidential requirements that will establish \textit{prima facie} discrimination. Claimants will need to prove that (a) they have suffered less favourable treatment, (b) which manifests or is likely to manifest significantly

\(^{125}\) Equality Act 2010, Section 13.
\(^{126}\) Equality Act, Section 149.
\(^{128}\) \textit{Essop and others v Home Office (UK Border Agency)} [2017] UKSC 27, [1].
\(^{129}\) Obermeyer and others (n 33).
within a protected group of people, and (c) that the harm suffered is disproportionate when compared with others in a similar situation.\textsuperscript{130} Once these requirements are met, the burden of proof shifts to the alleged offender who can then either justify the contested practice or refute the claim.\textsuperscript{131} This section will first explore the difficulties of establishing algorithmic discrimination by showing that claimants will struggle to prove that they are part of a disadvantaged group and identify a legitimate comparator. It will then consider how algorithmic bias challenges the way discrimination is usually proved by requiring more reliance on (and scrutiny of) statistical evidence. It is argued that where algorithms, (and not humans) discriminate, the mechanisms to prove discrimination are ill-equipped to protect patients. This necessitates more appropriate detection methods and evidential requirements, to which it is proposed that the error rate parity statistical measure of fairness could be used as a minimum standard of evidence in algorithmic discrimination cases in the medical context.

Establishing Discrimination: Disadvantaged Group, Comparator, and Algorithmic Groups

Compared to human discrimination, algorithmic discrimination is more abstract, opaque, unintuitive, and intangible.\textsuperscript{132} These characteristics make it challenging to establish that claimants have been unfairly discriminated against because patients may not realise that they are within a disadvantaged group. It is easier to detect discrimination in a human face-to-face experience; for example, where a doctor may regularly dismiss or refuse to investigate the pain or symptoms reported by female or ethnic minority patients.\textsuperscript{133} Being able to overtly observe contrasting treatment between different patient groups would make it easier to understand when patients are being treated fairly or being discriminated

\textsuperscript{130} Equality Act, Section 149.
\textsuperscript{131} \textit{Akerman-Livingstone v Aster Communities Limited} [2015] UKSC 15.
against by human doctors.\textsuperscript{134} Integrating algorithms into medical practice, however, makes this comparative element much harder for patients.

It cannot be assumed that algorithms will discriminate in ways that mimic human biases and relate to recognisable patterns of discrimination.\textsuperscript{135} Medical algorithms are valued precisely because of their ability to find unintuitive connections and patterns between data, which could result in algorithms making decisions based on their own biases towards ‘algorithmic groups’.\textsuperscript{136} Algorithmic groups may transcend the legally protected characteristics but still suffer levels of disparity which would be deemed discriminatory if applied to the protected grounds.\textsuperscript{137} For example, an algorithm might discriminate against ‘video gamers’ by giving them lower credit scores\textsuperscript{138} or target suicidal advertisements at ‘sad teens’.\textsuperscript{139} Since these characteristics do not map onto the legally protected groups, it is less likely that they will be given protection under discrimination law or that members of such algorithmic groups will be aware that discrimination has occurred in the first place. Algorithmic groups do not have the same sense of community, common identity, or social saliency as protected groups to be able to recognise that they are part of a disadvantaged group or identify who would be a legitimate comparator.\textsuperscript{140} Furthermore, algorithmic groups do not evoke the same sense of moral wrong which is triggered by protected groups as there is a lack of related historical oppression. This may make it more challenging for such cases to be pursued with the same level of commitment.\textsuperscript{141} Therefore, algorithmic groups will introduce new patterns of discrimination. This means legislators should reassess and broaden the scope of the Equality Act 2010 to recognise these

\textsuperscript{134} Wachter (n 35).
\textsuperscript{136} Wachter (n 135) 1.
\textsuperscript{137} Wachter (n 135) 5.
\textsuperscript{140} Wachter (n 38).
\textsuperscript{141} ibid.
instances of algorithmic discrimination that are not based on intuition and membership of historically oppressed groups.

Proving Discrimination: Statistical Evidence

To prove that a medical algorithm is discriminating against a group, it would be necessary to provide convincing and relevant evidence that its members have suffered a disadvantage in comparison with the comparator group. A comparator group would consist of patients who are similar to the claimant in all relevant aspects, except for the characteristic which is the basis of the discrimination claim. A simple example of this would be if an algorithm misdiagnosed cancer patients based on their postcode. The comparator group of patients would have the same biomarkers or relevant characteristics indicating the cancer diagnosis apart from their address.

The specific evidence used to support differential treatment towards patient groups will depend on the nature and circumstances of the case. While this may seem like a large onus on the claimant, statistical evidence is often used in other discrimination cases to prove disparity in relation to claims about the gender pay gap or redundancies based on age or racial segregation. The use of statistics will be important in indirect discrimination cases to determine a correlation between protected attributes and the elements of a contested rule. For example, racial bias in the cardiomyopathy triage algorithm was identified by illustrating the stronger correlation between (a) medical expenditures by race and the algorithm’s risk score, and (b) the biomarkers of health and the algorithm’s risk score. Since algorithmic discrimination is less intuitive, and the decision-making processes are less transparent, greater reliance must be placed on statistical evidence in proving disparate treatment even if this might be time-consuming for claimants.

However, integrating complex statistical evidence practices (such as regression analyses and ML models) will be a challenge for courts which already

---

142 Asda Stores Ltd v Brierley and others [2021] UKSC 10; Seldon v Clarkson Wright and Jakes [2012] UKSC 16; Mandla (Sewa Singh) and another v Dowell Lee and others [1983] 2 AC 548.
143 Obermeyer and others (n 33).
struggle to deploy complex statistical evidence. The Royal Statistical Society comments that the use of statistical evidence and probabilistic reasoning in the UK courts represents a long history of misunderstandings and miscarriages of justice. The case of Sally Clark is a well-known example of how reliance on a flawed application of probabilistic reasoning led to a wrongful conviction. Where defendants are the stronger parties to the case (as would be the case for technology and pharmaceutical companies or health institutions that develop medical algorithms), it is foreseeable that defendants will be able to manipulate statistics to present a more convincing case to avoid liability. For example, a recent Office for National Statistics report reveals that statistics relied on by the Government about the risk of teachers contracting COVID-19 in the second national lockdown was based on flawed and misleading data. The Government was able to justify in-person teaching by distorting data which actually proved that teachers were 28% more likely to contract COVID-19 than other key workers.

Data feminism uncovers how standard practices in data science serve to reinforce existing inequalities around the world. From this perspective, data feminist scholars have criticised the presentation of statistics from powerful institutions where underlying hierarchical relations can unfold in the data used to support their ideologies and practices.

To date, the use of statistics in UK courts has been at a relatively basic level, and have only been relied on where parties have taken the time and resources to submit these. There has been no proactive encouragement on the

---


145 ibid.


149 R v Secretary of State for Employment Ex Parte Seymour Smith and others [2000] UKHL 12.
part of the judiciary to use and understand statistical reasoning within judgments. This paper argues that this will need to change as algorithmic discrimination may only be observable at a statistical level. At the same time, courts must ensure that the greater use of statistics is undertaken with a rigorous analysis of the validity and reliability of the evidence in relation to the contested discriminatory rule. Achieving this would require better collaboration between lawyers, software engineers, and statisticians to ensure that proving algorithmic discrimination is technically feasible and can be properly understood by the judiciary to not to reinforce existing structural biases.

**Discrimination Recommendation: Error Rate Parity as a Minimal Standard of Evidence**

Where algorithmic discrimination is less intuitively identified and establishing disadvantaged and comparator groups is more complex, dependence on statistics to establish and prove discrimination will be necessary. To date, the court has preferred to specify what fairness requires on a case-by-case basis. However, since prior case law on algorithmic discrimination does not exist, there are no clear legal rules and principles that can be implemented into the design of algorithms to detect and mitigate discrimination.

There is a wealth of research on statistical measures of fairness among the data science and ML community, but this has been detached from both the legal and medical literature. This section explores how discrimination law can accommodate technical standards of fairness and locates these within the high-risk healthcare context. While using fairness metrics may appear to be a more burdensome task, it is contended that it will improve legal certainty, thereby helping claimants avoid unnecessary litigation and costs. Furthermore, the current unpredictability in how judges will handle cases of algorithmic discrimination renders the risk involved in the use of medical algorithms incalculable, which might hinder investment into these innovative and beneficial technologies. This section will consider three of the most well-researched and supported fairness metrics: fairness through unawareness, demographic parity, and error rate parity.

---

This section will establish the different ways that algorithmic discrimination can be measured and proven by claimants, and the problems associated with each when applied to bioethical principles. A thorough analysis will lead the author to argue that error rate parity is the most suitable fairness metric to be used as a minimal standard to determine whether a biased medical algorithm is discriminatory.

**Fairness Through Unawareness**

Fairness through unawareness (‘FTU’) states that if algorithms are blinded from protected characteristics while making decisions, the output decision is unlikely to be discriminatory. This metric proposes that fairness can be achieved by treating individuals as distinct from their membership in a particular group. FTU closely resembles current discrimination law jurisprudence and has been supported as a metric to determine the discriminatory output of algorithms used to train recruiting tools\(^{152}\) and determine jail sentences\(^ {153}\). Within these sectors, it is broadly accepted that algorithms should not be making predictions based on protected characteristics as there is no meaningful correlation between these factors. However, the same is not true within the medical context. Sometimes, medical decisions should be based upon protected characteristics given that race, gender, and age can affect health status.\(^ {154}\) A medical diagnostic algorithm which disregarded age and gender differences in patients with heart failure symptoms would not only be less accurate in terms of diagnosis, but would also reinforce health inequalities by offering groups of patients worse treatment.\(^ {155}\) Therefore, FTU would not work within the medical context.

---


\(^{154}\) Kristen Sethares and Elizabeth Chin, ‘Age and gender differences in physical heart failure symptom clusters’ (2021) 50 Heart & Lung 6.

\(^{155}\) ibid.
healthcare context because it would wrongly classify some algorithms as discriminatory even though their decisions were justifiably based on protected characteristics.

The FTU metric also overlooks more subtle forms of discrimination through factors that implicitly relate to the protected characteristics — like in indirect discrimination cases. For example, the algorithm that used healthcare expenditure as a proxy for a patient’s symptoms when calculating cardiomyopathy risk would not be considered discriminatory, according to FTU, despite clearly providing worse treatment to black patients. This metric ignores that historical inequalities and social determinants of health can also form the basis of an algorithm’s biased recommendation. Therefore, FTU is also unsatisfactory from a legal perspective as it fails to recognise significant cases of indirect discrimination.

The bioethical principle of beneficence means that doctors should always act for the benefit of the patient. This is at odds with the aims of FTU because eliminating protected characteristics from the decision-making process can be against the best interests of patients, as detailed above. This also relates to the principle of non-maleficence which requires that doctors do not harm patients. As illustrated, FTU has the potential to cause harm to patients through misdiagnoses. The principle of justice requires that medical resources and benefits are distributed fairly and equitably which means that decisions should sometimes be made based on an individual’s or group’s needs and circumstances, including protected characteristics if necessary. Therefore, FTU is an inadequate minimum standard of evidence from a legal, technical, and bioethical perspective.

**Demographic Parity**

The motivation is mathematical convenience: the aim is to make two numbers (i.e., recall) as close to equal as possible between two groups (i.e., white and black patients), solely to

---

156 Obermeyer and others (n 33).
158 ibid.
159 Hellman (n 151) 811.
satisfy a mathematical definition that says a system is fair when these two numbers are equal.¹⁶⁰

Demographic Parity (‘DP’) requires the proportion of individuals from each demographic group who are positively identified or selected by the algorithm to be roughly equal.¹⁶¹ For a diagnostic algorithm which predicts lung cancer, DP would require that diagnosis between the genders should be roughly equal.¹⁶² DP is useful because it provides a clear indication that different demographics are being treated differently (whether justifiably or not) by an algorithm’s decision. The problem with this, however, is that it does not consider that the prevalence of diseases will differ between different groups. For example, the incidence of breast cancer is significantly different between genders.¹⁶³ Therefore, a cancer diagnostic algorithm that has a significantly skewed positive diagnosis for women would reflect an accurate assessment. Yet, under DP, this would be evidence of discrimination because the positive diagnoses were not roughly equal across all gender groups. To satisfy DP, it might be required that female patients receive high rates of misdiagnosis to match the levels of breast cancer incidence in the male population. This would cause harmful misdiagnoses for women, but not be considered discriminatory. This problem is commonly termed the ‘levelling down objection’.¹⁶⁴

A partial solution to this issue would be to improve the system’s recall. In the context of cancer screening, it might be argued that false negatives are much more harmful than false positives; the latter means that the patient may

¹⁶² ibid.
¹⁶⁴ Mittelstadt, Watcher, Russell (n 160) 6.
suffer distress and will undertake unnecessary health checks or scans, whereas the
former means that future cases of cancer will go undiagnosed which risks fatality. Hence, an algorithm could be altered to change its predictions for the cases that it is least confident in diagnosing. By changing low confidence ‘low risk’ cases to ‘high risk’, the algorithm could diagnose more cases of cancer that would have previously been misdiagnosed, minimising preventable deaths. While this may have the potential to improve the accuracy of diagnosis, the lack of representative datasets means that it will prove difficult to find additional data to distinguish between the low and high-risk cases without introducing significant inaccuracies. Therefore, DP applied to medical algorithms causes high levels of inaccuracies which will render output recommendations even more harmful towards patients.

While on the surface DP may seem to be provoking the central bioethical principle of justice, this is only per strict egalitarian equality. DP achieves fairness by making everyone worse off, which does not align with the principles of equity. Fulfilling the principle of justice would seek to improve patient outcomes, not simply reduce them to a common level of harm. Given this, it is clear that the principle of non-maleficence is breached given that unnecessary pain and suffering can materialise in an attempt to equalise positive rates amongst groups. Most significantly, however, DP clearly conflicts with beneficence. Beneficence requires that medical algorithms not just avoid harm but actively promote the welfare and health of patients. A medical algorithm which conforms to DP would arbitrarily misdiagnose patients by favouring formal equality over harm reduction, clearly deviating from the principle of beneficence. Thus, DP is not suitable as a minimum standard of evidence in discrimination given its inherent inaccuracies when applied to the medical context which means it violates bioethical principles.

Error Rate Parity

Error Rate Parity (‘ERP’) measures whether an algorithm is making the same number of errors (both false positives and false negatives) across different subgroups of the population. Within medicine, ERP can be used to evaluate whether an algorithm is providing equal quality of care to different demographic groups. Grote and Keeling argue that ERP is the most promising candidate as a

---

165 Mittelstadt, Watcher, Russell (n 160) 5.
metric of fairness within the context of healthcare. The advantage of ERP over DP is that it does not assume that the ratio of false positives and false negatives is identical for different groups, but only that the harms are distributed fairly. According to this metric, if the algorithm which diagnoses heart disease generates a slightly higher rate of false negatives for both male and female patients, the outcome is fair since the harms have been distributed equally across the two groups.

While the distribution of harm may be equal according to ERP, the type of harm that is distributed between patient groups may not be equal. False negative results which are largely directed at historically mistreated groups could further widen health disparities for patients. Misdiagnoses would be more likely to appear among those marginalised patients, who will struggle the most to rectify the erroneous decisions through further medical attention due to the barriers to accessing healthcare. Concerning gender, if the tendency is that females receive higher levels of false negative results, this could further entrench the healthcare gender gap, as female patients already receive a lower standard and quality of healthcare than other groups. Therefore, the distribution of false negatives among marginalised groups will be important to consider when determining whether discrimination has occurred so that health disparities are not further widened by algorithmic bias.

As with DP, equalising the error rates across all the groups may result in making predictions worse for some groups. However, a disparity in error rates is more meaningful because it can provide additional normative motivations to interrogate how data and systemic biases have contributed to a higher

concentration of false positive or negative errors amongst specific groups. Hence, an appealing feature of using ERP as a minimum evidential requirement is that it will encourage the broader aims of health and data ethics.\textsuperscript{169} The act of measuring and reporting group-specific error rates can create an incentive for developers to minimise error rates by collecting more representative datasets and better samples and building more transparent models. Therefore, ERP appears to be the metric most compatible with legal and technical goals, as it provides the most appropriate balance between equity, normative goals, and practicality.

ERP has the strongest commitment to justice since it requires that there be a fair distribution of harm among the patient population. It does not determine fairness based on protected characteristics but instead upholds that discrimination is determined by reference to whether patients are unfairly disadvantaged by an algorithm’s predictions. It encourages the designers of algorithms to take positive steps to prevent harm to patients by engaging with accurate and representative data, which reflects the principle of beneficence. However, it is essential to acknowledge ERP’s susceptibility to the ‘levelling down’ objection, a valid concern that challenges this metric’s alignment with the principle of non-maleficence. This highlights how in striving to equalise error rates among all groups, an algorithm may be compelled to make inaccurate predictions for a group with high accuracy, leading to unnecessary and harmful output predictions.

To address this, it is crucial to elaborate on how the ‘levelling down’ objection can be mitigated. Developers can play a crucial role by making a concerted effort to ‘level up’ algorithmic predictions across all groups. This will involve collecting and using data from diverse groups, engaging in algorithmic auditing processes, and establishing continuous monitoring systems to adapt to changes in environments. What is distinguishable about ERP compared to DP and FTU is that its main objection can be mitigated, while problems with the other metrics cannot be addressed in the same way as they are inherent to the fairness criteria itself. Unlike with DP or FTU, it may be possible to reduce disparities in ERP through adjustments in the decision threshold or other techniques without significantly harming the model’s overall performance.

---

\textsuperscript{169} Barocas, Hardt and Narayanan (n 151).
Error Rate Parity: A Minimal Standard of Statistical Evidence for Discrimination

Accordingly, it is proposed that ERP should be used as a base-level standard for statistical evidence in cases of discrimination involving biased medical algorithms. ERP would be able to address the instances in which an algorithm’s decision is based on a proxy feature beyond the protected characteristics without introducing damaging inaccuracies into algorithmic decisions. From a practical perspective, ERP as a minimum standard of statistical evidence for discrimination has the effect of making litigation easier and more certain, because (a) claimants will have a coherent strategy and evidential standard to confidently raise claims in court, (b) controllers or developers will be able to refute potential claims, and (c) judges will have a starting point from which to investigate potential discrimination cases. ERP will also encourage system developers and controllers to pre-emptively minimise and address biases within their algorithms. At present, the AI industry lacks clear and consistent legally binding equality requirements that can be translated into algorithmic design to prevent discrimination. Therefore, using ERP as a minimum standard of statistical evidence for discrimination will not only ensure that injured patients are suitably compensated, but will also foster preventative measures to minimise the effects of algorithmic bias.

Nonetheless, it is still a useful minimum standard of evidence that the courts can use, notwithstanding that other fairness metrics can also be used by parties. It is also important to point out that statistical fairness metrics alone will not indicate algorithmic bias. Instead, ERP should be used to support or refute a claim of discrimination which the judiciary can then assess contextually. Despite its limitations and conditions, it is maintained that ERP provides the most suitable starting point for the court to move towards establishing and proving discrimination using statistical evidence. It might be thought that the court should require that medical algorithms satisfy all fairness metrics to avoid discrimination claims, but this is statistically impossible. Therefore, ERP is the most relevant and applicable fairness metric that should be used as a minimum standard of evidence in discrimination cases. This will not only benefit defendants by

encouraging ethical data practices but also claimants by resolving the challenges of establishing and proving discrimination. It will also aid the judiciary by creating greater guidance to interrogate claims of algorithmic discrimination.

**EX-ANTE LEGAL PROTECTIONS: MEDICAL DEVICES REGULATIONS 2002**

Presently, medical algorithms that qualify as medical devices are regulated by the Medical Devices Regulations 2002 (MDR). The MDR was not drafted to regulate algorithms, but rather, medical devices like diagnostic equipment, implantable devices, and surgical instruments. However, the distinct and intangible nature of algorithms raises unique challenges that do not fit within this existing framework. Hence, there are no appropriate standards and guidance that specifically apply to the regulation of biased medical algorithms. It is questionable whether this framework is the most appropriate to use to regulate medical algorithms, given that the requirements of *in vitro* devices will demand very different safeguards to a software-based tool. Nevertheless, given that the Government has decided not to establish any new AI regulations, it is likely that the MHRA will adapt the MDR accordingly to regulate medical algorithms.

This section will argue that the MDR fails to regulate many *general-purpose AI systems* which are applied to the medical context. This imposes responsibility on downstream developers to mitigate the presence of biases which could cause serious patient harm. It is recommended that the statutory wording of the MDR which defines the scope of medical devices is amended from ‘intended purpose’ to a ‘reasonably foreseeable purpose’ to ensure that upstream providers pay greater attention to the limits of their foundation models, and the biases that could emerge in high-stake environments.

---

171 Medical Devices Regulations 2002.
172 Henderson, Flood and Scassa (n 24) 17.
The Medical Devices Regulations 2002

General-purpose AI Adapted for Healthcare Application

General-purpose AI (‘GPAI’) refers to algorithms that can perform a wide range of tasks without being designed for a specific product or purpose. For example, the BakeryScan is a GPAI system that utilises ML technology for pattern recognition. It was originally intended to identify different types of bread and pastries but has since been adapted to detect cancerous cells, with a 99% accuracy score.\textsuperscript{173} Despite this new use case, the BakeryScan would fall outside the scope of the MDR because it was not originally intended for medical image detection.\textsuperscript{174}

The MDR assumes that all medical devices will have a fixed intended purpose, whereas the appealing feature of GPAI is that they will have underdetermined purposes. For instance, Open AI’s GPT-3,\textsuperscript{175} Google’s BERT,\textsuperscript{176} and Meta’s OPT\textsuperscript{177} are increasingly being adapted for clinical use cases through fine-tuning on biomedical datasets. These datasets are BioBert for biomedical text mining,\textsuperscript{178} G-BERT for treatment recommendation,\textsuperscript{179} Med-BERT for disease prediction,\textsuperscript{180} and ClinicalBERT for predicting hospital

\textsuperscript{174} Medical Devices Regulations 2002, Section 2(1).
\textsuperscript{175} Diane Korngiebel and Sean Mooney, ‘Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery’ (2021) 4 npj Digital Medicine 93.
\textsuperscript{176} Jacob Devlin and others, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’ (2019) 1 North American Chapter of the Association for Computational Linguistics.
\textsuperscript{179} Junyue Wang and others, ‘Pre-training of Graph Augmented Transformers for Medication Recommendation’ (2019) 28 International Joint Conference on AI.
\textsuperscript{180} Laila Rasmy and others, ‘Med-BERT: pre-trained contextualised embeddings on large-scale structured electronic health records for disease prediction’ (2021) 4 npj Digital
readmission.\footnote{Kevin Huang, Jaan Altosaar, and Rajesh Ranganath, ‘ClinicalBERT:Modeling Clinical Notes and Predicting Hospital Readmission’ (2020) CHIL 2020 Workshop.} Outside the intended purpose of these GPAI systems, these downstream medical applications fall outside the MDR’s scope. There is no obligation for either upstream providers or downstream developers to conduct risk assessments, post-marketing surveillance or pass conformity assessments for these medical applications of GPAI. Without such safeguards, biases that were not present in the original model which could change the output and behaviour of systems could be unregulated. This relates specifically to bias introduced from medical data due to the domain-specific language, unstandardised nature and pre-existing systemic biases found within the healthcare system. For example, a recent study revealed that the medical records of black patients contained more markers of disbelief compared to white patients.\footnote{Mary Catherine Beach and others, ‘Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women’ (2021) 36 Journal of General Internal Medicine 1708, 1711.} Since these biases were not accounted for in the original foundation model, this increases the risk profile of the algorithm and could cause unintended harm to patients. Thus, the problem with the MDR in its application to GPAI is that there is no clarity in terms of what happens once software has moved beyond its original intended purpose to the medical context, and who should comply with the MDR following its downstream medical application.

Who Should be Responsible for Regulatory Compliance of the Medical Applications of GPAI?

At present, if the upstream providers of GPAI do not wish for the purpose of their foundation models to be a medical device, they are dispensed of all liabilities and responsibilities under the MDR.\footnote{Daniel Tietjen and Ennio Schwind, ‘AI software ChatGPT- actually a medical device?’ (Taylorwessing.com, 27 March 2023) <https://www.taylorwessing.com/en/insights-and-events/insights/2023/03/ki-software-chatgpt> accessed 6 April 2023.} Upstream providers refer to large companies that offer the infrastructure of GPAI, such as Meta, OpenAI, and Alphabet. As a result of the regulatory burdens and expenses of compliance with MDR (as well as other sector-specific regulations), it is likely that developers of GPAI will avoid promoting the diverse application of their software for practical and commercial reasons. Providers will prefer to shift the divergent...
application of GPAI to downstream developers to release themselves of the regulatory responsibility for the risks and harm that materialise in single-use cases. OpenAI (the upstream provider of ChatGPT) has avoided alluding to the intended purpose of the system, and there is no mention of its medical application in promotional material, statements, or on its website.\textsuperscript{184} Of course, this is expected as it is challenging for developers to foresee all the possible applications of foundation models in different domains. Just within the medical context, there have been hundreds of downstream applications of ChatGPT, ranging from clinical decision support to patient communication and medical education.\textsuperscript{185} Given the extremely high frequency and variety of GPAI applications, relying on upstream providers to adequately monitor each use case could be considered an overly burdensome task. Upstream providers will also lack the medical expertise to contextually interrogate downstream outputs for bias.\textsuperscript{186}

Upstream providers will, however, have the greatest knowledge and understanding of the original foundation model which forms the basis in which biases emerge. Russell argues that it makes the most sense to assess the fairness of GPAI at its source since the large-scale vendors will be the entities with the data and design information to carry out conformity assessments.\textsuperscript{187} Even if downstream developers were under such an obligation, the underlying data and code of GPAI might be protected by trade secrets and patents, so a thorough assessment of its risk potential in the healthcare context might be unavailing. For

\begin{itemize}
  \item \textsuperscript{184} OpenAI, ‘Chat-GPT’ (OpenAI.com) <https://openai.com/chatgpt> accessed 28 January 2024.
\end{itemize}
example, Epic’s algorithm that manages US electronic health records is covered by multiple patents.\textsuperscript{188} Furthermore, upstream providers are increasingly creating research capacities focussed on their product applications in healthcare, such as Google Health and Microsoft’s AI for Health. This will offer upstream providers better knowledge of bias identification in medical applications. Although it is difficult to decide where the responsibility should lie for compliance with the MDR in relation to the medical applications of GPAI, this article argues that it should rest with upstream providers: they are in a better position to assess the limits of their foundation models and to absorb the responsibility and cost of compliance with the MDR. There is a concern that this burden might deter medical innovation because of the time-consuming and high costs of compliance. However, GPAIs (like medical chatbots) can have direct interaction and influence on patients without professional oversight so it is necessary that they are closely regulated to prevent harmful misdiagnosis and mistreatments. Recognising that upstream providers of GPAI are responsible for compliance with the MDR in the event of downstream medical applications of their models will encourage developers to think more carefully about the distribution of their models and their safety implications.

**MDR Recommendation: From ‘Intended Purpose’ to a ‘Reasonably Foreseeable Purpose’**

To establish who is responsible for compliance with the MDR for downstream applications of GPAI to the medical context, it is suggested that Regulation 9(1) of the MDR should be amended by changing the statutory wording from ‘intended purpose’ to a ‘reasonably foreseeable purpose’.\textsuperscript{189} If the medical application of GPAI is a reasonably foreseeable purpose of the model, the upstream provider will be required to comply with the MDR. This would require that upstream providers conduct risk assessments, have quality management systems, and operate post-market surveillance of their systems. A ‘reasonably foreseeable purpose’ better reflects the dynamic nature of GPAI in which its applications are varied, but this does not unduly place an impossible


responsibility on providers to determine all uses of this rapidly developing technology in the abstract. The wording is left deliberately broad to be considered on a case-by-case basis according to the expertise, evidence, and research available at the time to upstream providers. Regulators can then determine whether the use of GPAI could have been reasonably foreseen in a certain context. It is important that a greater duty be placed on upstream providers to consider these high-risk applications to prevent biases from causing harm in subsequent applications. This change is necessary to address the instances in which GPAI systems are later adapted to the medical setting but had previously fallen outside the scope of the MDR as this was not the original intended purpose of the software.

If upstream providers deem medical applications too high-risk, providers will be encouraged to set limits on how far developers can experiment with their foundation model without having to comply with extra responsibilities. Given that the MDHR will have close scrutiny of how these technologies are being used by patients, doctors, and researchers, they will be in a good position to enforce providers to set these limits if they are not proactively addressed. There is a fear that this may deter innovative medical applications of GPAI. However, given the high-stakes environment in which these applications exist, such preventative restrictions are necessary to ensure patient safety. It is important that GPAI applications in the medical context are attended to tentatively since harmful biases can be introduced in multiple stages of the value chain of GPAI, and no single entity has the knowledge to interrogate all these thoroughly. Nevertheless, since upstream providers are in the best position to do so with their thorough knowledge of the foundation model, they should be encouraged to work with downstream developers to explore how novel biases could emerge from the integration of large language models to domain-specific data sets. The partnership between NVIDIA (AI developer) and Paige.AI (health tech start-up) did this by leveraging the expertise and resources of both companies to create more ethical and effective cancer diagnosis tools. Therefore, extending

---

the scope of the MDR to devices that have a ‘reasonably foreseeable purpose’ in the medical context will give upstream providers a greater responsibility to define the boundaries of their software, and encourage better dialogue within the technology and healthcare communities to determine these limits.

It is important to contextualise this recommendation within the wider application of the MDR which captures diagnostic equipment, surgical tools, and implantable devices. The author does not deem that this change of wording will have a substantially negative impact on the regulation of these devices, given that they are usually developed specifically for medical purposes. In any case, widening the scope of the MDR might enhance patient safety further by capturing devices such as smartwatches and health apps that have been adapted to monitor health conditions and track symptoms.192 The Ada Lovelace Institute has also recommended that the EU AI Act determine risk categories of AI systems based on their ‘reasonably foreseeable purpose’ under Article 3(12).193 If implemented, it will be useful for the UK to observe how the EU interprets this standard, as technology grows in competency and use over time.

It must be acknowledged that this change of statutory wording will likely not capture the application of all GPAI to the medical context if this is not reasonably foreseeable by upstream providers. Some GPAI systems develop unexpected capabilities that would be impossible for providers to foresee, especially in the development stages of the system. As GPAI is trained on more data, these systems become more accurate and effective at identifying more complex patterns and relationships between data. This means that medical purposes for GPAI could be revealed much later, and it would not be a fair balance of responsibility to hold upstream providers under a duty to comply with the MDR for these applications. However, as GPAI is increasingly being embedded into healthcare settings, and the potential capabilities of these systems

193 Küspert, Moës, and Cunlop (n 117) 8.
become more apparent, what is conceived of as ‘reasonably foreseeable’ will evolve to reflect a broader range of medical GPAI applications. For example, in its infancy, the application of large language models to healthcare was not completely realised. However, it is now conceivable that they can be utilised for clinical decision support, medical chatbots, and the digitalisation of health records. Thus, it is not proposed that this change in wording will ensure that the application of GPAI to the medical context will always be regulated under the MDR. It is, however, a step in the right direction to ensure that the safe application of GPAI is encouraged by the law.

**SYSTEMIC CHANGES: THE LAW CANNOT DO IT ALONE**

So far, this paper has been dedicated to how hard law instruments will adapt to address the effects of algorithmic bias. Yet algorithmic bias is complex. Multiple sources of bias are likely to arise in the value chain of a medical algorithm, ranging from algorithm design flaws to non-representative and deficient data. While ex-post legal instruments could adequately redress individuals after they have experienced algorithmic bias, ex-ante provisions can only go so far as to prevent the biases from existing in the first instance. There have been proposals to tackle the causes of algorithmic bias through soft law measures, such as best practices and ethical guidelines. But since these are not legally binding, they rely on the spirit and enthusiasm of developers to comply with industry standards. Furthermore,

---

195 Roger McNamee, ‘Big Tech Needs to Be Regulated. Here Are 4 Ways to Curb Disinformation and Protect Our Privacy’ (TIME, 29 July 2020)
soft law measures usually require actions of companies that are currently not feasible. For example, while guidelines might encourage developers to use more representative training data, this can only happen if the possibility of attaining such data exists. Therefore, it appears that the law (through hard and soft measures) is not satisfactory alone to address the effects of algorithmic bias.

This paper turns to discussing systemic changes which are necessary to enable the law to both remedy and prevent algorithmic bias. It is argued that (a) greater data sharing is required to address data bias, and (b) a more diverse AI community is required to tackle coding bias. It is suggested that systemic change will not only minimise algorithmic bias but could also enhance medical treatment for those who have been systematically mistreated by the healthcare system. These systemic changes should be urgently encouraged to enable the law to address the effects of algorithmic bias and promote health equity.

Data Sharing and AI Workforce Diversity

Currently, there is an over-reliance on sourcing data from clinical trials, established research institutions and technology, which overrepresent Western, predominately white, European patients. In order to address this, more open science practices should be implemented amongst the medical and technology communities to develop datasets that represent the diverse patient population. Making data open and accessible ensures that medical information from more diverse populations can be used to train algorithms to be more sensitive to the needs and responses of different patient groups. Data sharing also mitigates sample bias by encouraging data to be collected from origins which have varying accessibility barriers. For example, medical data from rural health clinics will reveal different insights to that of medical health apps because they have different consumers. This proves useful when integrating medical data from more marginalised communities into mainstream algorithms that are being widely deployed in healthcare. Therefore, it is proposed that data sharing practices


196 Skin Cancer Foundation (n 58).
198 ibid.
should be encouraged during the development process of medical algorithms to ensure that algorithms are safer and more accurate.

Although there are restrictions on the sharing of sensitive healthcare data under the Data Protection Act 2018, data protection laws are more enabling than research communities typically believe.\textsuperscript{199} Data protection law permits the collection, processing, or use of special category data (genetic data, biometric data, health data) so long as this is per the conditions for safe use.\textsuperscript{200} There is also a concern that data sharing will not be prioritised by profit-driven technology companies who will be reluctant to share their datasets and seek data from other researchers. However, many companies are understanding the growing importance of data sharing to unlocking profitable products. For example, University College London academics recently developed an AI system that can predict heart attacks, stroke, and Parkinson’s disease in patients. The tool evolved from a collaboration between Moorfields Eye Hospital in London and Google’s DeepMind. Incentivising data sharing practices through levies and grants could also be a way to encourage organisations to collaborate to develop more accurate and fair medical algorithms. The UK Government has recently funded a partnership between Health Data Research UK, King’s College London, and the NHS AI Lab to create fairer generative AI models using the anonymised records of more than 1 million patients.\textsuperscript{201} Therefore, even though data sharing will require more time, effort, and incentivisation from private and public bodies, it will enable medical algorithms to be trained on data that is more representative of the patient population and rectify biases within models to improve patient safety.

The way algorithms are coded also contributes to algorithmic bias.\textsuperscript{202} A more diverse technology workforce is vital to prevent historical biases and stereotypes being replicated in algorithms. A report from Tech Nation found that

\textsuperscript{199} Centre for Data Ethics and Innovation, Review into bias in algorithmic decision-making (UK Government 2020) 9.

\textsuperscript{200} Data Protection Act 2018, Schedule 1.


\textsuperscript{202} Sara Hooker, ‘Moving beyond algorithmic bias is a data problem’ (2021) 2 Patterns 4.
only 19% of people within the technology industry are women, and less than 14% of women have authored AI-related research papers.\textsuperscript{203} A greater diversity of perspectives within the algorithmic development process will offer a greater chance that biases are identified by those who would be affected by the decision outcomes. Prioritising the voices of those in marginalised groups when framing problems and defining the parameters of medical algorithms is essential to address algorithmic bias in the design stage. Acknowledging that the biases of algorithms are largely attributable to the lack of diversity within the AI field is essential to warrant that algorithms are developed to serve all patients.

Data sharing practices, coupled with a more diverse workforce, could generate new algorithms that improve medical outcomes for marginalised patients. By identifying the unexplored health needs of those who have been traditionally mistreated by the healthcare system, medical algorithms have the potential to enhance, not just mitigate, health equity. The ‘FemTech’ phenomenon is an example of this, which focuses on improving medical treatment for female-specific conditions, such as contraception, fertility, and menstrual health.\textsuperscript{204} By investing in more data to reveal insights about female health, FemTech companies (which are usually led by a female workforce)\textsuperscript{205} are deploying algorithms which provide sensitive and accurate recommendations to women. For example, Ascertain has just developed an algorithm to predict the chance of a pregnant woman developing preeclampsia which can be used to prevent maternal mortalities.\textsuperscript{206} Therefore, while systemic changes are important to enable the law


\textsuperscript{205} Lucy Wilson, ‘Femtech Companies Closing the Gender Health Gap’ (Beauhurst, 30 June 2022) <https://www.beauhurst.com/blog/uk-femtech-startups/> accessed 27 April 2023.

to *mitigate* algorithmic bias, data sharing and improving the diversity of the AI workforce could also do more to *enhance* health equity.

**CONCLUSION**

This paper has examined how the UK’s hard law framework can address the effects of algorithmic bias in healthcare. By investigating the causes of algorithmic bias, it has analysed how hard law instruments can be revised to ensure that algorithms do not perpetuate existing biases and jeopardise patient safety. The author is sensitive to the fact that there are more challenges inherent in the application of hard law frameworks to algorithmic bias, and the recommendations that have been provided are not unassailable. However, this paper has aimed to focus on the most pressing issues, accompanied by recommendations that are appealing from a legal, technical, and bioethical perspective.

The paper began by discussing the (a) reasonable person standard of care, and (b) the foreseeability criterion in the common law tort of negligence. It explored how the application of biased medical algorithms will cause inconsistencies in the assessment of reasonableness and create a wide liability loophole when attempting to compensate patients injured by black box systems. To address this, it proposed that the tort law framework develop a reasonable algorithm standard of care to reflect the novelties of algorithmic decision-making and incentivise developers to ensure that the causes of algorithmic bias are mitigated in the design stages. To address the foreseeability loophole, it is suggested that where black box and opaque medical algorithms are used, for harm to be unforeseeable, the defendant must be able to provide evidence that they have maximised the extent to which the algorithm may be causally interpretable. This will prevent defendants from escaping liability for injuries caused by overly complex algorithms by requiring that they attempt to provide explanations as to why the algorithm predicted a certain outcome.

In relation to the Equality Act 2010, concerns were raised about how disadvantaged groups will suffer difficulties when establishing and proving algorithmic discrimination. To make this process easier and fairer, a new
minimum standard of evidence based on error rate parity can be used as a standard measure for the judiciary to assess discrimination claims. In the next section, the failure of the Medical Device Regulations to govern general-purpose AI that has been adapted to the medical context was critiqued. It was suggested that the defined scope of the Regulations be amended from ‘intended purpose’ to ‘a reasonably foreseeable purpose’ to ensure that upstream developers monitor the emergence of bias in medical applications of general-purpose AI. Finally, while this paper has mainly been dedicated to exploring how hard law instruments will address the effects of algorithmic bias, it is acknowledged that both soft and hard law measures have their limits. Thus, in the final section, the paper highlighted that systemic changes are necessary to enable the law to address algorithmic bias at its core and enhance healthcare equity.

The recommendations to amend hard law provisions will be useful to the medical, legal, and AI communities who can work together to develop effective hard law frameworks that ensure patient safety, promote accountability, and prevent algorithms from perpetuating existing biases. Collaboration amongst these communities is essential to ensure that lawyers are equipped with the knowledge and tools necessary to understand the complexities of algorithms and their implications in medicine. Greater clarity around these legal provisions will encourage innovation in medical algorithms which can transform healthcare and drastically improve patient outcomes. For this to succeed, the law needs to embrace new ideas to protect patients against negligence, discrimination, and harm caused by algorithmic bias.

As the use of medical algorithms continues to grow, future research should focus on exploring other fairness metrics that can be used within the discrimination framework that align with legal standards and bioethical principles to make the process of proving algorithmic discrimination accessible to patients. Moreover, more research is required to explore how data and privacy laws could be amended to facilitate data sharing practices to decrease reliance on outside funding. Private and public grants are not a sustainable solution to tackling the wide-scale issue of non-representative datasets. By implementing both preventative and compensatory measures, the law can effectively address algorithmic bias and promote equitable healthcare outcomes for all patients.