Cambridge Law Review (2020) Vol V, Issue ii, 58–87

The Curious Case of Doctor Ultron: How Well is English Law Currently Suited to Manage the Inherent Risks Associated with Black Box AI Medical Diagnostics?

Nikhil Dutt Sundaraj

I. Introduction

Hollywood has given us plenty to cherish over the years, and interestingly, a lot of modern technological inventions may very well have seen their creative inception on the silver screen. Not least among these inventions is artificial intelligence (AI) of which we have seen countless iterations, rarely benevolent. From Terminator's Skynet to Marvel's Ultron, artificial intelligence has been portrayed as a potentially uncontrollable force which could easily overrun its human creators.

So, how close are we to that? Well, at present we have AI in use in many industries and to an extent, some have replaced several tasks which used to be manned. However, those tasks tend to be relatively simple, like how a chess software program can play against a human opponent according to the rules of the game. However, what may seem a step closer to the 'Ultron' mould is a dynamic AI. Dynamic AI programs solve a problem using past input-output data alone in order to identify certain features common to each input that tend to influence the occurrence of a certain output. They do this *without* a set of preprogramed rules on how to arrive at that outcome.[1] In the context of medical diagnostics, this means that an AI which is meant to detect signs of skin cancer (for instance) would be fed past data and images of various skin nevi and the corresponding diagnosis,

[1] Anon, "Artificial Intelligence in Medicine: Machine Learning" (*IBM*) <https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine> (accessed 29 April 2020).

and then form its own 'process' to determine how future input data would be diagnosed.

These AI programmes function using predictive algorithms which enable it to diagnose a condition based on past data alone. Their use is incredibly attractive because AI's ability to sift through vast quantities of data saves a lot of time and money. This lightening of practitioners' menial yet arduous workload would free them up to attend to other cases or tasks with which AI is unable to help and allow junior doctors to learn more about those.[2] An AI program would effectively be a supplementary member of the medical team. Furthermore, their large workload capacity would make them useful allies in the fight against global pandemics like Covid-19 where mass testing and diagnosis is essential.[3]

Another important feature of these AI is their 'black box' nature. They are called as such because the exact mechanism and overarching 'thought process' the AI undergoes as the algorithms interact is unknown even to their software programmers.[4] As such, the doctors, hospitals and health authorities themselves do not know exactly how the AI arrives at a diagnosis, merely that it *does*. This is a distinctive feature of black box AI called 'opacity'.[5] Of course, with this opacity comes a question of legal accountability. Who is to be held liable for harm done by a black box diagnostic AI? That itself is a tough question to answer as there are multiple parties involved in the production of the AI program and its use to arrive at a diagnosis.[6] There is an inherent risk of harm for the patient that could arise from a device malfunction, meaning its failure to meet its performance specifications or otherwise perform as intended, especially if is leads to misdiagnosis by the AI.[7] Those risks must be dealt with, not least by the Law. Such misdiagnoses could themselves be caused by malfunctions such as contextual bias, software glitches, or improper use of the machine.[8] The parties in question that could contribute

[2] Varun H Buch, Irfan Ahmed. & Mahiben Maruthappu, "Artificial intelligence in medicine: current trends and future possibilities" (2018) British Journal of General Practice 68(668), 143–144.

[3] Wim Naudé, "Artificial Intelligence against COVID-19: An Early Review" (*Medium*, 12 April 2020) <https://towardsdatascience.com/artificial-intelligence-against-covid-19-an-early-review-92a8360edaba> (accessed 21 April 2020).

[4] Yavra Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation" (2018) Harvard Journal of Law & Technology 31(2).

[5] Dallas Card., "The 'black box' metaphor in machine learning" (*Medium*, 6 July 2017) <https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0> (accessed 21 April 2020).

[6] US FDA Code of Federal Regulations (CFR) Title 21, s 803.3(k).

[7] William Nicholson Price II, "Black-Box Medicine" (2015) 28 Harv JL & Tech 419.

[8] Max Tegmark, "Benefits & Risks of Artificial Intelligence" (*Future of Life Institute*, 30 May 2016) <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1> (accessed 21 April 2020).

to the risks are doctors, hospitals/medical centres and software developers/ manufacturers.

For the purposes of this discussion, 'dealing with' the risks involves: a) minimising the chance of the harm occurring, and b) taking appropriate action in the event that it occurs. 'Taking appropriate action' means firstly, adequately compensating the claimant and vindicating their rights and secondly, holding the wrongdoer sufficiently responsible with a sanction which is appropriately weighed against the harm done.

Furthermore, in dealing with the risks, the legal regime should be encouraging good medical practice while not overly discouraging the use and adoption of black box AI in the medical industry. As has been outlined by numerous academics and commentators in their works, regulatory agencies have a tendency to be very restrictive with the use of new technologies in fields as important as medicine and can be very heavy-handed in the potential penalties imposed for breaches of regulations.[9] This arguably is not without good reason as the main objective of new medical technologies is to make saving lives easier. However, a balance needs to be struck between, on one hand, the need to protect people from harm and hold responsible parties to account, and on the other hand, fulfil the need for new medical technologies in the market and obtain their potentially vast benefits.

The scope of this discussion will concern English Law. This will examine the measures possible under Medical Negligence (the main area in which civil suits for personal injury claims exist), regulations of the medical profession (the General Medical Council) and hospitals (the Care Quality Commission), as well as the relevant product regulations (EU Commission Regulation 2017/745 – the 'Medical Device Regulation') and product liability laws (the Consumer Protection Act 1987).

This discussion will only consider semi-autonomous AI, not fully-autonomous AI. This has been done in order to keep the scope of the dissertation manageable. This dissertation will target AI machines that operate mainly in an advisory role to the doctors when they are trying to diagnose a patient, as opposed to making the final call in the diagnosis itself. While semi-autonomous AI may generate a probable diagnosis, they do not make the final call.

Here, it will be assumed that the damage suffered to the patient in question was done during the course of the machine's normal clinical operation, not during the clinical trial stage when the machine's fitness for use is still being

[9] William Nicholson Price II, "Regulating Black-Box Medicine" (2017) 116 Mich L Rev 421.

assessed using human testing (like a beta test). This is because The Association of the British Pharmaceutical Industry (ABPI) has published detailed guidelines on compensation in clinical trials, which have been adopted to apply to clinical trials of medical devices as well. These Guidelines, whilst not legally binding, hold that that it is ethically reasonable for the participants to accept some of the inherent risks in testing a new treatment. Thus in order to avoid the discussion on the legality of such an exclusion, which itself has a whole different test of reasonableness to ordinary medical negligence, the scope for this discussion has been limited to the use of black box AI in normal clinical operation only.[10]

In Nicholson Price's 2018 work titled 'Medical Malpractice and Black Box Medicine', he makes a claim that the 'traditional medical negligence regime makes no sense'. Price was making this claim because he does not think that a doctor should be held to that test when they cannot intervene and pit their substantive judgment against the machine's. While this contentious claim was made with reference to US Law, it does provide an avenue to open up the discussion to refer to the Law in any given legal system, or at least one that follows the Common Law. This is a starting point for a wider discussion about how the Law deals with the risks associated with black box AI.

Price is not alone in this claim. James Vincent echoed the words of Elon Musk who says that AI regulation is essential as he thinks it will prove to be an existential threat to humanity. While that may sound a bit Hollywood-ish, the central point about the need for the regulation of black box AI is an important one.[11]

Thomas Hornigold of the University of Oxford Department of Physics reiterated this point in the medical context. He explains that:

"The problems of black-box algorithms that make inexplicable decisions are bad enough when you're trying to understand why that automated hiring chatbot was unimpressed by your job interview performance. In a healthcare context, where the decisions made could mean life or death, the consequences of algorithmic failure could be grave."[12]

[10] Anon, "Synapse: Clinical trial liability – navigating the legal, regulatory and insurance" (*Taylor Wessing United Kingdom*, 30 January 2019) <https://www.taylorwessing.com/synapse/january19.html> (accessed 21 April 2020).

[11] James Vincent, "Elon Musk says we need to regulate AI before it becomes a danger to humanity" (*The Verge*, 17 July 2017) <https://www.theverge.com/2017/7/17/15980954/elon-musk-ai-regulation-existential-threat> (accessed 21 April 2020).

[12] Thomas Hornigold et al., "Life-or-Death Algorithms: Avoiding the Black Box of AI in Medicine" (*Singularity Hub*, 18 December 2018) <https://singularityhub.com/2018/12/18/life-or-death-algorithms-the-black-box-of-ai-in-medicine-and-how-to-avoid-it/> (accessed 29 April 2020).

Furthermore, academics such as Isaac Kohane have voiced concerns that the safety levels of these AI are as of yet unknown seeing as they are only now just coming into practice and have not been through similar vetting to older technologies which have been in use for far longer. Furthermore, their lack of similarity to existing tests from the point of view of the human-machine interface make them much harder to read than say, a blood test or imaging test.[13]

II. Thesis

Given the numerous concerns voiced, the acknowledged need for legal intervention and weight on the issue, as well as the desired balance between competing aspects of the issue, the question needs to be asked how well English Law is currently suited to manage the inherent risks associated with black box AI medical diagnostics, especially in light of its vast potential benefits. The stand taken is that English Law is currently well suited to deal with the risks associated with black box AI in medical diagnostics. This is because of three main points, as listed below:

a) English Law's way of dealing with these risks is far wider than mere medical negligence. It also has an all-encompassing system of professional regulatory agencies and product regulations.

b) The medical negligence regime in tort is flexible enough to address these risks.

c) These risks themselves are overstated.

III. Professional and Product Regulations

English Law's way of dealing with these risks is far wider than mere medical negligence. Aside from medical negligence, it also provides for a system of professional regulatory agencies for practitioners and hospitals, and product regulations for medical equipment. These arms of the legal regime are forms of accountability for the production and use of medical equipment, regardless of its technological sophistication. This researcher's opinion on the issue is that as systems of accountability, they fulfil three main purposes:

a) To encourage good and safe practice

[13] Eric Bender, "Unpacking the Black Box in Artificial Intelligence for Medicine" (*Undark Magazine*, 12 April 2019) <https://undark.org/2019/12/04/black-box-artificial-intelligence/> (accessed 21 April 2020).

b) To attribute responsibility to the party which caused the harm suffered by the claimant/patient

c) To compensate the claimant/patient for the harm they have suffered

The extent to which these arms of the legal regime fulfil these purposes (both individually and collectively) will be the basis of their evaluation.

A. Professional regulatory agencies

Professional regulatory agencies help very well with purposes 1 and 2. The General Medical Council (GMC) and Care Quality Commission (CQC) outline what constitutes 'good practice' and there are enforcement systems for them and an avenue of attributing responsibility for wrongs. However, they do not compensate for harm.

(i) The General Medical Council (GMC)

The GMC, which gets its statutory power and duties form the Medical Act 1983, sets standards of competence and conduct which healthcare professionals must meet in order to register and practise, this includes updating and/or producing new guidance. The GMC's main enforcement abilities are through the Medical Practice Tribunal Service (MPTS).[14]

1. Encouraging good and safe practice

How does the GMC encourage good practice among medical practitioners and attribute responsibility? Its exact mechanisms can be found in the Medical Act 1983. It can influence the standard of a medical education institution by setting the qualification standard for medical practitioners.[15] Furthermore, it establishes the standards and requirements for medical practitioners to qualify and develops and promotes postgraduate education, which includes keeping one's knowledge of the field current and up to date.[16] This is reinforced by its advisory role to members on standards of professional performance.[17] Thus, it could be argued that this regulation encourages doctors to learn about avant-garde medical treatment and

[14] MPTS, "Our Role" (12 April 2020) <https://www.mpts-uk.org/about/our-role> (accessed 2 November 2020).
[15] ss 5(1) and 5(2), MA 1983.
[16] ss 34(h)(1), MA 1983.
[17] s 35(b), MA 1983.

devices like black box AI, which would equip them with the knowledge to ensure that its use is smooth. This advice found in their regulations is readily available and accessible online not just for medical professionals, but the public as well. As such, the GMC is well placed to direct the profession on trends/changes in the standards or functions of AI-driven medical devices. This solidifies its efforts to promote consistency across the board in good and safe standards of practice. This would mean that all treatments, including those involving black box AI, are kept within those standards without exception, which ensures good and safe practice be undertaken by all medical professionals.

However, what is absent is any specific mention of standards for AI, or any medical software for that matter. This does not give any clarity on how practitioners are expected to apply standards of professional performance to medical AI, nor does it allow a standard for AI to be derived from comparable software already in use. This is problematic seeing as the GMC guidelines are generally seen as ethical standards which guide the development of legal standards addressing the same issues.[18] The law should be taking cue from them. In this case, the general standards of professional performance, especially the duties of competence, professional knowledge/skill, and taking prompt action if they think patient safety is being compromised, would have to be read compatibly with the use of black box AI. As such, if, with the practitioner's presumably current professional knowledge, he/she thinks a patient's safety might be compromised by the machine, he/she has a duty to act to prevent that. The monitoring of this would be done by a 'responsible officer' whose watchdog function within each healthcare body ensures that the practitioners within it stick to the GMC's standards of good and safe practice.[19] This would encompass black box diagnostic AI if used. This would guide the law to take an approach to black box AI liability that revolves around assessing the skill and knowledge doctors (who followed the regulations) should have regarding AI.

2. Attributing responsibility for harm

The GMC's enforcement arm is the Medical Practitioners Tribunal Service (MPTS). When cases are brought to the GMC by aggrieved patients and the GMC investigates, it can take regulatory action against the responsible practitioner through the MPTS, which serves as a 'court' of sorts to hear cases.[20] An important

[18] Radu Damaschin, "Ethical Challenges in the Practice of Law" (*Lexology*, 17 August 2018) <https://www.lexology.com/library/detail.aspx?g=5d4e260f-edda-4e89-9b95-e64e0b9a57a6> (accessed 19 November 2020).

[19] s 45(A)(2), GMC Guidelines.

[20] MPTS (n 14).

feature of the MPTS is its transparency in hearing cases. Every single case it hears and adjudicates is published on its website for public viewing.[21] This constant publishing of MPTS decisions reminds practitioners of the standards expected of them and the possible action they could face if they fall short. With respect to black box AI, it would be up to the practitioners themselves to make the call on whether using a particular AI diagnostic device would meet the standards of good and safe practice, but they are actively dissuaded from cutting corners or forgoing any steps which they could reasonably take in using that equipment.

The flipside here is that the deterrence-motivated publication of these hearings can have a detrimental effect on a practitioner's career, as everything would be publicly accessible. It is conceivable that the publication of a malpractice case to do with black box AI could dissuade other practitioners from using the technology due to the potential risks to their career if the machine makes a mistake in the diagnosis and they fail to catch it. A possible solution to this would need to be patient-centric, seeing as the GMC's accountability system seems to rely on patients to bring the claims forward.[22] This could involve patients more in choosing the method of their diagnosis and treatment, and disclosing any material risks involved, such that the patient knows exactly what is being done and the reasons behind it. This may serve to alleviate some of the fear and anxiety-exacerbated complaints received by the GMC.

Furthermore, the publishing of MPTS decisions online for public viewing shows that the GMC takes the issue of patient safety very seriously and believes in the importance of public knowledge about their medical professionals in whom the public would entrust their lives and wellbeing. This dissemination of knowledge is an element of attributing responsibility for harm that has an impact in the public sphere by making the members of the public aware of it. Moreover, the low out-of-pocket costs to claimants here ensures that cost is not a deterrent to one's desire to attribute responsibility for harm done to them.

3. Compensation for harm

Where the GMC seems to fall short is that it does not have power conferred by the Act or any other powers to compensate claimants directly for their harm or any logistical costs that they incurred in registering their complaints. However, compensation may be awarded in the form of criminal damages, but only if the

[21] s 35(B)(4), GMC Guidelines.

[22] Anon, "Trustworthy Artificial Intelligence (AI) in healthcare" (*Med Tech Europe*, 28 November 2019) <https://www.medtecheurope.org/wp-content/uploads/2019/11/MTE_Nov19_Trustworthy-Artificial-Intelligence-in-healthcare-1.pdf> (accessed 21 April 2020).

case goes to Court, like it did for the case of Dr. Bawa-Garba.[23]

(ii) The Care Quality Commission (CQC)

The Care Quality Commission (CQC) sets guidelines for what is expected of healthcare providers in England, and as such is the relevant professional guideline regime for hospitals. It gets its statutory powers from the Health and Social Care Act 2008 Regulated Activities Regulations 2014. They maintain 'fundamental standards' which are "the standards below which your care must never fall."

1. Encouraging good and safe practice

The CQC helps promote good and safe practice. It sets out the various duties of the hospital under Reg. 9-17. One of its most important functions is its imposition of the hospital's responsibility to ensure that a patient understands their available treatment choices and enable them to make informed decisions therein.[24] Furthermore, under the HSCA 2008, there must be 'nominated individuals' in the hospital who are responsible for supervising the management of the facility, having met the criteria and been vetted through submitted written qualifications.[25] Similar to the 'responsible officer' arrangement in the GMC guidelines, this attributes responsibility to hospitals and medical centres which do not adhere to CQC standards. Thus, if a hospital or medical centre uses black box diagnostic AI, it would be the duty of that nominated individual to make sure that the hospital's policies and procedures are structurally sounds and geared towards the patient's safety in accordance with the CQC standards.

2. Attributing responsibility for harm and compensation for harm

However, with regard to attributing responsibility and compensating, the CQC cannot itself prosecute but can only take 'regulatory action' against any hospital in breach.[26] This, similar to the GMC's own complaint mechanism against individual healthcare practitioners, normally involves fines and cannot compensate for any harm. In fact, the only course of action provided for by the CQC for patients to obtain any form of recourse is the complaint mechanism

[23] *Bawa-Garba v GMC* [2018] EWCA Civ 1879.

[24] ss 9(3)(c) and 9(3)(d), HSCA 2008.

[25] Reg 6, HSCA 2008 RA Regulations 2014.

[26] NHS Providers, "Briefing: Complying With The Health And Social Care Act" (30 March 2015) <https://nhsproviders.org/media/1058/complying-with-the-health-and-social-care-act-2008-regulated-activities-regulations-2014.pdf> (accessed 20 April 2020).

set up under Reg. 16.[27] Like the GMC, however, it can prosecute for breaches of its guidelines which would also constitute breaches of criminal law (e.g. failure to obtain informed consent for a procedure, which could amount to criminal battery). This could also achieve a compensatory function if one of the remedies awarded in a criminal trial is compensation. Again, though, it's only in very limited situations and in the case of black box AI, would depend on whether the law holds that the harm was criminal.

B. Product regulations

The product regulation regime helps with 1 (promoting safe and good practice), 2 (attributing responsibility) and 3 (compensation for harm). However, it only addresses and applies to manufacturers, importers and distributors. Barring any new legislation post-Brexit, the primary legislation for these purposes at the moment is EU Regulation 2017/745 – the Medical Device Regulation (MDR), and it is supplemented (in its mechanism of compensating claimants) by the Product Liability Directive 1985 and Consumer Protection Act 1987.

(i) Preliminary Matters

The MDR, incorporated into UK Law, places the onus on the manufacturers, importers and distributors to ensure that their devices are of a certain standard before they are permitted for use. Those parties are then accountable as such to The Medicines and Healthcare products Regulatory Agency (MHRA). Black box diagnostic AI can arguably fit under the remit of this regulation as a medical device seeing as software can be a medical device, even if black box AI are not referred to specifically.[28]

As a preliminary matter, it should be explained how black box diagnostic AI fits under the ambit of the MDR. As outlined above, it counts as a medical device, but that is only one part of the puzzle. A malfunction of black box AI needs to be covered by the legislation as that is what will most likely cause the harm. A malfunction of black box AI is likely to fall under the definition of 'device deficiency' which explicitly includes 'malfunction' as a device deficiency.[29] Owing to this, and the reaffirmation that any 'inadequacy' in the performance of the device would suffice as a malfunction, it is safe to say that an AI malfunction falls under the regulation.

[27] Reg 16, HSCA 2008 RA Regulations 2014.
[28] Art 2(1), MDR.
[29] Art 2(59), MDR.

Next up is to determine whom exactly the legislation would address. The MDR imposes liability on 'manufacturers', 'importers' and 'distributors'. Only the software manufacturers and programmers arguably fit the definition seeing as they fit the bill for "a natural or legal person who manufactures or fully refurbishes a device or has a device designed, manufactured or fully refurbished, and markets that device under its name or trade mark".[30]

(ii) Encouraging Good and Safe Practice

The MDR promotes good and safe practice and minimises the risk of harm to the patient. This is especially apparent in its Article 83 'postmarket surveillance plan', which could go a long way to minimising risks of issues such as contextual bias. Contextual bias is when the AI has a skewed idea of what a certain diagnosis should look like with respect to the input data because it had been trained on input data from only a particular context (e.g. a skin cancer detection AI that was only trained on Caucasian males and would react differently to images of similar skin conditions on African-American females).[31] The MDR's obligatory manufacturer-run postmarket surveillance plan mandates monitoring the performance of their AI after release.[32] This could include an obligation to monitor the performance of black box AI. Furthermore, this could also cover an obligation to look out for incidence of contextual bias induced misdiagnosis. Seeing as malfunctions in the AI such as contextual bias would only become apparent once the AI has entered operation, this postmarket surveillance would keep track of its performance and enable manufacturers to ensure that it is keeping up to standard.

However, one needs to consider that the constant monitoring of manufacturing processes and conduct of postmarket surveillance could incur very high costs for the manufacturer and could easily disincentivise them from entering the market as a barrier to entry.[33] Be that as it may, it is much better than the alternative of having the product go onto the market sans regulation and with possibly lax safety standards, have it malfunction and cause harm, and then incur costs in redesign, recalls, wastage of sunk costs, and expensive litigation. Furthermore, this is a weighing exercise between the monetary and logistical costs of operation on one hand against the potential business profits on the other, AND the ethical considerations of transparency, accountability and patient safety.

[30] Art 2(30), MDR.

[31] Price II (n 9)

[32] Art 83(d)–83(f), MDR.

[33] A. Michael Froomkin, Ian R. Kerr and Joelle Pineau, "When AIs Outperform Doctors: The Dangers of a Tort-Induced Over-Reliance on Machine Learning and What (Not) to Do About It" (2019) 61 Arizona Law Review 33.

The latter values are important as they do not just safeguard the patients' bodily rights, but also provide a mandate for the product's use in the market and its safety standards by signalling that the manufacturers are looking out for patients' safety.

(iii) Attributing Responsibility for Harm

The MDR takes a risk-based approach to its classification of medical devices. According to MDR, most black box AI would be classified as a class IIa device as it is transient; meaning it is meant to be used at intervals during treatment purely for a diagnosis; and is not meant to administer any treatment. However, the classification is 'heightened' to one of a higher risk category if the device is meant to provide information that could lead to a serious (class Iib) or irreversible (class III) deterioration of the patient's health. As outlined by Minssen, this is consistent with the approach used by the United States FDA in their own Software as a Medical Device (SaMD) regulations which classifies devices according to the potential harm that could befall the patient if it does not perform correctly.[34] This risk-based approach does its job in suitably keeping the patient's welfare at the heart of its structure for accountability. However, it may be criticised as suffering from a lack of specificity in classifying conditions which may be on the 'border' between each class. For instance, a device such as the app SkinVision which spots irregular patterns on a patient's skin may provide information on a condition that could range from relatively minor consequences to the patient (e.g. detecting liver spots) to a condition that could be potentially life-threatening (e.g. detecting cancerous moles).[35] As such, the device may be seen to 'jump' classes depending on the purpose of its use for each individual case. This potential 'jumping' of classes would make it very difficult to fit the device into a specific classification at the time of manufacture. More importantly, patients might not know what to expect from a device which has not been adequately classified according to a single category of risk, thus hindering the transparency factor of the device too. It is thus important for manufacturers to perhaps specify and narrow the intended purpose of each device they market at the time of its manufacture such that doctors and patients alike can know what to expect from a device that's meant to track liver spots versus one that's meant to diagnose cancerous moles, and the devices can be ensured to comply with their corresponding regulatory class requirements.

[34] Timo Minssen, "Regulatory Responses to Medical Machine Learning" (2020) 7 Journal of Law and the Biosciences 1.

[35] H. A. Haenssle et. al, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists" (2018) 29 Annals of oncology: official journal of the European Society for Medical Oncology 8, 1836–1842.

Notwithstanding, once properly classified, the class IIa classification brings with it the imposition of certain obligations that make the manufacturers accountable for the device's adherence to conformity standards. This 'conformity assessment' refers to the process demonstrating whether the requirements of this Regulation relating to a device have been fulfilled.[36] As such, the robustness of the MDR's regulatory regime would ensure that there are multiple mechanisms of accountability to firstly, ensure that the device adheres to the principles of good and safe practice and secondly, attribute responsibility to the manufacturers for ensuring this.

There is one recurring problem, however, which is how the regulation's seemingly inherent need for transparency can be fulfilled if the black box AI is opaque. In this case, manufacturers would have to opt for a different kind of transparency. It is not the transparency of the AI's mechanism and thought process which would be essential. What is essential is the transparency of the manufacturing process to ensure that strict standards are adhered to, and the transparency in the evaluation of the machine's effectiveness through rigorous clinical trials and postmarket surveillance with frequent reports. This is one big reason why the MDR can address black box medical AI without being overly burdensome on manufacturers. The level of transparency and monitoring it calls for is high, but the nature of that transparency is not incompatible with the opacity of black box AI. A patient-centric risk-based approach could be taken for the research, validation and approval of this AI. As part of the MDR's Art. 83 postmarket surveillance regime, the product monitoring should be conducted based on its clinical operation and patient treatment, tailored to each hospital's capabilities.

It should be noted that there is a possible exemption from the MDR, but only under very specific circumstances. Article 5 says that the requirements of this Regulation shall *not* apply to devices, manufactured and used only within health institutions established in the EU, provided that all of a list of conditions are met.[37] That list calls for a stringent series of documentation, research, testing, submission of findings and an urgent need for the product in the market at the material time for it to be exempt from regulation. Hence, it does not relax the strictness of the requirements but merely puts out an alternative route for the product to safely be used in the market if it is urgently needed (i.e. no substitutes exist). This does draw a good balance between protecting patients from harm while not hindering the outreach and access to this useful technology in the market. However, it may not

[36] Art 2(40), MDR.
[37] Art 5(5)(a)–5(5)(f), MDR.

necessarily apply to the UK market because multiple black box diagnostic AI have been put forth or proposed for use such as RenalytixAI for kidney disease diagnosis and Cambridge Cancer Genomics for oncology, thus there are indeed substitutes and the exemption would not apply.

(iv) Compensation for Harm

Compensation for harm done to the patient by a medical device is not provided for explicitly in the MDR. For compensation matters, we need to refer to the Product Liability Directive 1985 (Directive 85/374/EEC) and its domestic legislative implementation, the Consumer Protection Act 1987 (CPA 1987).

This is because Article 10 of the MDR states that liability for manufacturers is to be under/according to PLD 1985.[38] The UK has incorporated this directive into domestic law by implementing the CPA in 1987. Besides this, the CPA also serves as a crystallisation of the product liability rules which had existed only in the Common Law up to that point.

In order to determine whether a patient harmed due to misdiagnosis can claim compensation, the injuring party must fall under the CPA 1987. It appears rather clear that the manufacturer as outlined in the MDR would fall under the CPA as well. This is because the Act applies to the producer, importer or person with name/trademark on product, just like the MDR.[39] The type of harm (personal injury) is covered as it is included in s.5 of the Act.[40]

The compensation needs to be for a product defect. Is misdiagnosis a 'defect'? A 'defect' is when a machine does not perform at a level that one would be reasonably entitled to expect it.[41] Arguably, some common mechanisms that drive the machine's provision of a misdiagnosis (such as contextual bias and software glitches) are defects. A more precise iteration of the level of safety that the claimant is entitled to expect from the product can be found in the case of *Hastings v Finsbury Orthopaedics Ltd*.[42] The claimant/user is entitled to expect that subject to de minimis considerations, the product's level of safety would not be worse, when measured by appropriate criteria, than existing alternative products that would otherwise have been used.

This is where it gets tricky with black box diagnostic AI. The closest thing that one might get to an 'alternative' product is clinical diagnostic aid software.

[38] Art 10(16), MDR.
[39] s 2(2)(a)–(c), CPA 1987.
[40] s 5(1), CPA 1987.
[41] s 3(2), CPA 1987.
[42] [2019] 11 WLUK 399.

This, however, is different as the purpose of clinical diagnostic aid software is not to generate a diagnosis, but to highlight certain features in the data that help the practitioner generate the diagnosis themselves. The primary point of comparison between the two would be their ability to highlight key items to aid the doctor, but only the AI goes further to generate a diagnosis.[43] As such, there is an inherent difference in their functions; arguably too different to be called similar or alternative products. As such, it could be argued that in the absence of an alternative product against which the safety of black box diagnostic AI can be measured, the interpretation of 'safety' should be as close as possible to the statute seeing as for the sake of patient safety, a novel product should still be held to the standards of the law without exception. On that note, since it is reasonable to expect a diagnostic machine to diagnose something properly, its failure to do so should probably constitute a defect. This argument may seem simplistic but is unfortunately limited by the machine's opacity. Were we able to know the exact steps taken by the machine in arriving at its decision, it would be easier to audit its process at each step and determine what is reasonable for us to expect it to do at each step. Unfortunately, our knowledge of its function is only limited to its input and output. As such, any assessment of its function will have to hinge on that. Hence, in the absence of an argument to the contrary, the law's application would probably be as close as possible to the legal test's wording, which would entail expecting a reasonable machine to diagnose something properly. As more becomes known about how these AI operate, the legal standard can move with that knowledge and perhaps be a bit more specific in its definition of a 'defect' for black box AI.

There are several possible defences open to the manufacturer under the MDR. Firstly, it is a defence to say that the defect did not exist in the product at the relevant time of sale.[44] In the context of black box AI, however, the occurrence of the malfunction is not the defect; it is the propensity of the machine to malfunction which is the defect.[45] That arguably existed at the time of manufacture/programming.

Secondly, it could be argued that black box AI technology faults may be 'outside the scope of technical knowledge' or outside the remit of what was reasonably expected to have been discovered.[46] This is to say that the state of knowledge at the time was not such that a producer of products of the same

[43] Gurpreet Dhaliwal, "Expert Opinion Software for Medical Diagnosis and Treatment—Reply" (2014) JAMA Internal Medicine 174(4), 639.

[44] Art 4(1)(d), MDR.

[45] Ryan Benjamin Abbott, "The Reasonable Computer: Disrupting the Paradigm of Tort Liability" (2018) 86 George Washington Law Review 1.

[46] Art 4(1)(c), MDR.

description as the product in question might be expected to have discovered the defect; probably due to its inherent opacity. It is a potential lacuna which manufacturers could exploit. However, one could counter this with an argument that software faults such as glitches and coding errors are common knowledge among any software developer, black box or not, and that the manufacturer should be taking all reasonable steps to ensure that they do not develop software that is prone to malfunctioning. Furthermore, as a matter of policy, it seems unfair to place the risk on consumers to bear the burden of harm done by a machine (or any medical device for that matter) just because its manufacturers do not know exactly how their own machine works. If this is done, it would further worsen scepticism of black box AI and fuel the reluctance to use it. Similarly, with respect to driverless cars, product liability is there because the purpose of the car's software is to anticipate and avoid crashes in general, not anticipate and avoid crashes with that particular factual matrix.[47] As such, it is unfair to place the risk of harm on the driver of the car just because the software was only tested in collisions with trees rather than pedestrians during product trials.

C. Tying the professional and product regulations together

As such, it has been shown that the professional and product regulations can separately and together fulfil the three main purposes of accountability. This is greatly aided by the fact that their use is not exclusive. Claimants can bring claims in both areas (professional or product regulation) should they wish to. However, do these systems impose liability at the expense of flexibility? Nicholson Price outlines the problem with American FDA regulations which in his words; is rigid and stifles innovation in the field, compounded by apparent fear about the machine's 'thought process' due to opacity.[48]

The way to move forward, Price says, may be one that involves collaborative regulation. Centralised FDA regulation premarket with healthcare providers and hospitals providing the data for the device's postmarket regulation when it is in operation.[49]

If these suggestions sound familiar, it is because they are very similar to the system already imposed by the EU MDR. The MDR takes an approach that emphasises strict monitoring of performance and manufacturing standards, but

[47] Andrew D. Selbst, Negligence and AI's Human Users (2019) 100 Boston University Law Review 1315.

[48] Price II (n 9)

[49] ibid.

making an effort to not overly hinder access to the market. In fact, the postmarket surveillance called for under Article 83 would arguably promote the entry of black box AI products into the market as that would provide the platform on which their assessment can be based.

In summary, the product regulation regime would generally keep a robust and thorough check on the safety and qualify of black box AI entering the market and hold parties to account. However, they fall short in obtaining compensation from medical bodies and practitioners.

IV. Medical Negligence

Despite the nature of medical negligence in English Common Law which some may see as 'traditional', the medical negligence regime is more flexible than some commentators may think. It can be adjusted to apply to cases of black box AI. Medical negligence is a branch of the Common Law tort of negligence and as such has the same basic elements; the duty of care, breach, remoteness and foreseeability of harm, and causation.[50] The main focus of the analysis here is on the duty of care and remoteness/foreseeability considerations as those will be the most 'novel' when applied to the context of black box diagnostic AI, and how they work with the regulations.

A. The duty of care

The standard of care needs to first be clear. In medical negligence, this is known as the standard of the 'reasonable doctor'. The reasonable doctor standard can be distilled from two cases; namely *Bolam* and *Bolitho*. In order to distil a more precise duty of care in the context of black box AI, both the Bolam and Bolitho principles need to be read consistently with each other. The Bolam principle states that if a doctor reaches the standard of a responsible body of professional medical opinion, they are not negligent. [51] The Bolitho principle follows this, but adds that if the doctor/practitioner acted in a way that is so irrational and that a 'layman' would deem 'logically indefensible', then the Court is entitled to reject the body of medical opinion and hold the practitioner liable anyway.[52] Therefore, there must be a case made for the use of black box diagnostic AI to be viewed within the ambit of what the responsible body of medical opinion would deem as standard

[50] Stephan Trahair, "The History of Medical Negligence in the UK" (*Enable Law*, 5 March 2019) <https://www.enablelaw.com/news/expert-opinion/the-history-of-medical-negligence-in-the-uk/> (accessed 21 April 2020).

[51] *Bolam v Friern Hospital Management Committee* [1957] 1 WLR 582.

[52] *Bolitho v City and Hackney Health Authority* [1996] 4 All ER 771.

practice. This may seem like a chicken-and-egg situation in which the AI would need to be in use first before the standard can be developed. However, that is not necessarily the case because it is possible for the Bolam/Bolitho principles to be applied even to novel factual scenarios. They are not limited to their specific factual matrices, nor are they by law 'locked' according to the level of technology and knowledge present in medical practice in that time. This means that the notion of what constitutes the standards of a 'reasonable doctor' is malleable and can change with time and advancements in the field.[53] As such, there should be minimal fuss in adapting it to situations involving black box AI.

The question then arises as to how we can adapt the Bolam/Bolitho tests to apply to black box AI. This, similar to the exercise of the duty of care in the case of Bolitho itself, requires an interpretation of the test in light of the actual treatment being sought and applied. This may very well call for more specific duties that should be followed by the practitioner in order to minimise the chance of harm befalling the patient due to misdiagnosis. This could include a duty to properly evaluate each machine's fitness for use and acquire as much knowledge as reasonably possible about how the machine works, and its possible malfunctions. It would then follow that the practitioner has a duty contingent on that to take reasonable steps to prevent the machine from malfunctioning.

Furthermore, even though the diagnostic process is automated, the case could also be made for the doctor apply the skill and knowledge of a reasonable doctor with their experience in exercising their judgment for the diagnosis. To use an obvious example, if a machine spots a small tattoo on a patient's skin and mistakenly diagnoses it as a form of skin cancer, the doctor should not then blindly follow the diagnosis. The doctor should read what the machine diagnoses and exercise his own knowledge to overrule it if need be. Putting a non-cancer afflicted patient through cancer treatment could cause a great deal of physical and psychological suffering. Furthermore, the case of *Wagon Mound (no. 2)* states that loss will be recoverable where the extent of possible harm is so great that a reasonable man would guard against it, even if the chance of the loss occurring was very small.[54] This is a clear risk-based attribution of liability. In a case where the extent of harm due to a misdiagnosis is grave physical harm or even death, it should follow that the doctors would take all reasonable steps to prevent against it. While not all black box diagnostic AI would be diagnosing a condition that could lead to that extent of harm, this rationale would apply to those that do, such as

[53] Jonathan Morgan, "Torts and Technology" in Roger Brownsword, Eloise Scotford and Karen Young (eds), *The Oxford Handbook of Law, Regulation, and Technology* (OUP 2017).

[54] *Overseas Tankship (UK) Ltd v The Miller Steamship Co (Wagon Mound No. 2)* [1966] UKPC 10.

those meant to diagnose heart conditions, brain tumours, respiratory tract fibrosis or cancer.

This is especially pertinent since the doctors 'sign off' on every treatment, indicating that they administered it with their best knowledge and have assumed responsibility for the patient's wellbeing.[55] This is arguably analogous to scenarios when a student doctor, radiologist or nurse malperforms/misdiagnoses. Doctors, having signed off as supervisors overseeing the treatment and diagnosis as a whole, are still accountable for that. As such, they make the final call while the other members of the team are merely helping them obtain the information they need and carry out the procedures they deem suitable. Similarly, the AI merely helps the doctors with the diagnosis and cannot bear responsibility (legally or logically) for its actions. If the doctor has the responsibility for ensuring patients' safety, they should arguably bear the responsibility for ensuring that the AI used does not jeopardise said treatment.

The flexibility of the reasonable doctor standard is arguably acknowledged by the GMC itself. The fact that the GMC also uses the 'reasonable doctor' as the standard in its case handling shows that even the professional regulatory authority, which is generally seen as the go-to for guidance on how the law should develop, regards the current tort legal standard as suitable for attributing liability. Most notably, the GMC gauges the standard expected of the doctor in any case against the medical negligence legal standard of the reasonable doctor with the same seniority and type (e.g. senior cardiac surgeon against reasonable senior cardiac surgeon). Furthermore, that can be used as evidence in court to prove a breach of the practitioner's duty of care. This was shown in the case of *Bawa-Garba v GMC*.[56]

In this 2014-2015 case, the MPTS found that Dr. Bawa-Garba fell below the reasonable standard of a competent doctor of her level. When the case went to court after the tribunal hearing, the EWHC used the MPTS decision as evidence for the same.[57] This shows the compatibility between the standard of the reasonable doctor in the legal and professional contexts. When read compatibly with the duty imposed by the GMC on doctors to constantly update their skills, knowledge and expertise in their field, it could (and should) mandate maintaining reasonable working knowledge about how black box diagnostic AI function and ensuring that they adapt their standard of care to cover any harm that could

[55] AOMRC ed., "Guidance for Taking Responsibility: Accountable Clinicians and Informed Patients. *Academy Of Medical Royal Colleges*" (30 June 2014) <https://www.aomrc.org.uk/wp-content/uploads/2016/05/Taking_Responsibility_Accountable_Clinicians_0614.pdf> (accessed 29 April 2020).

[56] *Bawa-Garba v GMC* [2018] EWCA Civ 1879.

[57] [2018] EWHC 76 (Admin).

potentially be inflicted by it.

Besides the duties of the reasonable doctor, UK medical negligence law also imposes a duty to disclose the inherent risks associated with the proposed medical treatment or procedure from the case of *Montgomery v Lanarkshire Health Board*. This provides ample room for patients to be warned about the inherent risks associated with their proposed treatment, including possible side effects and complications. The duty to disclose exists when a reasonable person in the patient's position would be likely to attach significance to the risk, or the doctor is or should reasonably be aware that the particular patient would likely attach significance to it.[58] This is tricky because with the avant-garde nature of black box AI, it is difficult to tell what a reasonable person would know or apprehend about it, and thus difficult to know what level of risk they might perceive from that procedure. However, the harm that could befall the patient as a result of black box misdiagnosis is similar to the harm that could befall them with any faulty diagnosis, and thus the risk apprehended should be the same – the risk of misdiagnosis. The doctor's choice to go along with the diagnosis and administer the treatment is what causes the harm directly. So, since the current medical negligence regime covers ordinary misdiagnosis-related harm, there is no reason why it should not cover black box AI misdiagnosis as well.

B. Remoteness and foreseeability of harm

The standard for the foreseeability of harm in English Law comes from the case of *Wagon Mound (no. 1)*.[59] This posits that for the defendant to be held liable for the harm done to the claimant, it must have been reasonably foreseeable that said harm would flow from the breach of the duty of care. In the context of medical negligence, the harm that befalls the patient must be reasonably foreseeable by a reasonable doctor as consequence of using the AI (which gave the misdiagnosis). Arguably, it should still suffice for situations with black box AI. This is because the doctor's knowledge that the AI might possibly be faulty should lead them to anticipate a misdiagnosis and from then on, it is established precedent in *Khan v MNX* that harm is very much foreseeable due to wrong treatment meted out on a misdiagnosis.[60]

But what about a situation where involving contextual bias, where the machine has apparently functioned perfectly in training, but could (yet not necessarily will) misdiagnose patients in another context owing to its bias? It is arguable in that situation that there is essentially no way that the hospital or doctor

[58] *Montgomery v Lanarkshire Health Board* [2015] UKSC 11, 84–87.
[59] *Overseas Tankship (UK) Ltd v Morts Dock and Engineering Co Ltd (Wagon Mound No. 1)* [1961] UKPC 2.
[60] *Khan v MNX* [2018] EWCA Civ 2609.

could have foreseen the misdiagnosis because ostensibly, the machine works fine.

That may be conceivable as a possible scenario. However, as explained previously, practitioners would have a duty to maintain the utmost of their knowledge on how the machine could go wrong in practice, and take reasonable steps to prevent it as part of the Bolam/Bolitho regime. Since the issue of contextual bias is already making the rounds among developers, healthcare entities and academics, this should arguably be one of the 'ways' in which they should foresee the AI going awry in practice. Moreover, those reasonable steps could include familiarising the AI with data from the context in which it is meant to operate clinically so that it 'knows' its new situation. This does not require an extension/modification of the current law, merely a novel variation of how the duty is fulfilled in practice.

However, foreseeability is also an issue because due to black box opacity, it is impossible to tell if or when some harm might actually occur until the damage is done. This applies to situations of contextual bias and other errors too.[61] Basically, no one even knew of the issue of contextual bias until it had occurred, and people were being misdiagnosed in testing. Any analysis must be retrospective before the same issue is considered 'foreseeable' in subsequent cases.

Nevertheless, we need to classify this as one of the very human limitations of our interactions with computers. We are by design (the computer's and ours) unable to understand how specific algorithms interact and undergo a machine learning process in order to learn how to respond to a certain input. The counterfactual here which would argue for liability would claim that there is a scenario where the doctor/hospital using the AI program is able to gain a sophisticated enough understanding of the machine in order to accurately predict how and when/under what circumstances it would go wrong. That is simply impossible. Remember that the *Wagon Mound (no. 1)* standard calls for liability for what is *reasonably* foreseeable. While the word 'reasonably' may sometimes seem like a catch-all for anyone who tries to argue for absence of liability on novel technology, it corresponds with reality to say that there is no way this specific manner in which the malfunction happens would be foreseeable prior to it happening.[62] Holding doctors/hospitals responsible irrespective of fault here would effectively be strict liability. This would be undesirable as it manifestly disincentivises the use of the machine. Besides, strict liability is already a standard imposed on the producer via product regulations. Hence, imposing strict liability in negligence as well may overly discourage the use and development of the technology altogether.

[61] Tal Zarsky, "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making" (2015) 41 Sage Journals Science, Technology & Human Values 118.

[62] Selbst (n 47)

What type of harm then should fall under the 'reasonably foreseeable' harm that counts under the Wagon Mound test in the context of black box AI? Simply, the foreseeable type of harm is misdiagnosis. That could reasonably be foreseen if the diagnostic AI malfunctions. Therefore, taking reasonable steps to prevent it should include familiarisation with the machine's operation, ways it could go wrong, and constant monitoring of its results to ensure that patients have not been harmed due to misdiagnoses.

C. Hospitals' liability

How about the hospital's liability under negligence? Is there a test for the 'reasonable hospital'? Yes, but it is not the Bolam/Bolitho test. The question to be asked is "what would the reasonable hospital have done in that situation?" This seems rather vague and is not well-explained in the case law. It should be asked then whether that might provide a loophole through which hospitals could escape liability to an extent.

It probably would not because English Law tends to rely heavily on the doctrine of vicarious liability in situations involving medical negligence.[63] That is why we tend to get case names like 'Claimant X v Y Hospital/Health Authority'. This doctrine of vicarious liability can place the burden of compensating an injured claimant on the hospital itself. Owing to this doctrine of vicarious liability and coupled with the professional regulations of the CQC, hospitals are therefore incentivised to make sure that their equipment, facilities, staff and procedures are as effective as can be to treat patients so that they are not made vicariously liable.

D. Tying the regulations and medical negligence law together

Drawing a comparison between medical negligence law and the regulations (both professional and product), the regulations still suffer one drawback. Medical negligence is seemingly the ONLY way in which an injured claimant can *seek compensation* from a doctor or medical institution, as opposed to a manufacturer. On top of that, medical negligence encourages good and safe practice by being a disincentive to malpractice thereby motivating hospitals and practitioners to take all reasonable steps possible to ensure patient safety. It attributes responsibility by identifying the parties that caused the harm and through its vicarious liability mechanism, can hold both the doctors and hospitals in which they are employed to

[63] Clare Feikert, "Medical Malpractice Liability: United Kingdom (England and Wales)" *(Medical Malpractice Liability: United Kingdom (England and Wales) | Law Library of Congress*, 1 May 2012) <https://www.loc.gov/law/help/medical-malpractice-liability/uk.php> (accessed 20 April 2020).

account. It can compensate the claimant as mentioned earlier because the Court will decide upon the remedy in any civil case.

As such, it should be for claimants to decide which route they would rather take to sue for the harm done as the statute would make it easier to hold the manufacturer liable whereas if their goal is to go after the practitioner and medical centre AND receive compensation, the medical negligence route would be better. If they wish to incur minimal costs (i.e. no litigation costs) but still want to hold the responsible doctor or hospital to account, then the professional regulations complaint system via the GMC would be best, albeit without the possibility of compensation.

The existing tort regime is adept at compensating claimants for personal injury, which is the main type of harm for which patients would sue related to misdiagnoses, even in tricky situations when the material cause of the negligent harm is unclear. Supposing a patient had obtained two separate opinions from two separate doctors on a diagnosis; the tort regime is flexible enough to allow claimants to sue both using a several and proportionate liability argument. This was enshrined into Common Law by the case of *Barker v Corus*.[64] Both doctors would be liable in proportion to the amount of damage they caused, assuming causation could be made out for both.

Joint and several liability such as that found in *Fairchild v Glenhaven Funeral Services* may not be available.[65] This is because the Compensation Act 2006 only expressly applies it to asbestos related harm.[66] On the other hand, the Act does not expressly rule out joint and several liability for other reasons either, so it could be argued for a case of black box AI. Notwithstanding, this would need to be backed up by a strong policy argument like that which enabled joint and several liability to be enshrined for victims of asbestos-related harm. This was swift and fierce political backlash against the ruling in *Barker v Corus* which had discounted this liability for any sort of harm. Large numbers of workers, families, trade unions, and Members of Parliament had called for the reversal of the ruling on the basis that it would undermine full compensation for working people and their families. As such, it could be said that similarly wide-ranging and severe socio-political consequences would need to be foreseeable before Parliament or the Courts would think about awarding joint and several liability for harm done by black box AI. Another reason could be the difficulty in proving which party exactly caused the harm. This would be most pertinent in cases of contextual bias when it is unknown whether the manufacturer's training data or the hospital's own log of case data had fed the machine the bias. In that scenario, a joint and several liability regime may

[64] [2006] UKHL 20.
[65] [2002] UKHL 22.
[66] s 3(2)(b), Compensation Act 2006.

better compensate the claimant and attribute responsibility to each party which arguably should have taken steps to prevent the issue from surfacing.

As such, even with the novel and relatively unexplored nature of liability for black box AI, English medical negligence is flexible enough to adapt its tests and standards of liability to cover risks associated with black box AI. It may require further guidance, perhaps in the form of an explicit statement of the specific nature of *how* to fulfil these duties in relation to black box AI. Realistically, however, the judiciary and/or legislature could only do that with more (currently scarce) knowledge on how the AI works in practice. Nevertheless, for now, the current regime will suffice and has to develop incrementally. Furthermore, the professional regulations and doctors' expertise would be relied on in order to ensure that in the meantime, they keep their knowledge up to date with the level and type of technology they are using and apply that knowledge to the fulfilment of their duty of care as medical practitioners. Moreover, the inherent flexibility of the medical negligence regime's standard of care ensures that doctors and hospitals have enough leeway to incorporate black box AI into their procedures without being worried about potential liability, so long as they ensure that they take reasonable steps to prevent harm to the patient. Such due diligence is not too dissimilar from the standard that is expected of them when any new piece of medical technology enters operation.[67] That has helped the UK's medical industry manage to hence far keep relatively current and ensure that novel and beneficial methods are not kept from benefiting patients that need them. Hence, it would provide a suitable platform for the introduction of black box diagnostic AI into more routine use among UK hospitals.

What English Law seems to lack is any specific legislation (or case law) that addresses AI or black box AI and is tailored to its issues. However, it does not seem that there is an urgent need for that seeing as the main purposes of accountability are fulfilled by the existing areas. Additionally, the high costs of litigation may act as a barrier for aggrieved patients to claim compensation, as they would only be able to use the MPTS and CQC complaint system. However, this is the reality for most personal injury claims and it would be up to the patient upon advice from their lawyer to decide whether the quantity and certainty of compensation outweighs the cost of seeking it. As such, on the balance of things, English Law has all the appropriate bases covered. Furthermore, it does not seem that any part of the English legal regime would be so draconian in their policing of the use of new devices that it might stifle their adoption and growth in the industry. In fact, it is arguable that the professional regulations and reasonable doctor standard actively

[67] Morgan (n 53)

encourage the use of beneficial novel technologies such as black box AI.

V. Much Ado About (Almost) Nothing?

Finally, as apprehensive as some commentators may be about new and relatively untested technologies, it would appear that the risks associated with black box medical diagnostics are overstated.

A. Opacity is not a barrier against use

Opacity, despite being a much-discussed drawback of black box AI, is not necessarily unique to black box AI, or any type of software for that matter. Doctors do not necessarily have a deep understanding of the exact mechanisms behind how every single pill/treatment works for patients and yet they are still in regular use with minimal repercussions (barring certain expected side effects). Thus, these treatments represent 'black boxes' in themselves too. For instance, Paracetamol is one of the most common types of drug administered to treat a fever. Doctors know that it lowers body temperature and that it works, given its long and successful track record. However, the kicker is that neither the doctors nor the pharmacology experts who engineer it know exactly how it does what it does or how it works.[68] Nevertheless, Paracetamol remains the most prescribed over-the-counter and clinical form of analgesic in the world today and doctors who prescribe them are not expected to know how it works in order to prescribe it. As such, the risks associated with black box medicine associated with doctors' non-knowledge alone may be exaggerated.

However, the difference is that with most other types of treatment, even the other 'black boxes' like complex treatments, there is *someone* down the chain who knows how it works. With the aforementioned example of paracetamol, it could be argued that since most manufacturers seem to have their own 'secret compound' for their iteration of paracetamol-based medicine, these companies know which ingredients go into the compound and hence make the difference in the pill's function. As such, if parties want to conduct their research (as hospitals and doctors themselves often do in order to ascertain the possible side effects), they can do so. However, with black box diagnostic AI, even the software developers (who coded the algorithms) themselves do not know how the machine 'thinks' to formulate the diagnosis after they program it.[69] As such, the hospitals and doctors themselves are unable to analyse the mechanism behind the treatment to tell with

[68] Gerard A. McKay and Matthew R. Walters, "Clinical Trials and Drug Development," *Clinical Pharmacology and Therapeutics* (9th edn Wiley-Blackwell 2013).

[69] Price II (n 9).

certainty what the nature or extent of the risk is, unlike pharmacology companies which can clearly tell you the possible side effects from consuming paracetamol.

Nevertheless, when it comes to medical treatment, sometimes there is not always somebody who understands exactly *why* and *how* something works. Using the paracetamol example; the only way that pharmacology companies have been able to determine which compounds work to lower body temperature and which do not is through extensive testing (mainly on animals) and pure trial and error. Hence, it is inaccurate to say that paracetamol manufacturers have an exact handle on the bio-chemical processes that make their products work. Similarly, black box AI manufacturers do not have an exact handle on how their machines 'think' to generate any given output. To an extent, some trial and error (clinical testing) will be necessary to prove these devices' efficacy, and postmarket surveillance to ensure that it works for extended periods of use. Fortunately, both mechanisms are provided for in the MDR. A caveat should be added; that the trial subjects should be representative of the population that the machine will serve, or that the trials should take place in different contexts to identify whether the bias exists; to avoid contextual bias.

B. Technological advances could increase transparency

Furthermore, hence far there is no clear indication that things could go gravely wrong. Given the speed at which AI technology is advancing, software developers and programmers may yet develop a way to monitor the machine's substantive decision-making process. That seems to be happening as we speak. In a paper presented at the ACM CHI Conference on Human Factors in Computing Systems in 2019, researchers from MIT, the Hong Kong University of Science and Technology (HKUST), and Zhejiang University described a transparency-aiding tool that puts the analyses and control of automated machine learning methods into users' hands. This tool is called ATMSeer, and visualizes an AI's search process in an approachable user-friendly interface.[70] While it does not allow software engineers and users to alter the machine learning process as it is happening, it provides a hitherto unavailable method to track what the machine is doing and thereby make more substantive assessments of its output. For black box diagnostic AI, this would allow doctors to more easily track the diagnostic process and once they have received the diagnosis, 'audit' the machine's process to

[70] Rob Matheson, "Cracking open the black box of automated machine learning" (*MIT News*, 31 May 2019) <http://news.mit.edu/2019/atmseer-machine-learning-black-box-0531> (accessed 21 April 2020).

an extent to make a note of how it arrived at the conclusion it did.[71] However, the doctor would still need to exercise their own clinical judgment in determining the diagnosis' accuracy. Furthermore, it could aid manufacturers in their premarket assessments of the machines' performance by enabling them to break down the machine's thought process into digestible 'stages' at which point they can keep track of it and ensure the accuracy of its 'thoughts'.

VI. Limitations and Conclusion

A. What about more autonomous AI?

The discussion of this thesis does have its limitations. Firstly, this argument may change if it were to be used in cases with more intelligent/autonomous AI. It is difficult to determine at what point would the practitioners then stop being liable and shift the blame almost exclusively towards the manufacturers (and hospitals for using the AI in their medical treatment).

This could be a tricky issue for foreseeability especially seeing as we must distinguish between AI that exists for specified and limited purposes, like every form of AI currently on the market, and something closer to artificial general intelligence (AGI), sometimes called "strong AI." The latter is many years off and essentially unrelated to existing machine learning technologies. At that point, the foreknowledge that the AI could do anything at all could paradoxically increase the range of what is considered foreseeable. At least until then, category foreseeability (i.e. foreseeing the specific type of harm done which may be unrelated to the machine's primary function) should not be a concern.[72]

The MDR says that in the event that a modification is made to the software that changes its original performance, safety or intended use, or interpretation of data, a new UDI designation would be required.[73] – A fully autonomous AI that 'thinks' more like a human may fall under this category seeing as its 'thought process' on which its function is based would change as it learns and evolves. The manufacturer may be under an obligation to continually update the UDI, meaning it would also have to conduct near-constant postmarket surveillance on the fully autonomous AI.

There is also an issue of meaningful human control (MHC). The question is why and to what extent human control in AI is necessary or desirable for decision

[71] Thomas Wischmeyer, "Artificial Intelligence and Transparency: Opening the Black Box," *Regulating Artificial Intelligence*, vol 1 (Springer Nature Switzerland 2020).

[72] Selbst (n 47)

[73] Art 6(5)(2), MDR.

making in certain contexts.[74] The degree of MHC to set a threshold for liability of the human user of fully autonomous AI would be hard to determine.

Causation may be a much thornier requirement to fulfil when trying to attribute wrong in the case of a fully autonomous AI, especially when no doctor is present or the doctor does not provide any meaningful human intervention. It would also be hard to make out causation for the manufacturer as the choice to use the machine could constitute a novus actus and it is harder to prove that the manufacturer knew or reasonably foresaw ex ante what the machine would do unless they already knew it was faulty. Yavra Bathaee says a way to overcome this is by using a sliding scale of causation by applying the more traditional test for less autonomous AI and for more autonomous AI, and make it transparency-based.[75]

B. The brexit problem

Secondly, it is not clear how Brexit will affect this analysis, especially regarding the regulations: The MDR had direct effect in UK Law through the UK's assent to it as a (now former) member of the EU. Under the Brexit withdrawal agreement, the UK must continue to give effect to EU directives until 31st Dec 2020. The status of the MDR and its law under the English legal regime thereafter is unknown as of yet as the government has not given a clear indication of how it intends to regulate medical devices under the law post-withdrawal period. It could enshrine the MDR's regulatory regime under UK Law with an Act of Parliament, which should not change anything from the current regime (aside from the fact that applicants can no longer have their cases of breach of the regulation heard in an EU Court). They could come up with their own regulatory system, which would be time-consuming and cumbersome considering that devices have already been/are already being approved under the current regime and would need to be re-approved. It may very well be that the UK's regime of medical device regulation would remain very similar or the same post-withdrawal period as that seems to make the most sense.

The latest developments (as of 2 November 2020) state that from 1 January 2021 the Medicines and Healthcare products Regulatory Agency (MHRA) will be responsible for UK medical devices market that are currently regulated by the

[74] Liam G. McCoy et al., "On Meaningful Human Control In High-Stakes Machine-Human Partnerships" (*AI Pulse*, 26 September 2019) <https://aipulse.org/on-meaningful-human-control-in-high-stakes-machine-human-partnerships/> (accessed 21 April 2020).

[75] Bathaee (n 4)

EU system.[76] However, legislative changes in medical device regulation via The Medical Devices (Amendment etc.) (EU Exit) Regulations 2020 are still in the drafting stage and therefore have no effect yet.[77] Furthermore, the MDR originally had set 26 May 2020 as the date by which older devices had to be changed to meet MDR requirements. However, in light of the Covid-19 pandemic, this has been extended a year to 26 May 2021.[78] As such, older UK medical devices which had not yet been approved under the MDR need not seek approval under that, but under the new MHRA regime. However, previously-approved devices under the MDR would also need to be approved by the MHRA. Since the situation is still developing, the approval criteria are not clear yet. However, it is clear that the MHRA will take over the regulatory functions for medical devices. Given the benefits of the robust yet fair MDR system, the UK Parliament would do well to make its new regulations similar to the MDR. This would also ease the transition of MDR-approved medical devices in the UK to be re-approved under the MHRA.

C. Data protection

Thirdly, there is an entire facet of data protection issues which are also caused by the opaque nature in which black box medical diagnostics operate and the fact that large volumes of past patient data have to be used in order to train the machines. This raises concerns about consent and transparency, as well as the ownership of one's medical data.[79] While that is a pertinent and interesting discussion in itself, it was not analysed in this paper as the main 'harm' in focus was injury to the person caused by misdiagnosis. There is very little interplay between these two risks or forms of harm and thus the choice was made to not explore data protection issues here. Furthermore, the scope of the dissertation had to be manageable enough to fit within the word limit. A proper and thorough evaluation of English Law's ability to deal with data protection issues would have involved delving into other areas of law like intellectual property and legislation such as the GDPR and the European Convention of Human Rights. Most academics who write on the subject such as George Bouchagiar have written entire papers on the

[76] GOV.UK, "Regulating Medical Devices from 1 January 2021" (1 September 2020) <https://www.gov.uk/guidance/regulating-medical-devices-from-1-january-2021> (accessed 2 November 2020).

[77] s. 1 The Medical Devices (Amendment Etc.) (EU Exit) Regulations 2020

[78] Geoffrey Cardoen, "Getting Ready for the New Regulations." (Public Health, European Commission, 22 October 2020) <https://ec.europa.eu/health/md_newregulations/getting_ready_en> (accessed 2 November 2020).

[79] Minssen (n 34)

subject.[80] Trying to cram it into this paper would not have done the issue justice and would have detracted from this paper's ability to fully explore its main issue.

D. Concluding remarks

In all, English Law's ability to deal with the oft-overstated risks associated with black box medicine stands on rather solid ground. The inherently flexible and adaptable medical negligence regime is supplemented by statutorily founded professional regulations for practitioners and hospitals, and a robust product regulation regime. These together provide multiple systems for promoting good and safe practice, attributing responsibility for the device's operation and the risk of malfunction, and if something does go wrong, compensating the aggrieved claimant (normally the patient harmed). It does all of this without being overly burdensome on manufacturers who want to expand into the market or hospitals and doctors who wish to use the equipment in diagnoses. In fact, it could be said that the English legal regime promotes the measured integration of black box diagnostic AI into the industry without sacrificing accountability. Furthermore, with more advances in technology and improvements to our understanding of black box AI, it is becoming more expectation than hope that they will be viewed as allies in our medical treatment, more akin to the benevolence of Vision than the malice of Ultron.

[80] George Bouchagiar, "The Long Road Toward Tracking the Trackers and De-biasing: A Consensus on Shaking the Black Box and Freeing From Bias" (2019) *Review of European Studies* 11(1), 27.