International Journal of Law and Information Technology, 2019, 27, 171–203
doi: 10.1093/ijlit/eaz004
Article

OXFORD

Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications

Daniel Schönberger*

ABSTRACT

Artificial intelligence (AI) is perceived as the most transformative technology of the 21st century. Healthcare has been identified as an early candidate to be revolutionized by AI technologies. Various clinical and patient-facing applications have already reached healthcare practice with the potential to ease the pressure on healthcare staff, bring down costs and ultimately improve the lives of patients. However, various concerns have been raised as regards the unique properties and risks inherent to AI technologies. This article aims at providing an early stage contribution with a holistic view on the 'decision-making' capacities of AI technologies. The possible ethical and legal ramifications will be discussed against the backdrop of the existing frameworks. I will conclude that the present structures are largely fit to deal with the challenges AI technologies are posing. In some areas, sector-specific revisions of the law may be advisable, particularly concerning non-discrimination and product liability.

KEYWORDS: Artificial intelligence, medical law and ethics, fairness, data protection, autonomy, accountability, negligence, liability, product liability

INTRODUCTION AND METHODOLOGY

Introduction

Artificial intelligence (AI) has left the realms of science-fiction. It is perceived as the most transformative technology of the 21st century and beyond with huge societal and economic potential. A virtual race among countries for leadership in the domain has begun and the big tech companies are investing billions into AI research. Among other major fields like mobility and energy[1] healthcare has been identified as an early candidate to be revolutionized by AI technologies.[2] Indeed, various clinical and

[*] lic. iur. (UZH), Attorney at Law, LLM in Medical Law and Ethics (Edinburgh). E-mail: daniel.m.schoenberger@gmail.com.

Declaration of conflicting interests: The author is Head of Legal Switzerland & Austria at Google Switzerland GmbH, Zurich. These are his own opinions and not necessarily those of his employer. Acknowledgments go to Dr Edward (Ted) Dove, University of Edinburgh, for his valuable input on early versions of this article.

[1] See eg, Partnership on AI to benefit people and society, tenets <https://www.partnershiponai.org/tenets/> accessed 9 April 2019.

[2] Institute for Public Policy Research (IPPR), Better health and care for all: A 10-point plan for the 2020s The Lord Darzi Review of Health and Care, final report , 15 June 2018 (The Lord Darzi Review, 2018), <https://www.ippr.org/research/publications/better-health-and-care-for-all> accessed 9 April 2019.

patient-facing applications[3] have already reached healthcare practice. This includes GP at Hand, an application powered by Babylon Health that assesses known symptoms and risk factors to provide informed medical information as well as insights to stay healthy,[4] Corti a technology that is used to manage and optimize the emergency dispatch process in Copenhagen,[5,6] a deep learning algorithm developed by Google for detection of diabetic retinopathy in retinal fundus photographs[7] or the image-related technologies developed by DeepMind together with Moorfields Eye Hospital[8] as published in Nature.[9] Arguably, these and other applications will ease the pressure on healthcare staff, bring down costs and ultimately improve the lives of patients.

But revolutions rarely come without side-effects. Various concerns have been raised as regards the unique properties and risks inherent to AI technologies. It has even been suggested that AI would introduce a novel element to the healthcare environments and relationships it is applied to that is apt to entirely transform them.[10] With the technologies advancing at a high pace, calls for regulation are getting louder. In April 2018, the European Commission (EC) released an AI strategy for Europe[11] and mandated a High-Level Expert Group on Artificial Intelligence to support its implementation *inter alia* with regard to the ethical and legal implications.[12] However, to date, very little ethical and legal work related to AI in healthcare exists. This article aims at providing an early stage contribution with a holistic view on what enables the mentioned applications: their 'decision-making' capacities.

For this purpose, I will first introduce the relevant technologies as well as their unique technical challenges. In the core part of this article, I will discuss the possible ethical and legal ramifications against the backdrop of the existing frameworks. I will conclude that the present structures are largely fit to deal with the challenges AI technologies are posing. In some areas, though, punctual and sector-specific revisions

[3] For further areas of application, see M Fenech, N Strukelj, and O Buston, *Ethical, Social, and Political Challenges of Artificial Intelligence in Health*, Future Advocacy, Wellcome Trust, April 2018, Table 1 (Future Advocacy, 2018), <https://wellcome.ac.uk/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf> accessed 9 April 2019.

[4] NHS, GP at Hand, <https://www.gpathand.nhs.uk/> accessed 9 April 2019.

[5] Corti, Artificial intelligence that saves lives, <http://www.corti.ai/> accessed 9 April 2019.

[6] A Peters, *Having a Heart Attack? This AI Helps Emergency Dispatchers Find Out* (Fast Company, 11 January 2018), <https://www.fastcompany.com/40515740/having-a-heart-attack-this-ai-helps-emergency-dispatchers-find-out> accessed 9 April 2019.

[7] V Gulshan et al, 'Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs' (2016) 316(22) JAMA 2402–10.

[8] Moorfields Eye Hospital NHS Foundation Trust, Moorfields announces research partnership, 3 July 2016, <https://www.moorfields.nhs.uk/news/moorfields-announces-research-partnership> accessed 9 April 2019.

[9] J De Fauw et al, 'Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease', *Nature Medicine*, 13 August 2018.

[10] Future Advocacy, 2018 (n 3) 25.

[11] European Commission, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial intelligence for Europe, Brussels, 25.4.2018 Com(2018) 237 final (EC Communication on AI, 2018).

[12] European Commission, High-Level Expert Group on Artificial Intelligence, 24 June 2018,<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence> accessed 9 April 2019.

of the law may be advisable, particularly concerning non-discrimination and product liability.

Methodology and research questions

The previous months have produced a wealth of public policy papers from governmental and non-governmental organizations not only with international reach on AI more broadly but also related to healthcare specifically, including the reports of the House of Lords Select Committee on Artificial Intelligence[13] and House of Commons Science and Technology Committee,[14,15] the Communication[16] from the European Commission and accompanying Staff Working Document,[17] the statement of the European Group on Ethics,[18] the draft ethics guidelines on trustworthy AI by the High-Level Expert Group on Artificial Intelligence,[19] the European Parliament's Report on Civil Law Rules on Robotics,[20] the French Villani Report,[21] the report of the US National Science and Technology Council,[22] as well as the reports, briefs and papers from the Royal Society,[23] Future Advocacy,[24]

[13] House of Lords, Select Committee on Artificial Intelligence, Report of Session 2017–19, AI in the UK: ready, willing and able?, 16 April 2018 (House of Lords, 2018), <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf> accessed 9 April 2019.

[14] House of Commons Science and Technology Committee, Algorithms in decision making, Fourth Report of Session 2017–19, 15 May 2018 (House of Commons, 2018), <https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf> accessed 9 April 2019.

[15] House of Commons, Science and Technology Committee, Robotics and artificial intelligence, Fifth Report of Session 2016–17, 13 September 2016 (House of Commons, 2016), <https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf> accessed 9 April 2019.

[16] EC Communication on AI (n 11).

[17] European Commission, Staff Working Document, Liability for emerging digital technologies Accompanying the document Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Artificial intelligence for Europe, Brussels, 25.4.2018 SWD(2018) 137 final (EC Staff Working Document on Liability, 2018).

[18] European Group on Ethics in Science and New Technologies Brussels, Statement on Artificial Intelligence, Robotics and Autonomous Systems, 9 March 2018 (European Group on Ethics, 2018), <http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf> accessed 9 April 2019.

[19] High-Level Expert Group on Artificial Intelligence, Draft Ethics Guidelines for Trustworthy AI, Brussels, 18 December 2018, <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai> accessed 9 April 2019.

[20] European Parliament, Committee on Legal Affairs, Report with recommendations to the Commission on Civil Law Rules on Robotics, Brussels, 27 January 2017 (2015/2103(INL)) (European Parliament, 2017).

[21] C Villani, 'For a Meaningful Artificial Intelligence - Towards a French and European Strategy', 8 March 2018 (Villani Report, 2018), <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf> accessed 9 April 2019.

[22] Executive Office of the President, National Science and Technology Council, Committee on Technology, Preparing for the future of artificial intelligence, 2016 (US National Science and Technology Council, 2016), <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf> accessed 9 April 2019.

[23] Royal Society, 'Machine Learning: The Power and Promise of Computers that Learn by Example', 2017 (Royal Society, 2017), <https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf> accessed 9 April 2019.

[24] Future Advocacy, 2018 (n 3).

Reform,[25] Nesta,[26] the Nuffield Council[27] and the Lord Darzi Review.[28] These documents and related scholarly papers and scientific articles (amounting to a total of approximately 300 documents) were reviewed. As a general pattern, a repeating set of epistemic and normative concerns with the properties of AI technologies could be identified, which will be outlined in the next part. Subsequently, the following questions will be asked and suggestions be made to address them: How do the identified issues fall within the existing ethical and legal frameworks? Is there need for action? What could such action(s) look like? Given the nature of the concerns the debate will be grouped as follows, where ethics will inform the law: (i) fairness and discrimination; (ii) autonomy and information/access rights and (iii) moral responsibility and liability.

The legal analysis will be conducted at a European level. UK law will be considered to the extent it adds any specifics or where the law is not harmonized (eg negligence). Questions around data sharing and data governance, safety and efficacy aspects of medical devices regulation as well as intellectual property all present opportunities for future work and will not be covered in this article.

THE TECHNOLOGIES AND RELATED ISSUES

AI, algorithms and machine learning

The term 'artificial intelligence' was originally coined by John McCarthy and others defined as the 'science and engineering of making intelligent machines'.[29,30] No universally accepted definition exists, though. Against the European background of this article, I will work with the definition as recently suggested by the EC that refers to AI as 'systems that display intelligent behaviour by analysing their environment and taking actions - with some degree of autonomy - to achieve specific goals'.[31] Although, over the course of this article, I will challenge in particular the EC's notion of 'autonomy'. AI can be 'purely software based' ('non-embedded'), or 'embedded' in hardware devices.[32] In this article, I will deal with *non*-embedded AI

[25] E Harwich and K Laycock, 'Thinking on its Own: AI in the NHS, Reform', January 2018, Figure 1: AI methods (Reform, 2018), <https://reform.uk/research/thinking-its-own-ai-nhs> accessed 9 April 2019.

[26] J Loder and L Nicholas, 'Confronting Dr Robot, Creating a People Powered Future for AI in Health', May 2018 (Nesta, 2018), <https://media.nesta.org.uk/documents/confronting_dr_robot.pdf> accessed 9 April 2019.

[27] Nuffield Council on Bioethics, Artificial intelligence (AI) in healthcare and research, May 2018, <http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf> accessed 9 April 2019.

[28] The Lord Darzi Review, 2018 (n 2).

[29] J McCarthy et al, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence', August 31, 1955, *AI Magazine*, Volume 27, Number 4 (2006).

[30] J McCarthy, 'What is Artificial Intelligence?' Stanford University, 12 November 2007, <http://www-formal.stanford.edu/jmc/whatisai.pdf> accessed 9 April 2019.

[31] EC Communication on AI, 2018 (n 11) N 1.

[32] ibid.

only. AI combines various different techniques, such as expert systems or natural language processing.[33] In public discourse, these techniques are often referred to as 'algorithms', a mathematical construct describing a 'control structure accomplishing a specific purpose under given provisions'.[34] Notwithstanding, as Mittelstadt and colleagues note, algorithms are commonly understood more broadly, including also their technological implementation and final application.[35] A lot of the recent progress is owed to the expert area of machine learning (ML) and supervised learning in particular.[36] Following the helpful explanation of Burrell, an ML algorithm consists of two distinct algorithms operating in parallel, a 'learner' and a 'classifier'.[37] The learner learns from normally large sets of labelled 'training data' dictating the so-called 'ground truth' (eg historical prostate screens labelled by human experts using Gleason Score[38] 1–5).[39] The patterns derived from the training data by the learner result in a matrix of corresponding weights and are continuously tweaked by the algorithm during its 'learning' until the results become stable.[40] The final 'model' is then used by the classifier to analyse new input data, also referred to as 'features' (eg pixels from new prostate screens to be diagnosed) and determine the output classification as the actual goal of the exercise (eg Gleason 4).[41] The output usually includes a probability score (e.g. 0.8) reflecting the confidence of the algorithm as regards this 'prediction'. So-called deep learning makes use of artificial neural networks, a collection of simple trainable mathematical units (neurons), which collaborate to compute a complicated function. The neurons are organized in layers, where each layer learns from the layer below it. Thus, going up the ladder the network learns patterns of patterns continuously increasing the 'semantic density' in the data.[42] This probabilistic way of transforming data into 'knowledge' as the basis for a 'prediction' is what I will refer to as 'decision-making' in the following.

Various technicalities during this decision-making process give raise for concerns that might have ethical and legal import. The existing literature has been most comprehensively mapped out by Mittelstadt and colleagues. Accordingly, the following issues may be discerned: biased training data, inconclusive correlations, intelligibility, inaccuracy and discriminatory outcomes.[43]

[33] See eg, the table in Reform, 2018 (n 25) 12.

[34] RK Hill, 'What an Algorithm Is' (2016) 29(1) Phil & Technol 35–59.

[35] B Mittelstadt et al, 'The Ethics of Algorithms: Mapping the Debate' (2016) Big Data & Society 3(2), 1–21.

[36] See eg T Mitchell, *Machine Learning*, 1st edn (McGraw-Hill Education 1997).

[37] J Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', Big Data & Society, January–June 2016, 1–12.

[38] A score that indicates the grade of prostate cancer, see Prostate Conditions Education Council, Gleason Score, <https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score> accessed 9 April 2019.

[39] Burrell (n 37).

[40] ibid.

[41] ibid.

[42] See eg, J Schmidhuber, 'Deep Learning in Neural Networks: An Overview', 30 April 2014, arXiv:1404.7828 [cs.NE].

[43] Mittelstadt et al (n 35).

Concerns identified with the decision-making capacity of AI systems

Biased training data

'Technology is not neutral'[44] and its design is always the result of many diverse choices.[45] Since ML so heavily relies on the data an algorithm is trained on, the major issues are grounded in flawed training data. Informally, this issue is known as 'garbage in garbage out'.[46] Thus, if the underlying datasets 'reflect existing biases against minorities' or other vulnerable groups prevalent in society, the algorithms will inadvertently adopt and reproduce them in their outputs.[47] As Hardt notes, since particularly race and gender are latent in virtually every aspect of our society AI systems might well discover even very subtle encodings.[48] A study published in Science demonstrated that a semantic model that was trained on standard text from the internet reproduced many existing stereotypes, eg by associating domestic terms more with women and unpleasant terms more with African-American than with European-American names.[49] Now, you would think that such associations can be prevented by simply excluding sensitive variables like race, age or gender from the model. However, non-sensitive attributes like postal codes can easily overlap with these variables and act as so-called 'proxies'.[50]

Also, ML classifiers usually improve with the volume of data and there is naturally 'proportionally less data available about minorities'.[51] Consequently, as Hardt affirms, classifiers considering smaller groups usually come out worse, and the model trained on the larger population will not translate reliably to minorities (so-called 'sample size disparity').[52]

Inconclusive correlations

Even if the training data is unproblematic, an algorithm may still exhibit bias due to inconclusive correlations. Algorithmic decision-making usually builds on 'inductive knowledge' derived from 'sufficiently strong *correlations*' and not necessarily on 'established causations'.[53] In some cases, though, such correlative reasoning might be socially unacceptable.[54] An oft-cited case is the ProPublica study on COMPAS, an AI-powered tool used within the US states justice system to predict the recidivism

[44] A commonplace in the philosophy of technology, see eg PP Verbeek, 'What Things do, Penn State Press' (2005), 39, 46, 136. <http://www.psupress.org/books/titles/0-271-02539-5.html> accessed 9 April 2019.

[45] See eg J Bryson, 'Adventures in NI, Three Very Different Sources of Bias in AI, and How to Fix them', 13 July 2017, <https://joanna-bryson.blogspot.com/2017/07/three-very-different-sources-of-bias-in.html> accessed 9 April 2019.

[46] See eg, Future Advocacy 2018 (n 3) 30.

[47] A Romei and S Ruggieri, 'A Multidisciplinary Survey on Discrimination Analysis' (2014) 29(5) Knowl Eng Rev 582–638.

[48] M Hardt, 'How Big Data is Unfair', 26 September 2014, <https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de> accessed 9 April 2019.

[49] A Caliskan, JJ Bryson, and A Narayanan, 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases' (2017) 356(6334) Science 183–186.

[50] Romei and Ruggieri (n 47).

[51] Hardt (n 48).

[52] ibid.

[53] M Hildebrandt and BJ Koops, 'The Challenges of Ambient Law and Legal Protection in the Profiling Era' (2010) 73(3) The Modern Law Review 432, 428–460.

[54] House of Commons, (n 14) 21.

risk of criminal defendants.[55] The core feature of this technology derives a set of risk scores from a questionnaire consisting of 137 questions.[56] This catalogue includes the question: 'In your neighbourhood, have some of your family or friends been crime victims?'[57] Clearly, even if a strong correlation between recidivism and a 'yes' to this question could be found, it would be indefensible to deny probation to someone because their friend was a crime victim.

Intelligibility

A further issue identified in the literature is the so called 'opacity' of AI systems and the difficulty of humans to understand them.[58] According to Burrell they are opaque in a sense that at the receiving end we do not have any concrete notion of how an algorithm arrived at the specific classification from any given input.[59] ML algorithms are hence often portrayed as 'black boxes'.[60,61] Burrell explains opacity in ML algorithms as a product of their 'high-dimensionality and complexity'.[62] Another reason is their emergent nature and behavioural modification due to the way they learn during operation as described above.[63] Hence, an often heard concern is that not even the designers of these systems are able to provide a human comprehensible representation of them.[64] This provokes Tutt to note that the 'properties that make AI systems valuable also make them unpredictable and inherently hazardous'.[65] Because algorithmic decision-making can have life changing import, discussions have arisen around the need of 'intelligibility' of AI systems.[66]

Inaccuracy

In light of their probabilistic nature, ML models must be expected to fail in certain instances.[67] In a study published by Stanford researchers in Nature the classification of skin lesions to single out skin cancer was demonstrated to be on par with dermatologists.[68] However, no technology is perfect and tests of image classifiers show that

[55] J Angwin et al, 'Machine Bias', *Pro Publica*, 23 May 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 9 April 2019.

[56] ibid.

[57] COMPAS 'CORE' risk and needs assessment, <https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html> accessed 9 April 2019.

[58] Z Lipton, 'The Mythos of Model Interpretability', 6 March 2017, arXiv:1606.03490 [cs.LG].

[59] Burrell (n 37).

[60] Mittelstadt et al (n 35) 6.

[61] Often attributed to F Pasquale, *The Black Box Society* (Harvard University Press 2015).

[62] Burrell (n 37).

[63] A Markowetz et al, 'Psycho-Informatics: Big Data Shaping Modern Psychometrics' (2014) 82(4) Medical Hypotheses 405–11.

[64] Mittelstadt et al (n 35) 6.

[65] A Tutt, 'An FDA for Algorithms' (2017) 69 Admin L Rev 83.

[66] See the terminology used in House of Lords (n 13) 91.

[67] L Edwards and M Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably not the Remedy you are Looking for' (2017) 16 Duke L Technol Rev 18, 54.

[68] A Esteva et al, 'Dermatologist-level Classification of Skin Cancer with Deep Neural Networks' (2017) 542 Nature 115–18.

systems still confuse cats with frogs.[69] Obviously, a similar mistake in cancer diagnosis could have far more severe outcomes.

Inaccuracy can also be the direct consequence of bias. A recent study showed that the error rate of commercial facial-analysis programs on black women due to sample size disparity in one case were over 34 per cent (compared with 0.8 per cent on white males).[70,71] The textbook example, however, has become Google's labelling of black people as 'gorillas' in their Photos app.[72]

Finally, inaccuracy can occur, although technically the prediction is correct, but where model and outcome are inaccurate representations of the real-world equivalents. A study of 2015 analyses an ML research project to predict the risk of hospital attendants to develop pneumonia.[73] While its general accuracy was high, it instructed doctors to send patients with asthma home, although such patients are generally regarded as high risk pneumonia candidates. The study revealed that patients with a history of asthma historically were not only admitted to the hospital but also directly taken to the intensive care unit. Thus, the files of such patients rarely appeared within the 'requires further care' data the model was trained on and the algorithm consequently classified them as low risk.[74,75]

Unfair outcomes

In certain cases, algorithmic decisions can have disproportionate effects on a protected class even if based on unobjectionable knowledge.[76] Computer scientists speak of 'unequal ground truth',[77] where the 'best and formally accurate approximation of reality' is simply unfair.[78] To that end, the ProPublica study cited above also revealed that the likelihood for black defendants to be labeled as 'high risk' but not actually re-offend was twice as high as for white defendants. To make things worse, for white defendants the tool reversely attributed 'low risk' scores twice as often as for blacks, although they actually *did* re-offend.[79] The tool was obviously biased against black people with deleterious outcomes for black defendants' sentences. This

[69] J Wexler, 'Google AI Blog, Facets: An Open Source Visualization Tool for Machine Learning Training Data' 17 July 2017, <https://ai.googleblog.com/2017/07/facets-open-source-visualization-tool.html> accessed 9 April 2019.

[70] J Buolamwini and T Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 Proceedings of Machine Learning Research 1–15.

[71] L Hardesty, 'Study finds Gender and Skin-type Bias in Commercial Artificial-intelligence Systems', *MIT News*, 11 February 2018, <http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212> accessed 9 April 2019.

[72] T Simonite, 'When it Comes to Gorillas, Google Photos Remains Blind', *Wired*, 11 January 2018, <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/> accessed 9 April 2019.

[73] Caruana R et al, 'Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission' (2015) *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ACM 1721–30.

[74] ibid.

[75] K Crawford and R Calo, 'There is a Blind Spot in AI Research' (2016) 538 Nature 311–13.

[76] Mittelstadt et al (n 35) 5.

[77] ibid.

[78] S Shalev-Shwartz and S Ben-David, *Understanding Machine Learning, From Theory to Algorithms* (CUP 2014) 265.

[79] Pro Publica (n 56).

happened, because obviously the risk score as the target variable for some reason 'significantly correlated with race'.[80] This phenomenon referred to as 'redundant encoding' 'statistically stereotyped' black defendants.[81,82]

This part explained the relevant technologies and mapped out the potential issues with their decision-making capacities. In the core part of this article, I will now test how these issues fall within the existing ethical and legal frameworks.

ETHICAL AND LEGAL ASSESSMENT

Fairness and discrimination

The challenge in a healthcare context in particular

It is often advanced that algorithmic decision-making might lead to fairer and more inclusive outcomes than human judgment or decisions 'based on ad hoc rules'.[83] In a semantically neutral sense, 'algorithms are designed to discriminate' and to that end need to 'give weight to some factors over others'.[84] But as outlined above, the properties of ML systems bear the risk to reflect and exacerbate existing bias, which might unfairly affect members of protected groups based on sensitive categories like gender, race, age, sexual orientation, ability or belief.

To date, few cases have been described in the literature related to AI fairness in a specific healthcare context. In a recent Nature article, however, Zou and Schiebinger discuss the groundbreaking work of Esteva[85] and colleagues that used ML to detect skin cancer. It is highlighted that fewer than 5 per cent of the images this model was trained on were from individuals with dark skin.[86] Given the issues described above this seems problematic. Indeed, it is safe to assume that medical AI applications are especially susceptible to bias and discrimination. Rajkomar and colleagues discern four categories of possible bias in healthcare, bias in model design, in training data, in interactions with clinicians and in interactions with patients.[87] Vayena and colleagues particularly emphasize cases in which the data sources themselves do not reflect true epidemiology within a given demographic,[88] such as population data biased by the entrenched overdiagnosis of schizophrenia in African Americans.[89] Arguably, the most relevant case in this context relates to the issue of sample size disparity, where there is not enough data on a particular group, as outlined above. Indeed, as

[80] See for the problem, eg Royal Society (n 23) 114.

[81] T Calders and I Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures', in B Custers et al (eds), *Discrimination & Privacy in the Information Society* (SAPERE 2013) 43–57.

[82] See Royal Society (n 23) 92.

[83] See eg, Google, 'Responsible AI Practices', *Fairness*, <https://ai.google/education/responsible-ai-practices?category=fairness> accessed 9 April 2019.

[84] House of Commons (n 14) 32.

[85] Esteva et al (n 68).

[86] J Zou and L Schiebinger, 'AI can be Sexist and Racist — it's time to make it Fair' (2018) 559 Nature 324–6.

[87] A Rajkomar et al, 'Ensuring Fairness in Machine Learning to Advance Health Equity' (2018) 169 Ann Intern Med 866–72.

[88] E Vayena, A Blasimme, and IG Cohen, 'Machine learning in Medicine: Addressing Ethical Challenges' (2018) 15(11) PLoS Med e1002689.

[89] HW Neighbors et al, 'The Influence of Racial Factors on Psychiatric Diagnosis: A Review and Suggestions for Research' (1989) 25 Community Ment Health J 301–11.

the literature shows, major health inequalities notoriously not only persist across but also within countries,[90,91] tightly intertwined with social inequalities.[92] There is a 36-year gap in life expectancy between the poorest and the richest countries in the world,[93] but even within the city of London men in the richest parts on average live 18 years longer than men in the most deprived neighbourhoods.[94] There are ten times less physicians in low-income countries than in high-income countries and in countries with non-state funded healthcare systems costs are often prohibitive.[95] In many parts of the world, persistent gender inequalities limit women's access to healthcare.[96] Disparities in access to healthcare between urban and rural areas also exist in developed and high-income countries.[97] Stigma related to mental-illness,[98] addiction,[99] certain disease like HIV,[100] sexual preference or gender identity,[101] poverty,[102] as well as 'internalised stigma' in ethnic minorities[103] serve as further barriers to access. Literature and government data also suggest that large populations are underrepresented in clinical trials data, which seem to favour predominantly competent adult white men.[104,105,106] Clinical trials data on pregnant women is missing almost entirely.[107,108]

[90] World Health Organization, 'Monitoring Health Inequality', 2015, <http://apps.who.int/iris/bitstream/handle/10665/133849/WHO_FWC_GER_2014.1_eng.pdf?sequence=1> accessed 9 April 2019.

[91] OECD, Health Inequalities, <http://www.oecd.org/health/inequalities-in-health.htm> accessed 9 April 2019.

[92] R Hart, 'If you're not a White Male, Artificial Intelligence's Use in Healthcare could be Dangerous', 20 July 2017, <https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/> accessed 9 April 2019.

[93] World Health Organization, *World Health Statistics 2011* (World Health Statistics 2011), <http://www.who.int/whosis/whostat/2011/en/> accessed 9 April 2019.

[94] M Marmot, 'Social Justice, Epidemiology and Health Inequalities' (2017) 32(7) Eur J Epidemiol 537–46.

[95] World Health Statistics (n 93).

[96] G Mariani et al, 'Improving Women's Health in Low-income and Middle-income Countries. Part I: Challenges and Priorities' (2017) 38(12) Nucl Med Commun 1019–23.

[97] E Merwin, A Snyder and E Katz, 'Differential Access to Quality Rural Healthcare: Professional and Policy Challenges' (2006) 29(3) Fam Community Health 186–94.

[98] C Henderson et al, 'Mental Health-related Stigma in Health Care and Mental Health-Care Settings' (2014) 1(6) Lancet 467–82.

[99] MA Brondani et al (ed), 'Stigma of Addiction and Mental Illness in Healthcare: The Case of Patients' Experiences in Dental Settings' (2017) 12(5) PLoS One e0177388.

[100] TA Crowell et al, 'Stigma, Access to Healthcare, and HIV Risks among Men who Sell Sex to Men in Nigeria' (2017) 20(1) J Int AIDS Soc 21489.

[101] A Müller, 'Health for All? Sexual Orientation, Gender Identity, and the Implementation of the Right to Access to Health Care in South Africa' (2016) 18(2) Health Hum Rights 195–208.

[102] H Allen, 'The Role of Stigma in Access to Health Care for the Poor' (2014) 92(2) Milbank Q 289–318.

[103] J Owuor and J Nake, 'Race Equality Foundation, Internalised Stigma as a Barrier to Access to Health and Social Care Services by Minority Ethnic Groups in the UK', May 2015, <https://raceequalityfoundation.org.uk/wp-content/uploads/2018/02/Health-Briefing-36_1.pdf> accessed 9 April 2019.

[104] Hart (n 92).

[105] U.S. Food and Drug Administration, FDA Report, Collection, Analysis, and Availability of Demographic Subgroup Data for FDA-Approved Medical Products, August 2013, <https://www.fda.gov/downloads/RegulatoryInformation/LawsEnforcedbyFDA/SignificantAmendmentstotheFDCAct/FDASIA/UCM365544.pdf> accessed 9 April 2019.

[106] U.S. Census Bureau, Quick facts, <https://www.census.gov/quickfacts/fact/table/US/PST045217> accessed 9 April 2019.

[107] R van der Graaf et al, 'Fair Inclusion of Pregnant Women in Clinical Trials: An Integrated Scientific and Ethical Approach' (2018) 19 Trials 78.

[108] Hart (n 92).

Evidently, this overview does not claim completeness. However, it still supports the argument that historical health data accommodates underrepresentation of and bias against large populations. Naturally, where minorities and even whole populations are excluded from health services, no health records of them exist. The deterrent effects of the so called 'digital divide' in health have long been documented.[109,110] Unless health records are digitized, such data will remain excluded from any future AI development.[111] States and individuals who cannot afford the necessary technologies or do not have the required 'digital literacy' will stay at the sidelines.

Fairness

Literally, all public policy papers reviewed for this article qualify potential bias in and discrimination by AI systems as a major ethical concern[112] that might affect the access to as well as the results of healthcare.[113] As Hardt holds, accuracy in automated decisions seems to be a strong indicator for 'fairness'.[114] Hence, if a classifier is disproportionally inaccurate on minorities, the decision-making is unfair towards these groups. Sensitive criteria like gender, age, race and sexual orientation do not provide legitimate reasons to deviate from a formal understanding of justice in the sense of equal treatment; rather they suggest the need for special protection. Given the properties of AI systems, existing biases in healthcare might be deeply baked into the technologies that are designed to play a central role in future care, which could exacerbate social inequalities. 'Feedback loops' might perpetuate existing stigmatization and contribute to 'self-fulfilling prophecies'.[115,116] For example, as advanced by Char and colleagues, given the tendency to withdraw care in cases of extreme prematurity or brain-damage algorithms could conclude that these situations are always fatal and adjust their predictions accordingly with obviously lethal consequences for the concerned patients.[117]

From a purely utilitarian point of view, it could be argued that as long as the general population and health systems benefit overall even inherently biased systems should not be withheld from clinical practice. However, few would agree that this is a society they would want to live in, where the interests of minorities are simply rolled over. After all distributive justice requires a 'fair, equitable and proportionate' distribution of the benefits and burdens any new technology entails.[118] Accordingly,

[109] United Nations, Bridging the Digital Divide in Health, UN Chronicles, Vol XLVIII No 3, 2011, <https://unchronicle.un.org/article/bridging-digital-divide-health> accessed 9 April 2019.

[110] L López et al, 'Bridging the Digital Divide in Health Care: The Role of Health Information Technology in Addressing Racial and Ethnic Disparities' (2011) 37(10) Comm J Qual Patient Saf 437–45.

[111] Also, see Reform (n 25).

[112] See eg, Future Advocacy (n 3) 30.

[113] ibid.

[114] Hardt (n 48).

[115] M Leese, 'The New Profiling: Algorithms, Black Boxes, and the Failure of Antidiscriminatory Safeguards in the European Union' (2014) 45(5) Security Dialogue 494–511.

[116] Mittelstadt et al (n 35) 9.

[117] DS Char, NH Shah, and D Magnus, 'Implementing Machine Learning in Health Care - Addressing Ethical Challenges' (2018) 378 N Engl J Med 981–3.

[118] See eg TL Beauchamp and JF Childress, *Principles of Biomedical Ethics*, 5th edn (OUP 2001), 226.

AI systems in healthcare should not be launched unless they warrant a sufficient degree of equal opportunity especially for minorities and are aligned with the needs and requirements of the public health system. In any case, as Hart (distinct from Hardt) postulates, it should be avoided that the systems our future health might depend on are early and irrevocably 'calibrated for younger, more urban bodies'[119] or the already privileged classes of the Western world more generally.

While these seem to be straightforward claims in terms of bioethics it must be conceded that no universal definition of 'fairness' under any given circumstance exists. This makes a translation into computational parameters hard. Wattenberg and colleagues demonstrate this by example of so-called 'threshold classifiers' making a yes/no decision about loan-grants.[120] There are indeed various possible thresholds that could be regarded as 'fair' taking into account the default risks and profit interests of the decision-maker, the overall distribution of granted loans as well as the distribution relative to different groups.[121] Similar ethical *dilemmata* could be faced by care institutions within fixed-tariff systems on the basis of so-called 'diagnosis related groups (DRGs)'.[122] In such a system, the decision when to release a patient from the hospital is informed by various parameters like quality and safety of care, the job satisfaction of the care personnel, economic interests of the care institution and insurers as well as the interests of vulnerable subgroups.[123] Existing attempts to map the related ethical issues are scarce and have been described as challenging.[124] This seems an area where a cautious approach to algorithmic support could be beneficial. AI-powered modelling based on patients' electronic health records has already demonstrated to predict multiple medical conditions.[125] This could include features giving account to the interests just outlined above to predict a release date for patients. Under aspects of fairness a just distribution among different DRGs as well as vulnerable subgroups would have to be considered. While this is no trivial task, existing research suggests methods for 'equality of opportunity' that shifts the 'burden of statistical uncertainty' away from the concerned groups to the decision-maker in order to incentivize them to 'find more direct features and optimise classification accuracy'.[126] Also, in their latest work Rajkomar and colleagues recommend to incorporate principles of distributive justice into model design, deployment and evaluation to ensure equality in patient outcomes, performance and resource allocation.[127]

[119] See Hart (n 92).

[120] M Wattenberg, F Viégas, and M Hardt, 'Attacking Discrimination with Smarter Machine Learning' <https://research.google.com/bigpicture/attacking-discrimination-in-ml/> accessed 9 April 2019.

[121] ibid.

[122] See eg, for Switzerland in Swiss Federal Office of Public Health, SwissDRG tariff system, <https://www.bag.admin.ch/bag/en/home/versicherungen/krankenversicherung/krankenversicherung-leistungen-tarife/Spitalbehandlung/Tarifsystem-SwissDRG.html> accessed 9 April 2019.

[123] V Wild et al, 'Assessing the Impact of DRGs on Patient Care and Professional Practice in Switzerland (IDoC) – a Potential Model for Monitoring and Evaluating Healthcare Reform' *Swiss Med Wkly* 2015;145:w14034, 9 February 2015.

[124] ibid.

[125] A Rajkomar et al, 'Scalable and Accurate Deep Learning with Electronic Health Records' (2018) 1 npj Digital Medicine 18.

[126] M Hardt, E Price, and N Srebro, 'Equality of Opportunity in Supervised Learning', 7 October 2016 arXiv:1610.02413 [cs.LG].

[127] Rajkomar et al (n 87).

A mapping of non-discrimination laws in healthcare

Non-discrimination is one of the fundamental values of the EU, enshrined in Article 2 of the Treaty on European Union (TEU), Article 10 of the Treaty on the Functioning of the European Union as well as Article 21 of the EU Charter of Fundamental Rights (corresponding to Article 14 of the European Convention on Human Rights). However, given the EU's general lack of authority to legislate in healthcare, individuals will mainly have to rely on the Racial Equality Directive[128] and the Goods and Services Directive.[129] The latter only applies to health services that are provided against remuneration, but this includes instances where such services are paid via a compulsory insurance scheme.[130,131] Publicly funded health services provided under statutory social security schemes like in the UK fall within the scope of the Social Security Directive.[132]

According to Schedule 19 of the UK Equality Act 2010, any NHS Foundation Trust[133] is subject to the public sector equality duties (PSED) as defined in section 149 of the Equality Act 2010. This includes the avoidance of discrimination and advancement of equality of opportunity with regard to the protected characteristics age, disability, gender reassignment, pregnancy and maternity, race, religion or belief, sex and sexual orientation.[134,135,136] Otherwise, the Act is materially in line with the mentioned EU Equality Directives.

Discrimination laws and AI

The legal protections cited in the previous subsection cover direct and indirect discrimination[137] and intentionality or actual harm are no constitutive elements.[138] Direct discrimination, that is a less favourable treatment compared to others merely on the basis of protected characteristics, puts the focus on the individual.[139] However, per definition an AI model is about statistical patterns and never about individuals. Thus, direct discrimination in an AI context will rarely be the case and

[128] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, art 3(1)(e).

[129] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, Recital 12.

[130] See eg, ECJ, C-157/99 Smits and Peerbooms, 12 July 2001, para 57.

[131] See also, European Commission, Report on the application of Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services, Brussels, 5.5.2015, COM(2015) 190 final.

[132] Council Directive of 19 December 1978 on the progressive implementation of the principle of equal treatment for men and women in matters of social security (79/7/EEC), art 3.

[133] Within the meaning of s 30 of the National Health Service Act 2006.

[134] s 149(1)(a) and (b) of the Equality Act 2010.

[135] s149(7) of the Equality Act 2010.

[136] NHS, The NHS Constitution for England, updated 14 October 2015, <https://www.gov.uk/government/publications/the-nhs-constitution-for-england/the-nhs-constitution-for-england> accessed 9 April 2019.

[137] See art 2 of the Racial Equality Directive; art 2 of the Goods and Services Directive; art 4 of the Social Security Directive; ss 13, 14 and 19 in conjunction with s 149(8) of the Equality Act 2010.

[138] See also G Di Federico, 'Access to Healthcare in the European Union: Are EU Patients (Effectively) Protected Against Discriminatory Practices?', in L Rossi and F Casolari (eds), *The Principle of Equality in EU Law* (Springer 2017), 229–53, 237.

[139] P Craig and G De Búrca, *EU Law: Text, Cases and Materials* (OUP 2011) 918.

only be present, if a protected characteristic was used as a determining feature to differentiate a member of a class sharing that characteristic.[140] In contrast, under the Racial Equality and the Goods and Services Directives as well as the Equality Act 2010 discriminatory decisions that are mediated by an 'apparently neutral provision, criterion or practice' but still disproportionately affect certain protected groups are qualified as *indirect* discriminations.[141] In an AI context, quite obviously, the algorithm *itself* constitutes an 'apparently neutral criterion'. Thus, to the extent such criterion is indeed *not* neutral but perpetuates bias, 'algorithmically mediated discrimination' will generally fall within the scope of indirect discrimination.[142,143] Notwithstanding the foregoing, such application, having a disparate effect on a protected class, does *not* qualify as an indirect discrimination, if that 'practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary'.[144] This equates to a 'four-steps-proportionality test' comprising the elements of a 'legitimate aim', 'suitability, necessity and proportionality *stricto sensu*'.[145]

In the first and so far, only comprehensive paper that discusses algorithmic fairness against the backdrop of European non-discrimination laws, Hacker concludes that the first three steps of the test are taken with relatively little effort.[146] I would agree. As per the first element, arguably the predictive task of the algorithmic application, eg the allocation of transplants or any image classification in mammography, provides for the required 'legitimate aim'.[147] Concerning 'suitability' or also 'effectiveness', it seems correct to equate this second requirement with predictive accuracy. This is likely met by demonstrating a high grade of a classifier's overall accuracy.[148] The 'necessity' or also 'subsidiarity' element of the test requires that no similarly suitable but less invasive practice is available. Since it is exactly the 'superior effectiveness' that makes the adoption of algorithmic decision-making appealing,[149] also this step may be deemed as passed.

Thus, ultimately the test boils down to a proportionality test *stricto sensu*. Practices pursuing a legitimate aim that are suitable and necessary can still fail the overall test, if the aim is not compelling enough or the disparate impact on the protected group is in too harsh a contradiction with the values of non-discrimination

[140] See S Barocas and A Selbst, 'Big Data's Disparate Impact' (2016) 104 Calif L Rev 671–732, 694.

[141] art 2(b) of the Racial Equality Directive; art 2(b) of the Goods and Services Directive; s 19(1) and (2) of the Equality Act 2010.

[142] P Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law', Common Market Law Review (forthcoming) 10, <https://www.academia.edu/36494567/Teaching_Fairness_to_Artificial_Intelligence_Existing_and_Novel_Strategies_against_Algorithmic_Discrimination_under_EU_Law_Common_Market_Law_Review_forthcoming_> accessed 9 April 2019.

[143] I Žliobaitė, 'Measuring Discrimination in Algorithmic Decision Making' (2017) 31(4) Data Mining & Knowledge Discovery 1060–89.

[144] art 2(b) of the Racial Equality Directive; art 2(b) of the Goods and Services Directive; see also the materially similar provision in s 19(1) and (1) of the UK Equality Act 2010.

[145] Craig and De Búrca (n 139) 526.

[146] Hacker (n 142).

[147] ibid 17.

[148] ibid 18.

[149] ibid.

laws.[150,151] In the *Bilka and Cadman* cases, the CJEU adopted a 'structured approach' relying on a test of 'a real need on the part of the undertaking'.[152] A preferable notion of 'proportionality in the narrower sense' was articulated by the UK Supreme Court in *Homer*, holding that the 'assessment of whether the criterion can be justified entails a comparison of the impact of that criterion upon the affected group as against the importance of the aim to the [decision-maker]'.[153] To that end, it is particularly the point of 'overall-accuracy' that must be challenged. Indeed, as Hardt demonstrates a top ranking AI system with 95 per cent accuracy could mean that the 5 per cent error rate is evenly distributed over all participating groups, reflecting the uncertainty intrinsic to any inference based on historical data.[154] Or it could mean that the classifier shows perfect accuracy on the dominant population but is 50 per cent prone to error on minorities,[155] which would be inacceptable.

The issues around AI intelligibility might even exacerbate the disadvantages on minorities, when it comes to enforcement.[156] The law generally acknowledges the inherent evidence problems with discrimination cases by reversing the burden of proof once a *prima facie* case for an indirect discrimination could be established.[157] Hence, it is sufficient for a claimant to make plausible a discriminatory practice and it is up to the decision-maker to provide evidence for its legitimacy. However, without precise knowledge of the training data, the features and their relative weight in the model, a claimant will have difficulties to refute the assertions of the decision-maker, especially if they argue with convincing overall accuracy.[158,159] There is scarce CJEU case law suggesting that the opacity of the system and the refusal or inability to disclose the underlying factors shall be construed against the decision-maker.[160,161] But given the peculiarities of these cases, much legal uncertainty remains. Additionally, it must be remembered that under the EU Equality Directives—unlike the UK Equality Act 2010—the protected characteristics are limited to race and gender. Given AI's capabilities to detect even well-hidden subtleties many other groups potentially marginalized in healthcare are left exposed.

[150] L Hoffmann, 'The Influence of the European Principle of Proportionality upon UK Law', in E Ellis (ed), *About The Principle of Proportionality in the Laws of Europe* (Hart Publishing 1999) 107.

[151] I Smith and A Baker, *Smith & Wood's Employment Law* (OUP 2015) 347.

[152] CJEU, Case C-170/84, *Bilka - Kaufhaus GmbH v Karin Weber von Hartz*, 13 May 1986, para 37; CJEU, Case C-17/05, *BF Cadman v Health & Safety Executive*, 3 October 2006, para 32.

[153] *Homer (Appellant) v Chief Constable of West Yorkshire Police (Respondent)*, [2012] UKSC 15, on appeal from: [2010] EWCA Civ 419, para 24.

[154] Hardt (n 48).

[155] ibid.

[156] See Hacker (n 142) 22.

[157] art 8 of the Race Equality Directive and art 9 of the Goods and Services Directive; s 136 of the Equality Act 2010.

[158] J Grimmelmann and D Westreich, 'Incomprehensible Discrimination' (2017) 7 Calif L Rev Online 164–77, 168.

[159] Hacker (n 142) 23.

[160] CJEU, Case C-109/88, Handels- og Kontorfunktionærernes Forbund i Danmark v Dansk Arbejdsgiverforening, acting on behalf of Danfoss, 17 October 1989, paras 15 and 16.

[161] CJEU, Case C-415/10, *Meister v Speech Design*, 19 April 2012, para 42.

Having said all this, arguably only unreasonable efforts or infeasibility should relieve a decision-maker from applying an unbiased model.[162,163] However, in a healthcare context, exactly this might be the case, as available medical training data is very sparse. Revisiting the 'skin cancer example' cited above (Section 'The challenge in a healthcare context in particular'), if there are simply no images of lesions on dark skin available, the system may not be trained accordingly and a potential bias against darker skinned individuals due to underrepresentation seems unavoidable. Also, in their influential paper, Barocas and Selbst reveal the uncomfortable truth that in our society the 'relevant attributes which *should* be considered' in an objective decision-making process (eg the targeting of individuals susceptible to alcohol or tobacco addiction within a preventive health campaign), are 'meaningfully shaped by sensitive attributes' (eg the growing up in an underprivileged area that is often correlated with a particular ethnic or educational background).[164] Consequently, computer scientists have demonstrated that removing the problematic correlations with the sensitive category (eg ethnicity) or the respective proxy (eg postal code) will usually come at a 'significant cost in accuracy'.[165,166] How would you want to reach vulnerable groups, if you cannot identify them? Obviously, an inaccurate classifier—possibly reduced to arbitrariness—will likewise raise concerns under non-discrimination laws. Hence, if algorithmic decision-making is not to be banned by non-discrimination laws in certain sensitive areas like healthcare, society will have to agree on a workable 'trade-off between fairness and utility'.[167]

Further discussion

On the more positive side, 'ML fairness' is a heavily researched field[168,169] and approaches like 'fairness through awareness'[170,171] all place the burden to find a just allocation between all interests at stake on the decision-maker. However, any engineering attempt to 'put design- and policy constraints on an algorithm' to prevent disparate effects on protected groups necessarily entails that the sensitive characteristics are accurately recorded.[172] And as we have seen above, differentiation based on these criteria might even amount to *direct* discrimination. Arguably, this should not be the case, if sensitive variables are used to *avoid* discriminatory outcomes. However, the trade-off between accuracy and fairness described above always entails the probability to get the threshold slightly wrong. This is a litigation risk an AI developer may not

[162] Hacker (n 142) 18.

[163] Barocas and Selbst (n 140) 710.

[164] ibid.

[165] T Calders and S Verwer, 'Three Naive Bayes Approaches for Discrimination-free Classification' (2010) 21(2) Data Mining and Knowledge Discovery 277–92.

[166] Also see, Grimmelmann and Westreich (n 158).

[167] C Dwork et al, *Fairness Through Awareness*, 20 April 2011, arXiv:1104.3913.

[168] See eg Zou and Schiebinger (n 86).

[169] See eg Google, Machine Learning Fairness <https://developers.google.com/machine-learning/fairness-overview/> accessed 9 April 2019.

[170] Dwork et al (n 167).

[171] Also see M Feldman et al, 'Certifying and Removing Disparate Impact' *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), 259–68.

[172] Zou and Schiebinger (n 86).

want to face. Additionally, the processing of sensitive criteria enjoys special protection under data protection laws (see below Section 'Information and access rights').[173] These legal uncertainties might impede responsible development striving for fairness and the legislator should clarify the situation or introduce exemptions to remove these obstacles as soon as possible.

Notwithstanding the foregoing, the awareness within the research community is growing and the industry seems invested in an ethical advancement of their innovations as expressed by various codes of conduct[174] as well as initiatives like the Partnership on AI[175] or Google's People + AI Research that propagate a people centred approach.[176] Also, a focus on a diverse and multidisciplinary workforce may help with detection of potential bias at all stages of development and operation. Finally, as much as an inconsiderate application of AI could reflect or exacerbate existing bias, a prudent deployment can also help to *detect* bias in society and help address it.[177] The Geena Davis Institute for example used ML to demonstrate how little females are seen and heard in major movies, empowering women in their ambitions and career opportunities.[178] The ultimate test, though, should be, if AI on balance puts protected groups in a better position compared with the *status quo*, that is being subjected to conventional forms of bias that have been persisting within society for many decades.

We will now turn to the question to what extent individuals need to understand how an AI system arrived at a certain prediction in order for them to make autonomous and informed decisions.

Respect for autonomy, information and access rights

Respect for autonomy in the age of AI

There is an ambiguity around the use of the term 'autonomy' in the context of AI and so-called 'autonomous systems', which prompted the European Group on Ethics in Science and New Technologies (European Group on Ethics) to render a clarifying statement. They remind of the fact that 'autonomy' is a philosophical term rooted in human dignity that describes the intrinsically human capacity of individuals to 'legislate for themselves, to formulate, think and choose norms, rules and laws for themselves to follow'.[179] This essentially follows the notion of autonomy in biomedical ethics as normally attributed to Beauchamp and Childress.[180] Consequently, the

[173] See articles 9 and 22(4) of the General Data Protection Regulation (GDPR).

[174] See eg A Winfield, 'A Round Up of Robotics and AI ethics: Part 1 Principles', 23 December 2017, <http://alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html> accessed 9 April 2019.

[175] Partnership on AI, <https://www.partnershiponai.org/> accessed 9 April 2019.

[176] Google, People + AI Research, <https://ai.google/research/teams/brain/pair> accessed 9 April 2019.

[177] See eg, J Savulescu and H Maslen, 'Moral Enhancement and Artificial Intelligence: Moral AI?', in J Romportl, E Zackova, and J Kelemen (eds), *Beyond Artificial Intelligence, Topics in Intelligent Engineering and Informatics*, vol 9 (Springer, Cham 2015).

[178] Geena Davis Institute, Geena Davis Inclusion Quotient, The Reel Truth: Women Aren't Seen or Heard, An Automated Analysis of Gender Representation in Popular Films, 2017, <https://seejane.org/wp-content/uploads/gdiq-reel-truth-women-arent-seen-or-heard-automated-analysis.pdf> accessed 9 April 2019.

[179] European Group on Ethics (n 18) 9.

[180] Beauchamp and Childress (n 118) 57.

European Group on Ethics holds that ‘it is not appropriate to manage and decide about humans in the way we manage and decide about objects or data’.[181]

While ‘full autonomy’ is an ideal that will rarely be present, the principle of respect for autonomy still requires ‘substantial autonomy’, which entails intentionality, understanding and the absence of coercing influences that could unduly affect a person’s action.[182] Given the scope of this article, I will focus on the aspect of ‘understanding’ only, which might be directly impacted by the issue around ‘intelligibility’ of AI systems. In a healthcare relationship, respect for a patient’s autonomy is honoured by so-called ‘informed consent’. Ethically, the concept is an expression of self-determination encompassing that individuals must be given the opportunity to agree to and make choices between risks they are exposed to.[183] According to the UK Supreme Court’s ruling in *Montgomery*, this mandates that patients are informed of the ‘nature, risks, consequences and alternatives’ associated with a specific medical intervention.[184,185] Hence, the right of individuals to information disclosure is not an end in itself but always related to the prevalence of any ‘material risks’ that in the eyes of a reasonable patient would be regarded as significant for their decision-making.[186] Additionally, the provided information needs to be ‘empowering’, which involves a ‘meaningful dialogue’ between doctor and patient.[187] According to the UK Supreme Court, ‘bombarding the patient with technical information which she cannot reasonably be expected to grasp’ would be in conflict with this notion.[188]

Now, does the application of AI change anything to this? To begin with, the right to information disclosure remains contextual and depends on the materiality of the risks involved. The Royal College of Radiologists, for example, regards the risks for patients with conventional radiology as ‘very low’.[189] There are no reasons to believe that the introduction of ML algorithms in clinical radiology or oncology would alter this assessment. Certainly, the recent evidence provided to the House of Lords Select Committee does not raise any concerns about an increased risk exposure of patients.[190] Against the backdrop of the above cited case law, even AI applications in riskier areas would not add anything novel. An explanation of the inner workings of the respective algorithms would not empower patients to make an informed choice about a given treatment. Also conventional medical systems are highly complex and it would be absurd to provide patients with their technical details. After all, AI technologies simply represent more advanced techniques for analysing data.

[181] European Group on Ethics (n 18) 9.

[182] Beauchamp and Childress (n 118) 59.

[183] G Laurie, S Harmon and G Porter, *Mason and McCall Smith’s Law and Medical Ethics*, 10th edn (OUP 2016) 4.112.

[184] ibid 4.109.

[185] *Montgomery v Lanarkshire Health Board* [2015] SC 11 [2015] 1 AC 1430.

[186] ibid 49.

[187] Laurie et al (n 183) 4.139.

[188] *Montgomery v Lanarkshire Health Board* (n 185) 90.

[189] Royal College of Radiologists, *Standards of Patient Consent Particular to Radiology*, 2nd edn (2012), 5, <https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/BFCR%2812%298_consent.pdf> accessed 9 April 2019.

[190] House of Lords (n 13), The Royal College of Radiologists – Written evidence (AIC0146).

Information and access rights

So-called 'informational self-determination' and the human right to privacy as enshrined in Article 8(1) of the European Convention on Human Rights and Article 7 of the EU Charter of Fundamental Rights are further aspects of 'self-determination'. The main instrument to protect these rights is the General Data Protection Regulation (GDPR),[191] that—given its different scope—applies to the processing of personal data[192] in parallel to medical law and ethics. Personal data 'means any information relating to an identified or identifiable natural person'.[193] Patient data not only shared with healthcare professionals (HCP) but also user data processed by AI providers, eg of birth control apps, will normally even qualify as 'special category' personal data.[194] Hence, the processing of health data will generally require 'explicit consent' from the data subject.[195] The GDPR only defines 'regular' consent as any 'freely given, specific, informed and unambiguous' expression of agreement with the respective data processing.[196] According to the corresponding Guidelines of the Article 29 Data Protection Working Party (WP29), 'explicit' refers to 'the way' consent is given and requires an 'express statement'.[197] The rationale for the requirement of explicit consent is 'user control', which in the absence of meaningful information becomes 'illusory'.[198] Such information includes *inter alia* the controller's identity and the purpose of the related processing activity.[199] Of particular interest in an AI context, however, are Articles 13(1)(f), 14(1)(g) and 15(1)(h) in conjunction with Article 22 GDPR that provide for information and access rights in case of 'automated decision-making'. Accordingly, the data subject is to be provided with 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing'. This raises the question whether information disclosure imposed by the GDPR goes beyond what is to be disclosed under the concept of informed consent in terms of medical law and ethics. Indeed, in the literature it has been suggested that under this regime the 'algorithmic black box must be opened'.[200]

However, the majority opinion represented by Wachter and colleagues holds that the cited provisions mandate an explanation of the general 'system functionality' only.[201] This view is shared by the WP29 affirming that 'meaningful information about the logic involved' does not necessarily encompass 'a complex explanation of

[191] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Text with EEA relevance).

[192] art 1(1) GDPR.

[193] art 4(1) GDPR.

[194] arts 4(15) and 9(1) GDPR as well as Recital 35.

[195] art 9(2)(a) GDPR.

[196] art 4(11) GDPR.

[197] art 29 Data Protection Working Party, Guidelines on Consent under Regulation 2016/679, 17/EN WP259, 28 November 2017 (WP29, 2017a), at 18.

[198] ibid 13.

[199] ibid.

[200] B Goodman and S Flaxman, 'EU Regulations on Algorithmic Decision-Making and a "right to Explanation"', arXiv:1606.08813 [stat.ML], 28 June 2016.

[201] S Wachter, B Mittelstadt, and L Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) Int'l Data Priv L 76–99, 83.

the algorithms used or disclosure of the full algorithm'.[202] In particular, it has been acknowledged that, given the involved complexity, it can be challenging to understand, how an algorithm arrived at a specific decision in the first place.[203] In contrast to providing complex mathematical explanations about the decision-making algorithms, the WP29 suggests to provide for 'clear and comprehensive ways' of informing the data subjects.[204] As examples they list information about the categories of data that have been used and their relevance for the decision-making, as well as information related to the building of profiles or statistical groups and how they are relevant and the way a profile is used for a decision concerning the data subject.[205] A pragmatic and tangible approach is suggested by the UK Information Commissioner's Office (ICO) that includes visualization of the processing activities such as Venn diagrams and 'adjustable sliders' that demonstrate the weighting of different factors and their effects when altered.[206]

Further discussion

Even though some argue that opacity in AI disrespects the agency and autonomy of individuals,[207] because opacity would preclude a meaningful risk assessment,[208] the above findings about the way and scope of information to be provided *ex ante*[209] seem uncontentious. In particular, it is helpful that the disclosure requirements under medical as well as data protection laws are virtually congruent. Both regimes renounce detailed technical explanations about the inner workings of the applied algorithms, which would indeed contribute little to facilitate a meaningful choice about a therapy or treatment. The complementary data protection rights, however, may be seen as an extension of patient rights with specific regard to AI applications. It must be added, though, that the information and access rights as just described are subject to a set of restrictions and only apply if a decision is 'based solely' on automated processing, which produces 'legal' or 'similarly significant effects' concerning a data subject.[210] The meanings of these terms are not defined in the GDPR and bear considerable legal uncertainty,[211] not least in a context of healthcare. Article 22(1) GDPR seems to aim at transactions of economic life and Recital 71 mentions 'credit applications' and 'e-recruiting' as areas, where decisions can 'similarly significantly'

[202] art 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 17/EN WP251rev.01, 3 October 2017 (WP29, 2017), at 25.

[203] ibid.

[204] ibid 31.

[205] ibid.

[206] Information Commissioner's Office – Big data, artificial intelligence, machine learning and data protection version 2.0, March 2017 (ICO 2017), 87, <https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf> accessed 9 April 2019.

[207] A Rubel and KML Jones, 'Student Privacy in Learning Analytics: An Information Ethics Perspective' (2016) 32(2) Inform Soc 143–59.

[208] B Schermer, 'The Limits of Privacy in Automated Profiling and Data Mining' (2011) 27(1) Comp L & Sec Rev 45–52.

[209] For the discussion on *ex post* explanations, see Section 'The alleged "right to explanation"'.

[210] art 22(1) GDPR.

[211] See Wachter et al (n 201) 92.

affect data subjects. The WP29 lists 'decisions that affect someone's access to health services' as a further potential use case,[212] but stays silent on any decisions within a given care setting. However, since the threshold for 'significance' is the same as for 'producing a legal effect',[213] any decision within an existing doctor-patient relationship potentially impacting a person's health or wellbeing will normally qualify, because the said legal interests rank higher than economic interests. Further, the cited provisions only apply to decisions 'based solely' on automation. This term sparked off debates as to whether minimal human involvement merely 'rubber-stamping' an automated decision, could preclude data subject information and access.[214] According to the more convincing opinion, though, which is shared by the WP29, human involvement must be 'meaningful' and have 'actual influence' on the result.[215] Hence, to the extent that AI is only used as an augmenting tool, eg in radiology, and the physician still has 'the last call', Article 22(1) and subsequently Articles 13–15 GDPR cannot be invoked. On the other hand, where AI is used eg in dispatching processes with no or only nominal human oversight, these provisions apply.

Hence, in many day-to-day healthcare situations no legal obligations to inform the patients of the use of AI will exist, which seems justified in the light of the above. In contrast, where AI applications are deployed at the front-end, eg taking calls of patients for triaging or first advice, patient autonomy requires a clear notification that they are dealing with a machine.[216] Failing to do so is not only deceptive but will also diminish trust in AI applications more generally, as Google discovered lately when being accused of deceiving users for unleashing their virtual 'Duplex' assistant on restaurant personnel without clear disclosure.[217] Notwithstanding the foregoing, the concerns around objectification raised by the European Group on Ethics are not a matter of human-to-machine interaction *per se*. Autonomy, dignity and self-determination can all be thoroughly respected by a machine application as well as deeply disrespected within human-to-human interaction.[218]

Moral responsibility, accountability and liability

Moral responsibility

The final chapter within the core part of this article will deal with the question, who bears responsibility if an algorithmic decision leads to harm. In broad philosophical terms, the European Group on Ethics describes 'moral responsibility' as an integral

[212] WP29 (n 202) 22.

[213] ibid 21.

[214] See M Hildebrandt, 'The Dawn of a Critical Transparency Right for the Profiling Era', in J Bus et al (eds), Digital Enlightenment Yearbook 2012 (IOS Press 2012) 51.

[215] WP29 (n 202) 22.

[216] European Group on Ethics (n 18) 16.

[217] N Lomas, 'Duplex shows Google Failing at Ethical and Creative AI Design', *TechCrunch*, 10 May 2018, <https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/> accessed 9 April 2019.

[218] As vividly demonstrated in the notorious hospital scene of 'Little Britain' that coined the phrase 'computer said no', as it is often cited these days; in BBC, Little Britain, Series 4: Little Britain USA, 2008, see eg <https://www.youtube.com/watch?v=0n_Ty_72Qds> accessed 9 April 2019.

aspect 'of the conception of a person on which all our moral, social and legal institutions are based'.[219] The concept is further understood as the capacity of 'being in charge whether something succeeds or fails'[220] and being subject to corresponding 'reactions like praise or blame'.[221] The term combines several aspects of human agency,[222] of which 'accountability' and 'liability' will be my main focus in the discussion that follows.

Accountability. The European Group on Ethics briefly refers to 'accountability' as 'obligation to provide an account'.[223] Roughly, this encompasses at least two different categories, the one of justification in a sense of explainability of one's own actions and appeal,[224,225] as well as the one of redress including compensation and allocation of blame.[226] I will deal with the second set of aspects largely under the heading of liability. It seems clear, though, that meaningful redress depends on information about the wrong that happened in the first place. Against the opacity of AI systems accountability of decision-makers has been described as a primary challenge in the reviewed literature.[227] To that end, many scholars have emphasized the importance of transparency, and explainability of AI systems and an alleged 'right to explanation' under the GDPR in particular.[228]

The alleged 'right to explanation'. Transparency and explainability have been identified as major instruments to hold AI systems accountable, or their developers and owners, respectively.[229,230] Accordingly, a so-called 'right to explanation' in the context of automated decision-making was identified in the GDPR.[231] However, various academics have convincingly argued that such a right does not exist but rather a set of *ex ante* information and access rights about basic system functionality as demonstrated above. To the extent that a right to an *ex post* explanation of *specific decisions* is construed from the GDPR, Wachter and colleagues persuasively hold that this mistakenly links Articles 13(2)f and 14(2)g to safeguards as laid out in Article 22(3) GDPR and further conflates these binding provisions with the non-binding language

[219] European Group on Ethics (n 18) 10.

[220] V Dignum, 'Responsible Autonomy', arXiv:1706.02513 [cs.AI], 8 June 2017.

[221] A Eshleman, 'Moral Responsibility', in E Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition) <https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/> accessed 9 April 2019.

[222] ibid.

[223] ibid.

[224] Dignum (n 220) 6.

[225] V Dignum, 'The ART of AI — Accountability, Responsibility, Transparency, Medium', <https://medium.com/@virginiadignum/the-art-of-ai-accountability-responsibility-transparency-48666ec92ea5> accessed 9 April 2019.

[226] EJ Emanuel and LL Emanuel, 'What is Accountability in Health Care?' (1996) 124(2) Ann Int Med 229–39.

[227] See eg, Edwards and Veale (n 67).

[228] The debate was sparked off by Goodman and Flaxman (n 200).

[229] See eg, EC Communication on AI (n 11) 3.3.

[230] See eg, S Wachter, B Mittelstadt, and C Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2018) Harvard J L & Technol (v3), arXiv:1711.00399 [cs.AI].

[231] Goodman and Flaxman (n 202).

of Recital 71.[232] Indeed, the only mentioning of a right 'to obtain an explanation' is not included in the binding text of the GDPR and recitals only have interpretative character but are not binding themselves.[233] Furthermore, the semantics of Article 15(1)(h) GDPR that has been identified as another host to a right to explanation are of a forward-looking nature ('*envisaged* consequences') and do not allow the inference of anything beyond the explanation of system functionality.[234] This view is shared by Edwards and Veale[235] and the WP29 reaffirms that the GDPR - and Article 15(1)(h) in particular—only encompass the right to obtain *ex ante* 'information about the envisaged consequences of the processing, rather than an [*ex post*] explanation of a particular decision'.[236]

Human in the loop. It is widely accepted that moral responsibility as an intrinsically human property cannot be allocated or shifted to algorithms or machines, however sophisticated they may be.[237] AI systems exhibit 'autonomy' to some degree, in a sense that they are technically able to make predictions independently. Hence, in order to uphold moral responsibility and accountability of humans the European Group on Ethics requires 'meaningful human control' being maintained and that humans ultimately remain in control of the decision-making process.[238] This postulate is commonly referred to as 'human in the loop'. Some regard to this claim is paid by Article 22(3) GDPR that requires the implementation of 'suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision'. This obligation again is subject to the limitations as outlined above, and hence will not materialize in many of today's or near future AI applications.

Discussion. (a) The claims for maintaining 'meaningful human control' stem from the debate around lethal autonomous weapon systems[239] and an absolute framing in this regard seems appropriate indeed.[240] However, the discussion around autonomous driving has shown that 'full autonomy' in traffic might actually be safer than to require human intervention in critical situations, when the driver will be overly trusting in the system and almost certainly be distracted and miss the moment.[241] In healthcare, very much will depend on the context and ultimately on the safety and efficacy

[232] Wachter et al (n 201) 82.

[233] R Baratta, 'Complexity of EU Law in the Domestic Implementing Process' (2014) 2(3) The Theory and Practice of Legislation 293–308.

[234] Wachter et al (n 201) 84.

[235] Edwards and Veale (n 67) 48–53.

[236] WP29 (n 202) 27.

[237] European Group on Ethics (n 18) 10.

[238] ibid.

[239] See eg, B Docherty, 'Statement by Human Rights Watch on Meaningful Human Control in Lethal Autonomous Weapons Systems to the Convention on Conventional Weapons Group of Governmental Experts', Geneva, Human Rights Watch, 11 April 2018, <https://www.hrw.org/news/2018/04/11/statement-human-rights-watch-meaningful-human-control-lethal-autonomous-weapons> accessed 9 April 2019.

[240] See Future of Life Institute, Autonomous weapons: an open letter from AI & robotics researchers, <https://futureoflife.org/open-letter-autonomous-weapons/> accessed 9 April 2019.

[241] See eg, A Davies, 'The Very Human Problem Blocking the Path to Self-driving Cars', *Wired*, 1 January 2017, <https://www.wired.com/2017/01/human-problem-blocking-path-self-driving-cars/> accessed 9 April 2019.

of the respective applications. While today in the UK it might make sense to use AI as an augmenting technology, eg in radiology, and leave the final decision with the human specialist, things could already look differently in less developed countries with a shortage of specialist medical professionals. For example, where AI could help to detect diabetic eye disease at an early stage and help prevent blindness at a large scale,[242,243] it would be unethical to withhold such technology even if used without final specialist sign off by an ophthalmologist.

In view of this development, it would be desirable for the technologies to demonstrate ethical behaviour and be aligned with and respect our own human values. In 'computer ethics', this approach has been known as 'value sensitive design'[244] that has been further developed in 'machine ethics'.[245,246] Most importantly in a medical context, Anderson and Anderson designed an 'explicit ethical healthcare agent'.[247] That is an AI capable of representing the four principles of biomedical ethics and computing what is right or wrong in a given situation.[248,249] However, moral responsibility takes considerably more than the mere exhibition of ethical behaviour, namely the prevalence of consciousness and intentionality.[250,251] And although there are authors who believe in such possibilities,[252,253] these theories are mere speculation. That said, where human oversight is renounced this must be the result of a deliberate process, following formal guidelines and warranting approval by an ethical review board. The deployment of AI must remain accountable to the person ultimately in charge of the respective healthcare institution.

(b) While some authors perceive transparency and explainability as the fix for accountability in AI deployment,[254,255] others are more sceptical. Mittelstadt and

[242] See L Peng and V Gulshan, 'Deep Learning for Detection of Diabetic Eye Disease', 29 November 2016, <https://ai.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html> accessed 9 April 2019.

[243] V Gulshan et al, 'Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs' (2016) 316(22) JAMA 2402–10.

[244] See B Friedman and P Kahn, *Human Values, Ethics, and Design. The Human-Computer Interaction Handbook* (L. Erlbaum Associates Inc 2003) 1177–201.

[245] See eg W Wallach and C Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford Scholarship Online 2009).

[246] See also M Riedl and B Harrison, 'Using Stories to Teach Human Values to Artificial Agents', The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02, 2016.

[247] Originally coined by JH Moor, 'The Nature, Importance, and Difficulty of Machine Ethics' (2006) 21(4) IEEE Intelligent Systems 18–21.

[248] M Anderson and SL Anderson, 'Ethical Healthcare Agents', in M Sordo, S Vaidya and LC Jain (eds), *Advanced Computational Intelligence Paradigms in Healthcare - 3. Studies in Computational Intelligence*, vol 107 (Springer 2008) 233–57.

[249] M Anderson and SL Anderson, 'Machine Ethics: Creating an Ethical Intelligent Agent' (2007) 28(4) AAAI, AI Magazine 15–26.

[250] ibid 19.

[251] Sharkey A, 'Can Robots be Responsible Moral Agents? And why should we Care?' (2017) 29 J Conn Sci210–16.

[252] R Kurzweil, *The Singularity is near: when Humans Transcend Biology* (Penguin Books 2006).

[253] N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (OUP 2014).

[254] See eg, Tutt (n 65).

[255] See eg, S Wachter, B Mittelstadt, and L Floridi, 'Transparent, Explainable, and Accountable AI for Robotics' (2017) 2(6) Science Robotics eaan6080. <https://robotics.sciencemag.org/content/2/6/eaan6080> accessed 9 April 2019.

colleagues refer to transparency as a 'panacea' often naïvely being invoked to solve ethical concerns arising from new technologies.[256] Transparency is composed of 'accessibility' and 'comprehensibility' and Mittelstadt and colleagues identify issues with both of them.[257] Concerns around accessibility mainly relate to the fact that proprietary technologies powering AI systems will often be protected as trade secrets or by intellectual property rights as legitimate means to safeguard and incentivize investments into research and development of innovative technologies. To that end, Recital 63 of the GDPR makes a clear *caveat* that information and access rights should not adversely affect such rights 'and in particular the copyright protecting the software'. Further, an overly broad requirement of transparency could itself compromise privacy by revealing personal data hidden in underlying data sets.[258] And finally, disclosure of system functionality also encompasses security risks, and potentially exposes the systems and their users to 'gaming by bad actors'.[259,260]

More fundamentally, though, as shown above AI systems are optimized for performance and 'lag in common sense reasoning',[261] as would be the condition precedent for any kind of human comprehensible explainability. In many instances, making AI more explainable will mean to limit its complexity, which will almost certainly adversely impact performance.[262] Hence, as already seen with fairness, requests for explainability do not come without trade-offs.[263] In the face of all the above, Edwards and Veale warn to not create a so-called 'transparency fallacy'.[264] Rather it seems advisable to spend some more time on the concept of 'explanation' itself. As Doshi-Velez and colleagues note, explanation of the logic behind a certain decision may help to prevent similar errors in the future as well as provide a basis for dispute resolution.[265] To that end, also *human* decision-making may require explanation. However, Tutt and also Sandvig and colleagues hold algorithms to higher standards and request explainability virtually without exception. While we knew how to deal with human decision making, in their arguing we would not have a well-defined conception of how this translates to AI.[266,267] I do not quite agree with their conclusion. As Doshi-Velez and colleagues correctly observe, society cannot and does not require

[256] Mittelstadt et al (n 35) 6.

[257] ibid.

[258] PB de Laat, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?, Philosophy & Technology* (Springer 2017) 1–17.

[259] ibid.

[260] J Naughton, 'Good Luck in Making Google Reveal its Algorithm', *The Guardian*, 13 November 2016, <https://www.theguardian.com/commentisfree/2016/nov/13/good-luck-in-making-google-reveal-its-algorithm> accessed 9 April 2019.

[261] J McCarthy, 'Programs with Common Sense', RLE and MIT Computation Center, 1960, <https://stacks.stanford.edu/file/druid:yt623dt2417/yt623dt2417.pdf> accessed 9 April 2019.

[262] See eg, R Seseri, 'The Problem with "explainable AI"', *TechCrunch*, 14 June 2018, <https://techcrunch.com/2018/06/14/the-problem-with-explainable-ai/> accessed 9 April 2019.

[263] ibid.

[264] Edwards and Veale (n 67) 67.

[265] F Doshi-Velez, 'Accountability of AI Under the Law: The Role of Explanation' (November 3, 2017), Berkman Center Research Publication Forthcoming; Harvard Public Law Working Paper No 18-07 <https://arxiv.org/abs/1711.01134> accessed 9 April 2019, 2.

[266] Tutt (n 65) 105.

[267] C Sandvig et al, 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 2014,

an explanation in every instance since explanations are not free.[268] Accordingly, they suggest to hold AI to similar standards as humans in terms of when and what kind of explanations can be required *according to the law*.[269] This is a reasonable approach that clearly defines the situations wherein an explanation is owed and materially confines an explanation to the factors that were important in a decision and whether such factors were determinative.[270] This position seems to be shared by Reed, who also agrees with Doshi-Velez and colleagues that from an engineering perspective such explanation is feasible.[271] As a matter of fact, substantial engineering efforts are invested into 'AI interpretability' and a lot of progress in understanding AI systems has been made lately.[272,273] Especially promising are visualization methods,[274] some of them being available open source to the whole research community.[275]

I will now assess the second aspect of accountability that is liability including negligence as well as product liability in particular.

Liability

Negligence. In order to successfully bring a negligence claim against an HCP in the UK, the claimant must establish the existence of a 'duty of care' of the HCP, the 'breach of this duty' and a 'causality' between the breach and the suffered harm.[276] Where an HCP is involved the existence of a duty of care can normally be taken for granted. As for the standard of care, ever since *Bolam v Friern Hospital Management Committee* this has been that of the 'ordinary skilled doctor'.[277] If there is more than just one option, an HCP does not act negligently provided that the applied treatment is 'in accordance with a practice accepted as proper by a responsible body of medical men skilled in that particular art'.[278] However, progress will not happen if medical reasoning will always stay on the conservative side. Accordingly, common law is open to risk taking, even if the new treatment is endorsed only by a subspecialty ('super-specialists')[279] of a responsible body[280] and is not to be considered unreasonable.[281] Within this framework, the application of novel and innovative techniques is

<https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf?_ga=2.185713710.666911272.1534001293-185958133.1533215146> accessed 9 April 2019.

[268] Doshi-Velez et al (n 265) 3.

[269] ibid.

[270] ibid 7.

[271] C Reed, 'How Should we Regulate Artificial Intelligence?' (2018) 376(2128) Philos Trans A Math Phys Eng Sci. pii: 20170360. <https://www.ncbi.nlm.nih.gov/pubmed/30082306> accessed 9 April 2019.

[272] See eg, A Binder et al, 'Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers', 4 April 2016, arXiv:1604.00825 [cs.CV].

[273] See eg, Google, 'Responsible AI Practices', *Fairness*, <https://ai.google/education/responsible-ai-practices?category=fairness> accessed 9 April 2019.

[274] C Olah, A Mordvintsev, and L Schubert, 'Feature Visualization, How Neural Networks Build Up their Understanding of Images', 7 November 2017 <https://distill.pub/2017/feature-visualization/> accessed 9 April 2019.

[275] Eg Facets <https://pair-code.github.io/facets/> accessed 9 April 2019.

[276] See eg Laurie et al (n 183) 809–10.

[277] *Bolam v Friern Hospital Management Committee* [1957] 1 WLR 583.

[278] ibid.

[279] Laurie et al (n 183) 834.

[280] *De Freitas v O'Brien and Connolly* [1995] 6 Med LR 108.

[281] *Bolitho v City and Hackney Health Authority* [1998] AC 232.

possible or even desired, however, always taking into account all relevant factors under the given circumstances.[282,283]

That said, harm caused to patients due to an erroneous prediction by an AI system *alone* does not yet amount to negligence on part of the HCP. Rather the question is whether it was reasonable for a responsible medical (subspecialty) body to apply the respective AI technology in the given situation and also in the same manner. For example, in the 'pneumonia case' described in Section 'Inaccuracy' the sole reliance on the algorithm's classification of asthma patients as low risk without further examination or specialist oversight would certainly represent a negligent act, particularly because with conventional treatment such patients are directly sent to the intensive care unit. The finding could be different, though, eg where an HCP used a properly marketed and CE marked AI-powered app to calculate the exact dose of medication for a patient and the tool produces the wrong result. Although a lot will still depend on how obvious the error was, the use of established AI tools might well exonerate the HCP.

The element of causation is very much subject to the specific case and will not be further discussed. However, in the light of the opacity inherent in AI systems, it might indeed be an insurmountable burden for a patient to prove not only causation but the breach of a duty of care in the first place. These concerns are reinforced by the fact that unveiling the algorithmic technicalities to patients would neither be empowering nor feasible in many instances. Nevertheless, it would be unacceptable under any title to place the consequences of missing evidence on the claimant and allow the decision-maker to hide behind their black box. The doctrine of *res ipsa loquitur* could help. Hence, if the 'facts speak for themselves',[284] the doctrine 'infers negligence on the part of the defendant'[285] by establishing a *prima facie* case that must then be rebutted by the decision-maker.[286] The UK courts have applied the doctrine somewhat reluctantly.[287] Nevertheless, according to Laurie and colleagues it has been demonstrated to be most useful concerning instances involving machinery or complicated processes, of which a patient has little understanding and no explanation is offered by the defendant.[288] Thus, the doctrine seems almost perfectly tailored to the application of AI and in a dispute it would place the onus on the decision maker to providing a meaningful explanation.

Product liability. To the extent harm is caused due to the defect of medicinal products or medical devices the strict liability (no fault) scheme under the Product Liability Directive[289] and its national implementations (eg the UK Consumer Protection Act

[282] Eg, *Cooper v Royal United Hospital Bath NHS Trust* [2004] All ER (D) 51.

[283] Eg *Simms v Simms* [2002] 2 WLR 1465.

[284] As goes its literate translation.

[285] Laurie et al (n 183) 877.

[286] *Cassidy v Ministry of Health* [1951] 2 KB 343.

[287] See eg, *Ratcliffe v Plymouth & Torbay Health Authority* [1998] PIQR P170 (CA).

[288] Laurie et al (n 183) 878–9.

[289] Council Directive of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (85/374/EEC) (Product Liability Directive).

1987) might apply in parallel to a negligence action. Under the product liability regime, a producer is 'liable for damage caused by a defect in his product'.[290] A 'product' is defined as 'movables' or 'goods',[291] and a 'defect' is any deviation from the standard of 'safety which a person is entitled to expect' all things considered.[292] It seems non-controversial to say that damage caused by products controlled by *embedded* software normally falls within the scope of product liability.[293] In contrast, there has been considerable debate recently at the European level about its applicability to apps and other *non*-embedded software.[294,295] Although in the light of the very clear definition as just described not even consumer protection organizations suggest that these 'objects' could be understood as 'products' within the meaning of the said Directive 'at least not without succumbing to an extensive interpretation'.[296] Notwithstanding this, there are indications that the work of an 'Expert Group on liability and new technologies' as instantiated by the EC in March 2018 indeed does include non-embedded software.[297] Also, by application of a 'functional interpretation' Wagner suggests an analogous treatment of software under Article 2 Product Liability Directive. In essence, Wagner argues that the inclusion of 'electricity' under the term 'product' according to the last sentence of the cited provision was a mere example of an intangible object that is to be treated like a corporeal asset, and that software would be another, 'and even better example'.[298] Still according to Wagner, because at the time the Product Liability Directive was adopted in July 1985 the distribution of software had only been known in connection with a physical storage medium the framers had had no reason to include media-free, merely downloadable software but would clearly do so, if the said Directive were to be enacted today.[299] For these reasons, the Product Liability Directive in its current reading would already extend to non-embedded software and even computer programmes loaded via the cloud.[300]

Of course, all this ignores that the equation of storage media (eg DVD) and software, which would include also the mere code under a product liability regime has never been a convincing one, let alone one that had ever been confirmed by the CJEU. Phenomenologically, software is much more related to services than products. This is

[290] art 1 Product Liability Directive.

[291] art 2 Product Liability Directive; s 1(2) of the Consumer Protection Act 1987.

[292] art 6(1) Product Liability Directive; s 3(1) of the Consumer Protection Act 1987.

[293] See eg, EC Staff Working Document on Liability (n 17) 17.

[294] ibid 11.

[295] European Commission, Evaluation of Council Directive 85/374/EEC on the approximation of laws, regulations and administrative provisions of the Member States concerning liability for defective products, final report, 2018 (EC Evaluation, 2018), <https://publications.europa.eu/en/publication-detail/-/publication/d4e3e1f5-526c-11e8-be1d-01aa75ed71a1/language-en> accessed 9 April 2019.

[296] See eg, BEUC, Review of product liability rules, Brussels, 25 April 2017, at 3, <http://www.beuc.eu/publications/beuc-x-2017-039_csc_review_of_product_liability_rules.pdf> accessed 9 April 2019.

[297] See European Commission, Launch of call for experts for group on liability and new technologies, 9 March 2018, <https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=615947> accessed 9 April 2019.

[298] G Wagner, Robot Liability, 19 June 2018, <https://ssrn.com/abstract=3198764> accessed 9 April 2019 or <http://dx.doi.org/10.2139/ssrn.3198764> accessed 9 April 2019, 11.

[299] G Wagner, ‚Produkthaftung für Autonome Systeme‘ (2017) 217(6) Archiv für die civilistische Praxis (AcP) 707–65, 59.

[300] ibid.

particularly true for cloud applications, as the capital 'S' in IaaS, PaaS and SaaS (infrastructure-, platform- and software as a service) reinforces without doubt. Services are clearly beyond the scope of the Product Liability Directive[301] and any suggestions that 'intangibles' and non-embedded software in particular would fall within its scope are misconceived. Products defined as 'movables' naturally entail tangibility. Also, the fact that 'electricity' was explicitly added to the definition of 'products' by Directive 1999/34/EC[302] shows that this was an exception to the rule of tangibility. At that time (1999), software and downloadable code had already been broadly known and applied phenomena. The fact that software had not been included along with electricity mandates that the legislator did not intend to do so. In addition, any liability for 'defective software'—that arguably could include the concept of 'defective information'—presents a very difficult proposition under aspects of Article 10 of the European Convention on Human Rights and could not only curtail the right to freedom of expression but also the underlying creative space itself. Thus, an extension of the Directive's scope by mere interpretation or any kind of 'soft law' would not be tenable but require carefully following due legislative process.

Discussion

(a) As the example above showed, it might well be possible that the proper use of duly marketed AI tools that subsequently fail will exonerate the HCP from allegations of negligence. To the extent that such technology does not qualify as a 'product', the patient will also lose any product liability claim. Under aspects of a fair risk distribution this situation is not sustainable. There are essentially two options to close this gap. The first one is to extend the duty of care and either include the producer of the AI system directly or make the HCP vicariously liable for the producer. Although the latter possibility should be dismissed for not being 'fair, just or reasonable'.[303] Indeed, it would be paradoxical to exonerate the HCP from using the technology but make them *indirectly* liable for the acts or omissions of the AI provider they have no control over. A direct general tort claim by the patient against the AI provider would have to be assessed carefully especially in relation to 'sufficient proximity' and 'reasonable foreseeability'.[304] This would go beyond the scope of this article in medical law and ethics. But it seems safe to say that the AI provider is generally subject to national extra-contractual tort or delict laws. More importantly, it would leave patients a lot worse off than with a strict liability claim under product liability.

Hence, as per the second option, products that qualify as medical devices normally fall within the scope of the Product Liability Directive.[305] AI-powered apps and

[301] CJEU, Case 495/10 (*Centre hosptialier universitaire de Besancon v Dutrueux*), 21 December 2010, para 39; CJEU, 10.5.2001, Case 203/99 (*Veedfald v Arhus Amtskommune*), 10 May 2001, para 17; Brüggemeier G, Tort Law of the European Union, Wolters Kluwer, 2015, para 298.

[302] Directive 1999/34/EC of the European Parliament and of the Council of 10 May 1999 amending Council Directive 85/374/EEC on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products.

[303] See the standard set in *Caparo Industries plc v Dickman* [1990] UKHL 2, [1990] 2 AC 605.

[304] See *Goodwill v British Pregnancy Advisory Service* [1996] 2 All ER 161.

[305] See eg, CJEU, Cases C-503/13 and C-504/13, *Boston Scientific Medizintechnik GmbH v AOK Sachsen-Anhalt - Die Gesundheitskasse, Betriebskrankenkasse RWE*, 5 March 2015.

non-embedded software used for diagnosis or treatment may qualify as 'medical devices' within the meaning of Article 1(2)(a) of the Medical Devices Directive,[306] like eg the Babylon technology used within the NHS.[307] This is even more true under the new Medical Devices Regulations (EU) 2017/745[308] and 2017/746,[309] according to which 'software in its own right, when specifically intended by the manufacturer to be used for one or more of the medical purposes set out in the definition of a medical device, qualifies as a medical device'[310] or in vitro diagnostic medical device, respectively.[311] It is important to note, though, that there is no link between the Medical Devices Directive or the Medical Devices Regulations and the Product Liability Directive, which would automatically translate non-embedded software into 'products', even if they qualified as (in vitro diagnostic) medical devices. A different treatment of (*in vitro* diagnostic) medical devices merely based on their tangibility, however, is hardly defensible and should be reconsidered. Much rather the decisive criteria should be the risk classification the said regulation attributes regardless of physical aspects. Medical devices regulation is the primary basis for safety and efficacy controls over AI systems used within healthcare.[312] Considerable doubts can be raised whether the overall risk-landscape justifies opening up the Product Liability Directive for non-embedded software across industries *in general*. Thus, strict liability should rather be introduced on a sector-specific basis within the medical devices regulation also taking into account the so-called 'development risks'[313] in the context of (*in vitro* diagnostic) medical devices in the form of non-embedded software. Alternatively, (*in vitro* diagnostic) medical devices could be included in Article 2 Product Liability Directive by reference. The details of such regulation, however, go beyond the scope of this article.[314,315,316]

(b) The EC is currently evaluating the existing product liability framework for its fitness to deal with so-called 'emerging digital technologies' that *inter alia* include AI.[317] While this work is largely determined by the hardware components of the

[306] Council Directive 93/42/EEC of 14 June 1993 concerning medical devices [1993] OJ L169/1.

[307] GP at hand, <https://www.gpathand.nhs.uk/> accessed 9 April 2019.

[308] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Medical Devices Regulation).

[309] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (In Vitro Diagnostic Medical Devices Regulation).

[310] Recital 19 in conjunction with art 2 Medical Devices Regulation.

[311] Recital 17 in conjunction with art 2 In Vitro Diagnostic Medical Devices Regulation.

[312] See eg, A den Exter, *Research Handbook on EU Health Law and Policy*, in Tamara K Hervey et al (ed) (Edward Elgar Publishing, Incorporated 2017) 255.

[313] art 7(e) of the Product Liability Directive.

[314] See eg, H Harvey, 'How to get Clinical AI Tech Approved by Regulators', Towards Data Science, 7 November 2017, <https://towardsdatascience.com/how-to-get-clinical-ai-tech-approved-by-regulators-fa16dfa1983b> accessed 9 April 2019.

[315] European Commission, MEDDEV 2.1/, July 2016, 'Guidelines on the Qualification and Classification of Standalone Software used in Healthcare within the Regulatory Framework of Medical Devices'.

[316] Medicines and Healthcare products Regulatory Agency, Guidance: Medical device stand-alone software including apps (including IVDMDs), published 8 August 2014, last updated 20 June 2018.

[317] EC Staff Working Document on Liability (n 17).

internet of things (IoT) and 'advanced robotics', AI-specific concerns mainly centre around the element of 'unpredictability' due to the alleged 'self-learning' capabilities of AI systems.[318] It is feared that these systems would develop unethical or even illegal behaviours not foreseeable to their human developers and thus offer them an easy way out of liability.[319] An oft-cited example is Microsoft's chatbot 'Tay' that learned racist and sexist language and needed to be withdrawn on launch day.[320] Under a negligence regime there might be limited space for a developer to argue that such 'self-learning' would be beyond what they can reasonably control. However, these notions have largely been rejected by numerous leading experts as based on an 'overvaluation of the actual capabilities' of such systems as well as 'a superficial understanding of unpredictability and self-learning capacities'.[321] Indeed, a 'Tay-like disaster' can be avoided by rigorous testing and by 'freezing' the algorithm and disabling 'online-learning', or at least by confining such continuous learning from real-time data input to a trusted environment of carefully selected users. Where such measures are omitted, a strong indication for negligence will exist. Arguably, under the regime of existing product liability law such unpredictability would not prevent an AI developer from becoming liable, if their AI over time moved away from a safety standard that may reasonably be expected by the respective users. This further aspect also speaks for closing the gap between the Medical Devices and the Product Liability Directives as suggested above.

(c) With progress happening fast the standard of care might shift accordingly. Soon we may not discuss negligence in connection with the use of AI but rather with the *omission* to use it. If the technologies become ubiquitous via apps or cloud services and their capacities trump those of a human HCP, diligent treatment should include them.[322] Even though a global adoption will not have to take place overnight, it would be recommended for HCPs to stay informed about major AI developments in their field.[323] Arguably, the greatest potential for negligence cases will be during the transitioning phase,[324] when AI technologies gain traction but have not been fully established yet. Also, concerns have been raised about situations, where human decisions conflict with algorithmic decisions.[325] Existing and near future applications are meant to augment an HCP's abilities and assist them in their decision making, rather than replace them. In contrast, Price discusses applications of so called 'black-box-medicine' where due to their opacity the HCP cannot be seen merely as the last

[318] ibid 10.

[319] See eg, T Simonite, 'When Bots Teach Themselves to Cheat', *Wired*, 8 August 2018, <https://www.wired.com/story/when-bots-teach-themselves-to-cheat/> accessed 9 April 2019.

[320] J Vincent, 'Twitter taught Microsoft's AI Chatbot to be a Racist Asshole in Less than a Day', *The Verge*, 24 May 2016, <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist> accessed 9 April 2019.

[321] The signatories, Open letter to the European Commission, Artificial intelligence and robotics, 2018 (Open Letter, 2018) <http://www.robotics-openletter.eu/> accessed 9 April 2019.

[322] See the considerations of T Yang and R Silverman, 'Mobile Health Applications: The Patchwork Of Legal And Liability Issues Suggests Strategies To Improve Oversight' (2014) 33(2) Health Affairs 222–27.

[323] See the standard set in *Crawford v Charing Cross Hospital* [1953], *The Times*, 8 December, CA.

[324] Also see, Laakmann AB, 'When should Physicians be Liable for Innovation?' (2015) 36 Cardozo L Rev 913.

[325] Future Advocacy (n 3) 34.

step in a sequence of care, like the suggestion of an off-label use of an approved therapy or medicinal product by an AI system.[326] Accordingly, Price does not think the traditional standard is appropriate.[327] However, unless there is unambiguous guidance from regulators or overseeing bodies, the last call for the application of a certain technology remains with the HCP. While the latter is liable for its decision, AI producers bear the liability for the safe and efficient functioning of their technologies.[328] Medical devices regulation will have to provide for adequate safety, quality and post-market surveillance standards. To that end, HCP's should be encouraged to a reasonable use—in accordance with the respective design and label constraints, that is—of properly tested and marketed medical devices involving AI without being made liable for their malfunctioning. Ultimately, a risk-utility test should be applied that should be deemed as passed, if an AI application proves at least as safe and effective as conventional methods or mere human decision making.

(d) For the sake of completeness, the proposals of some authors shall be mentioned to attribute legal personhood—or even rights in their own capacity[329]—to advanced autonomous systems.[330] Such legal entities, it is suggested, should be endowed with sufficient assets to cover liability in case they do harm.[331] These ideas, however, have been broadly dismissed for being unnecessary, impractical, unethical and open to abuse.[332,333] Indeed, not everything needs to be changed and existing statutes and case law will remain appropriate to deal with many new challenges. For example, in a very topical case users of birth control apps have reported unwanted pregnancies because of wrong predictions about their fertility cycles.[334,335] However, in *Richardson v LRC Products Ltd*, a UK case from 2000, a product liability claim over a burst condom had been rejected, essentially because the risks inherent with the use of condoms were deemed as common knowledge.[336] Arguably, this reasoning may be transposed even *a fortiori* to any temperature based method of contraception, AI-powered or not.

CONCLUSION

From a considered view of the relevant technologies it can be concluded that the current laws and ethical concepts are largely suited to deal with most concerns that have been raised in the reviewed literature particularly around bias, opacity and failure to

[326] WN Price, 'Medical Malpractice and Black-box Medicine', in I Cohen et al (eds), *Big Data, Health Law, and Bioethics* (CUP 2018).

[327] ibid.

[328] Future Advocacy (n 3), King D, cited at 35.

[329] DJ Gunkel, *Robot Rights* (The MIT Press 2018).

[330] European Parliament (n 20).

[331] ibid.

[332] J Bryson, M Diamantis and T Grant, 'Of, For, and By the People: The Legal Lacuna of Synthetic Persons', (2017) 25 Artif Intell Law 273–91.

[333] Open Letter (n 321).

[334] O Sudjic '"I Felt Colossally Naïve"': the Backlash against the Birth Control App', *The Guardian*, 21 July 2018, <https://www.theguardian.com/society/2018/jul/21/colossally-naive-backlash-birth-control-app> accessed 9 April 2019.

[335] See Läkemedelsverkets, Swedish Medical Products Agency, 30 January 2018, <https://lakemedelsverket.se/Alla-nyheter/NYHETER—2018/Lakemedelsverkets-tillsynsarende-av-Natural-Cycles-fortgar/> accessed 9 April 2019.

[336] *Richardson v LRC Products Ltd* [2000] 59 BMLR 185.

accurately model the real world. Some prompt clarifications, however, would be desirable. Responsible approaches to engineering already shift the onus of statistical uncertainty and opacity to the AI developers. The law should follow suit and clarify that the lack of explanation shall always be to the detriment of the decision-maker. Further, the existing gap between non-embedded software that qualifies as medical devices but not as products within the meaning of product liability should be closed. The law should also allow developers to build policy constraints around sensitive categories like gender, age or race, in order to avoid unfair outcomes without exposing them to the risk of legal actions. Ultimately, responsible and diversified engineering practices—examples of which were given in this article—should be incentivized and prioritized above hard and premature regulation that could well stifle many useful innovations benefiting patients and the health system in this still nascent field.