THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE DOCTOR-PATIENT RELATIONSHIP

Report commissioned by the
Steering Committee for Human Rights
in the fields of Biomedicine and Health (CDBIO)

Author: Brent Mittelstadt

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE DOCTOR-PATIENT RELATIONSHIP

By Brent Mittelstadt, Senior Research Fellow and Director of Research at the Oxford Internet Institute, University of Oxford, United Kingdom

All requests concerning the reproduction or translation of all or part of this document should be addressed to the Directorate of Communication (F-67075 Strasbourg Cedex).

All other correspondence concerning this document should be addressed to the Directorate General of Human Rights and Rule of Law.

© Council of Europe, December 2021

TABLE OF CONTENTS

1 ESSENTIAL ELEMENTS

1. In response to a call by the Committee on Bioethics (DH-BIO)[1] to work on trust, safety, and transparency, this report investigates the known and potential impacts of AI systems on the doctor-patient relationship. This impact is framed by the human rights principles referred to in the European Convention on Human Rights and Biomedicine of 1997, otherwise known as the "Oviedo Convention," and its subsequent amendments.

2. The deployment of AI in clinical care remains nascent. Clinical efficacy has been established for relatively few systems when compared to the significant research activity in healthcare applications of AI. Research, development, and pilot testing often do not translate into proven clinical efficacy, commercialization, or widespread deployment. The generalization of performance from trials to clinical practice generally remains unproven.

3. A defining characteristic of medicine is the 'healing relationship' between clinicians and patients. This relationship is augmented by the introduction of AI. However, the role of the patient, the factors that lead people to seek medical attention, and the patient's vulnerability are not changed by the introduction of AI as a mediator or provider of medical care. Rather, what changes is the means of care delivery, how it can be provided, and by whom. The shift of expertise and care responsibilities to AI systems can be disruptive in many ways.

4. The potential human rights impact of AI on the doctor-patient relationship can be categorised according to six themes: (1) Inequality in access to high quality healthcare; (2) Transparency to health professionals and patients; (3) Risk of social bias in AI systems; (4) Dilution of the patient's account of well-being; (5) Risk of automation bias, de-skilling, and displaced liability; and (6) Impact on the right to privacy.

5. Concerning (1), as an emerging technology the deployment of AI systems will not be immediate or universal across all member states or healthcare systems. Deployment across institutions and regions will inevitably be inconsistent in terms of scale, speed, and prioritisation.

6. The impact of AI on clinical care and the doctor-patient relationship remains uncertain and will certainly vary by application and use case. AI systems may prove to be more efficient than human care, but also provide lower quality care featuring fewer face-to-face interactions.

7. The inconsistent rollout of AI systems with uncertain impacts on access and care quality poses a risk of creating new health inequalities in member states.

[1] Committee replaced by the Steering Committee for Human rights in the fields of Biomedicine and Health (CDBIO).

8. Article 4 of the Oviedo Convention addresses care provided by healthcare professionals bound by professional standards. It remains unclear whether developers, manufacturers, and service providers for AI systems will be bound by the same professional standards.

9. Careful consideration must be given to the role played by healthcare professions bound by professional standards when incorporating AI systems that interact directly with patients.

10. Concerning (2), transparency and informed consent are key values in the AI-mediated doctor-patient relationship. The complexity of AI raises a question: how should AI systems explain themselves, or be explained, to doctors and patients? This question has many possible meanings: (i) How does an AI system or model function? How was a specific output produced by an AI system? (ii) How was an AI system designed and tested? How is it governed? (iii) What information is required to investigate the behaviour of AI systems? Answers to each of these questions may be necessary to achieve informed consent in AI-mediated care.

11. In cases where AI systems provide some form of clinical expertise, for example by recommending a particular diagnosis or interpreting scans, this requirement to explain one's decision-making would seemingly be transferred from doctor to AI system, or at least to manufacturer of AI system. The difficulty of explaining how AI systems turn inputs into outputs poses a fundamental challenge for informed consent. Aside from the patient's capacity to understand the functionality of AI systems, in many cases patients simply do not have sufficient levels awareness to make free and informed consent possible. AI systems use unprecedented volumes of data to make their decisions, and interpret these data using complex statistical techniques, both of which increase the difficulty and effort required to remain aware of the full scope of data processing and clinical analysis informing one's diagnosis and treatment.

12. AI systems interacting directly with patients should self-identify as an artificial system. Whether the usage of AI systems in care settings should always be disclosed to patients by clinicians and healthcare institutions is a more difficult question.

13. Concerning (3), AI systems are widely recognised as suffering from bias in their inputs, processing, and outputs. Biased and unfair decision-making often occurs not for technical or regulatory reasons, but rather reflects underlying social biases and inequalities. For example, samples in clinical trials and health studies have historically been biased towards white male subjects meaning results are less likely to apply to women and people of colour.

14. Social biases in AI systems can lead to unequal distribution of outcomes across patient populations and protected demographic groups. Western societies have long been marked by significant social inequality. These historical and contemporary trends influence the training of future systems. Without

intervention, these patterns in access to healthcare opportunities and resources will be learned and reinforced by AI systems.

15. Detecting biases in AI systems is not straightforward. Biased decision-making rules can be hidden in 'black box' models. Simply anonymising health data may not be an adequate solution to mitigate biases due to the influence of historical inequality and the existence of strong proxies for protected attributes (e.g., post code as a proxy for ethnicity). The various challenges of social bias, discrimination, and inequality suggest health professionals and institutions face a difficult task in ensuring their usage of AI systems does not further existing inequalities and create new forms of discrimination.

16. Concerning (4), the development of trust in a doctor-patient relationship may be inhibited by technological mediation. As a mediator placed between the doctor and patient, AI systems can inhibit tacit understanding of the patient's health and well-being and encourage both clinician and patient to discuss health solely in measurable quantities or machine interpretable terms.

17. Concerning (5), to ensure patient safety and replace the protection offered by human clinical expertise, robust testing and validation standards should be an essential pre-deployment requirement for AI systems in clinical care contexts. Evidence of clinical efficacy does not yet exist for many AI applications in healthcare, which has justifiably proven a barrier to widespread deployment.

18. Concerning (6), AI poses several unique challenges to the human right to privacy and complementary data protection regulations. These rights seek to provide individuals with greater transparency and control over automated forms of data processing. They will undoubtedly provide valuable protection for patients across a variety of use cases of medical AI.

19. The Oviedo Convention sets out a specific application of the right to privacy (Article 8 ECHR) which recognises the particularly sensitive nature of personal health information and sets out a duty of confidentiality for health care professionals.

20. Ethical standards need to be developed around transparency, bias, confidentiality, and clinical efficacy to protect patient interests in informed consent, equality, privacy, and safety. Such standards could serve as the basis for deployments of AI in healthcare that help rather than hinder the trusting relationship between doctors and patients.

21. Where AI can be observed to have a clear impact on rights and protections set out in the Oviedo Convention, it is appropriate for the Council of Europe to introduce binding recommendations and requirements for signatories concerning how AI is deployed and governed. Recommendations should focus on a higher positive standard of care with regards to the doctor-patient relationship to ensure it is not unduly disrupted by the introduction of AI in care settings.

22. The Council of Europe could set standards for what and how information about the recommendation of an AI system concerning a patient's diagnosis and treatment should be communicated to the patient. These standards should likewise address the doctor's role in explaining AI recommendations to patients and how AI systems can be designed to support the doctor in this role.

23. The capacity of AI to replace or augment human clinical expertise utilising highly complex analytics and unprecedented volumes and varieties of data suggests its impact on the doctor-patient relationship may be unprecedented.

24. The degree to which AI systems inhibit 'good' medical practice hinges upon the model of service. If AI is used solely to complement the expertise of health professionals bound by the fiduciary obligations of the doctor-patient relationship, the impact of AI on the trustworthiness and human quality of clinical encounters may prove to be minimal. At the same time, if AI is used to heavily augment or replace human clinical expertise, its impact on the caring relationship is more difficult to predict. It is entirely possible that new, broadly accepted norms for 'good' care will emerge through greater reliance on AI systems, with clinicians spending more time face-to-face with patients and relying heavily on automated recommendations. The impact of AI on the doctor-patient relationship nonetheless remains highly uncertain. We are unlikely to see a radical reconfiguration of care in the next five years in the sense of human expertise being replaced outright by artificial intelligence.

25. A radical reconfiguration of the doctor-patient relationship of the type imagined by some commentators, in which artificial systems diagnose and treat patients directly with minimal interference from human clinicians, continues to seem far in the distance.

26. Going forward, the ideal model of clinical care and AI deployment in healthcare is one that utilises the best aspects of human clinical expertise and AI diagnostics.

27. The doctor-patient relationship is a keystone of 'good' medical practice, and yet it is seemingly being transformed into a doctor-patient-AI relationship. The challenge facing AI providers, regulators, and policymakers is to set robust standards and requirements for this new type of 'healing relationship' to ensure patients' interests and the moral integrity of medicine as a profession are not fundamentally damaged by the introduction of AI.

2 INTRODUCTION

Technological solutions such as artificial intelligence (AI) are increasingly seen as a potential solution to growing resource pressures in medicine, healthcare, and biomedical research. AI systems promise novel means to evaluate and improve the quality of clinical care, undertake biomedical research and investigate new therapeutics and pharmaceuticals, and expand care offerings to previously underserved populations.[2] A key driver of innovation and adoption is the belief that AI may relieve health professionals from "certain time-consuming clerical tasks and could increase their time for caregiving practices."[3] Medical decision-making and care are increasingly supported by expert and robotics systems that assist in record management, diagnosis, treatment planning, and delivery of interventions. Home and social care are similarly transformed through the introduction of remote monitoring and management systems. Health can increasingly be monitored, modelled, and managed based on data representations of the patient, supplementing or replacing verbal accounts and face-to-face physical care.[4]

A unique impact of AI and other emerging data-intensive and algorithmic technologies is their capacity to augment and support human decision-making by recommending the best action to take in a given situation, the best interpretation of data, and so on.[5] But these systems can also be used to outright replace human decision-making, expertise, and face-to-face clinical care. Natural language processing applications such as OpenAI's GPT-3, for example, suggest a future in which initial patient contact and even triage can be handled in part or entirely by artificial conversational agents. AI systems are already used by clinicians and hospitals for clinical and operational decision-making, seen for instance in risk prediction, discharge planning, diagnostics, and decision-support systems.[6] Developments in deep learning likewise suggest a future in which drug discovery and biomedical research are increasingly driven by computational systems capable of intelligent behaviour.[7] Recent advances in the pharmaceuticals to treat a rare form of brain cancer or Deepmind's breakthrough in

[2] World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance* (2021); ITALIAN COMMITTEE FOR BIOETHICS, *Artificial Intelligence and Medicine: some ethical aspects* (2020), http://bioetica.governo.it/en/opinions/joint-opinions-icbicbbsl/artificial-intelligence-and-medicine-some-ethical-aspects/ (last visited Nov 30, 2021).
[3] COUNCIL OF EUROPE, *Artificial intelligence in health care: medical, legal and ethical challenges ahead* (2020).
[4] Brent Mittelstadt et al., *The Ethical Implications of Personal Health Monitoring*, 5 INTERNATIONAL JOURNAL OF TECHNOETHICS 37–60 (2014).
[5] George A. Diamond, Brad H. Pollock & Jeffrey W. Work, *Clinician decisions and computers*, 9 JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY 1385–1396 (1987); James G. Mazoué, *Diagnosis Without Doctors*, 15 J MED PHILOS 559–579 (1990).
[6] Rebecca Robbins & Erin Brodwin, *Patients aren't being told about the AI systems advising their care*, STAT (2020), https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/ (last visited Nov 9, 2021).
[7] World Health Organization, *supra* note 1.

protein folding via AlphaFold already show the potential of the state of the art in medical AI.[8]

While the promise of AI is clear, a significant area of uncertainty concerns its impact on the practice of healthcare, and in particular the doctor-patient relationship. Medical expertise is no longer the sole domain of trained medical professionals and researchers; rather, AI technologies create opportunities to provide care through a mix of public and private, professional and non-professional, and human and technological stakeholders.

In response to the growing recognition of these opportunities and risks of AI on the practice of medicine and clinical care by the Council of Europe, and the call by the Committee on Bioethics (DH-BIO) to work on trust, safety, and transparency in this context,[9] this report investigates the known and potential impacts of AI systems on the doctor-patient relationship. This impact is framed by the human rights principles referred to in the European Convention on Human Rights and Biomedicine of 1997 otherwise known as the "Oviedo Convention," and its subsequent amendments. Human rights principles regarding health may require certain standards to be met in the doctor-patient relationship which can be disrupted, displaced, or at least augmented by the usage of AI in clinical care.

The report is structured as follows:

- **Section 2** provides background and context concerning definitions of AI and related technologies, common ethical challenges posed by AI systems, and a brief historical overview of human rights principles regarding health in the context of the Oviedo Convention.
- **Section 3** reviews types of AI technologies in medicine, focusing in particular on AI systems aimed at augmenting clinical care and the patient experience.
- **Section 4** proposes a theoretical framework for the doctor-patient relationship based in human rights and connecting the aims of medicine to the standards of good medical practice as developed by medicine as a formal profession.
- **Section 5** then proposes several categories of current and potential impacts of AI systems on the doctor-patient relationship, focusing on issues of bias, inequality in access to care, opacity and transparency, patient autonomy and safety, clinician responsibility and automation bias, and the human right to privacy.
- **Section 6** concludes with recommendations aimed at bolstering human rights protections in the context of AI and the doctor-patient relationship.

[8] Diana M. Carvalho et al., *Repurposing vandetanib plus everolimus for the treatment of ACVR1-mutant diffuse intrinsic pontine glioma*, CANCER DISCOV (2021), https://cancerdiscovery.aacrjournals.org/content/early/2021/09/20/2159-8290.CD-20-1201 (last visited Nov 30, 2021); John Jumper et al., *Highly accurate protein structure prediction with AlphaFold*, 596 NATURE 583–589 (2021).
[9] COUNCIL OF EUROPE, *supra* note 2.

3 BACKGROUND AND CONTEXT

Concepts such as artificial intelligence (AI), machine learning, algorithm, and AI system have a wide array of meanings across academic, policy, and public discourse. Unhelpfully, the concepts are often used interchangeably.[10] For the sake of clarity, some definitions and distinctions will be offered.

Artificial intelligence refers to the demonstration of intelligence by a machine, wherein intelligence is understood in terms of its expression in humans and animals. As an academic field artificial intelligence studies "intelligent agents" or "computational intelligence", understood as systems that perceive their environment and take actions that maximize their chances of achieving their goals.[11] Machine learning can be understood as a specialised type of AI in which the agent, or computer program, improves its performance at some task through experience. Machine learning systems use "prior knowledge together with training data to guide learning."[12]

In simple terms, machine learning can be thought of as a type of software that learns from a training dataset, wherein labels are created and applied by human labellers according to prior knowledge. A classic example is an image recognition program which is taught to distinguish between classes of objects. In this case the training dataset would consist of a series of pre-labelled images from which the system can derive classification rules to apply to new images or datasets.

Algorithms can be understood as core components of machine learning and artificial intelligence systems that guide the processes of learning and turning input data into outputs. In mathematical terms an algorithm can be understood as a mathematical construct with "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions."[13] For clarity, a simpler definition can be offered: an algorithm is a well-defined sequence of steps that produce an output from some set of inputs.

A machine learning algorithm can be understood as a type of algorithm in which a part of the sequence of steps has been learnt rather than pre-defined. For example, a machine learning algorithm used for classification tasks develops classes that can generalise beyond the training data.[14] The algorithm creates a model to classify new inputs. A machine learning model is the internal data of the algorithm that is fitted to input data to improve performance.

Image recognition technologies, for example, can decide what types of objects appear in a picture. The algorithm 'learns' by defining rules to determine how new inputs will be classified. The model can be taught to the algorithm via hand labelled inputs (supervised learning); in other cases, the algorithm itself defines best-fit models to

[10] Robin K. Hill, *What an Algorithm Is*, 29 PHILOS. TECHNOL. 35–59, 36 (2015).
[11] David Poole, Alan Mackworth & Randy Goebel, *Computational Intelligence* (1998).
[12] Tom Mitchell, *Machine learning* (1997).
[13] Hill, *supra* note 9 at 47.
[14] Pedro Domingos, *A few useful things to know about machine learning*, 55 COMMUNICATIONS OF THE ACM 78–87 (2012).

make sense of a set of inputs (unsupervised learning).[15] In both cases, the algorithm defines decision-making rules to handle new inputs. Critically, a human user will typically not be able to understand the rationale of decision-making rules produced by the algorithm.[16]

Popular and policy definitions of these terms often do not follow these technical definitions which can cause confusion. The World Health Organization (WHO), for example defines artificial intelligence as "the performance by computer programs of tasks that are commonly associated with intelligent beings." Definitions of this type are on the one hand problematically broad, insofar as they turn on the definition of "intelligence" and scope of behaviours of "intelligent beings," and thus cannot be used to classify a particular system or AI or not-AI alone. With that said, the openness of the definition can also be helpful in policy terms by enabling additional systems to be captured beyond the state-of-the-art at the point of drafting.

Regardless of their limitations, policy definitions of AI are arguably more important than technical definitions if our concern is with harmonisation across regulatory and policy frameworks. The 'Artificial Intelligence Act' (AIA), a proposed horizontal risk-based regulatory framework proposed by the European Commission, offers a particularly broad definition of AI that promises to be an influential international policy going forward[17]:

> "'Artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Appendix I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."

Appendix of the AIA offers a non-comprehensive list of techniques and approaches that can be considered AI, which encompasses machine learning, logic and knowledge-based approaches, and a variety of statistical methods:

> "(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

[15] Bart W. Schermer, *The limits of privacy in automated profiling and data mining*, 27 COMPUTER LAW & SECURITY REVIEW 45–52 (2011); Martijn Van Otterlo, *A Machine learning view on profiling*, PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN–PHILOSOPHERS OF LAW MEET PHILOSOPHERS OF TECHNOLOGY 41–64 (2013).

[16] Andreas Matthias, *The responsibility gap: Ascribing responsibility for the actions of learning automata*, 6 ETHICS INF TECHNOL 175–183, 179 (2004).

[17] EUROPEAN COMMISSION, *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*, 2021/0106(COD) (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (last visited Oct 27, 2021).

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods."

As this definition shows the AIA's definition of 'AI system' does not align strictly with the technical definitions offered above. For example, in this definition machine learning is treated as a component of AI rather than as a specialised type of AI. To avoid ambiguity, we offer the following working definition of 'artificial intelligence system' for the purposes of this report:

> 'Artificial intelligence systems' refers to standalone or hardware-embedded software that acts as an intelligent agent or displays computational intelligence. An AI system can consist of one or more algorithms or models, but typically refers to complex systems in which multiple algorithms or models work together to perform a complex task.

Public discourse is currently dominated by concerns with a particular class of AI systems that make decisions and recommendations about important matters in life. These systems augment or replace analysis and decision-making by humans and are often used due to the scope or scale of data and rules involved. The number of features considered in classification tasks can run into the millions. This task replicates work previously undertaken by human workers, but on a much larger scale using qualitatively distinct decision-making logic. These systems make generally reliable (but not necessarily correct) decisions based upon complex rules that challenge or confound human capacities for action and comprehension.[18] In other words, this report addresses AI systems whose actions are difficult for humans to predict or whose decision-making logic is difficult to explain after the fact.

Common ethical challenges in AI

Prior review of the ethical challenges facing AI has identified six types of concerns that can be traced to the operational parameters of decision-making algorithms and AI systems. The map reproduced and adapted in Figure 1 takes into account:

> "decision-making algorithms (1) turn data into evidence for a given outcome (henceforth conclusion), and that this outcome is then used to (2) trigger and motivate an action that (on its own, or when combined with other actions) may not be ethically neutral. This work is performed in ways that are complex and

[18] Brent Mittelstadt et al., *The ethics of algorithms: Mapping the debate*, 3 BIG DATA & SOCIETY (2016), http://bds.sagepub.com/lookup/doi/10.1177/2053951716679679 (last visited Dec 15, 2016). The remainder of Section 2.1 draws heavily from the findings and ethical framework proposed in this landscaping study.

(semi-)-autonomous, which (3) complicates apportionment of responsibility for effects of actions driven by algorithms."[19]

From these operational characteristics, three epistemological and two normative types of ethical concerns can be identified based on how algorithms process data to produce evidence and motivate actions. The proposed five types of concerns can cause failures involving multiple human, organisational, and technological agents. This mix of human and technological actors leads to difficult questions concerning how to assign responsibility and liability for the impact of AI behaviours. These difficulties are captured in traceability as a final, overarching, type of concern.

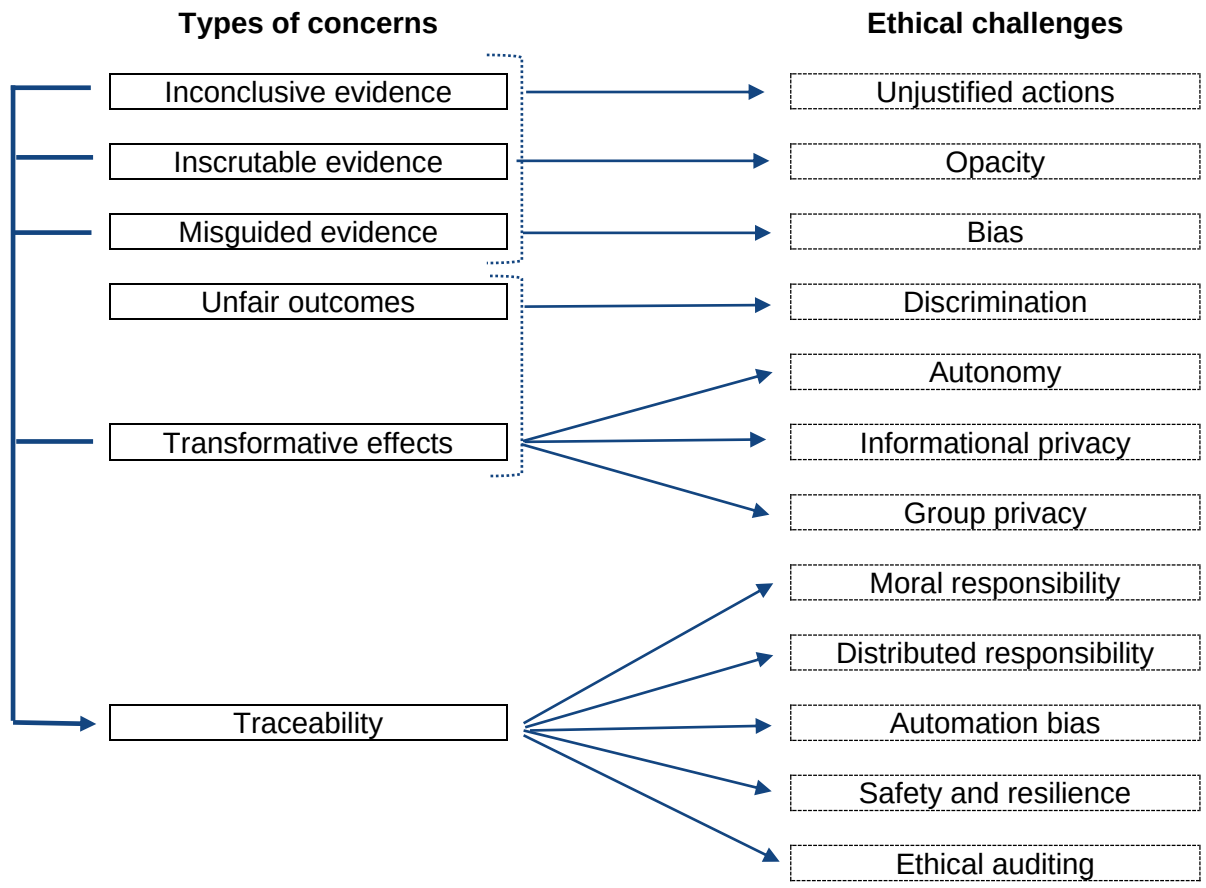

Figure 1 – Types of ethical concerns and challenges raised by algorithms (adapted from Mittelstadt et al., 2016)

The three aforementioned epistemological concerns with decision-making algorithms and AI systems can be defined as follows:

- **Inconclusive evidence** – When algorithms draw conclusions from the data they process using inferential statistics and/or machine learning techniques, they produce probable[20] yet inevitably uncertain knowledge. Statistical learning

[19] *Id.*

[20] The term 'probable knowledge' is used here in the sense of IAN HACKING, THE EMERGENCE OF PROBABILITY : A PHILOSOPHICAL STUDY OF EARLY IDEAS ABOUT PROBABILITY, INDUCTION AND STATISTICAL

theory[21] and computational learning theory[22] are both concerned with the characterisation and quantification of this uncertainty. Statistical methods can identify significant correlations, but correlations are typically not sufficient to demonstrate causality,[23] and thus may be insufficient to motivate action on the basis of knowledge of such a connection. The concept of an 'actionable insight' captures the uncertainty inherent in statistical correlations and normativity of choosing to act upon them.[24]

- **Inscrutable evidence** – When data are used as (or processed to produce) evidence for a conclusion, it is reasonable to expect that the connection between the data and the conclusion should be intelligible and open to scrutiny.[25] Given the complexity and scale of many AI systems, intelligibility and scrutiny cannot be taken for granted. A lack of access to datasets and the inherent difficulty of mapping how the multitude of data and features considered by an AI system contribute to specific conclusions and outputs cause practical as well as principled limitations.[26]

- **Misguided evidence** – Algorithms process data and are therefore subject to a limitation shared by all types of data processing, namely that the output can never exceed the input. The informal 'garbage in, garbage out' principle illustrates this phenomenon and its significance: conclusions can only be as reliable (but also as neutral) as the data they are based on.[27]

The three epistemic concerns detailed thus far address the quality of evidence produced by an algorithm that motivates a particular action. Normative concerns can be attached to these actions as well. There are two such potential normative concerns:

- **Unfair outcomes** – Algorithmically driven actions can be scrutinised from a variety of ethical perspectives, criteria, and principles. The normative acceptability of the action and its effects is observer-dependent and can be assessed independently of its epistemological quality. An action can be found discriminatory, for example, solely from its effect on a protected class of people, even if made on the basis of conclusive, scrutable and well-founded evidence.

- **Transformative effects** – The impact of AI systems cannot always be attributed to epistemic or ethical failures. Much of their impact can appear initially ethically neutral in the absence of obvious harm. A separate set of

---

INFERENCE (2006). where it is associated with the emergence of probability and the rise of statistical thinking (for instance in the context of insurance) that started in the 17th Century.

[21] GARETH JAMES ET AL., AN INTRODUCTION TO STATISTICAL LEARNING (2013).

[22] Leslie G. Valiant, *A theory of the learnable*, 27 COMMUNICATIONS OF THE ACM 1134–1142 (1984).

[23] PETER GRINDROD, MATHEMATICAL UNDERPINNINGS OF ANALYTICS: THEORY AND APPLICATIONS (2014).

[24] Boaz Miller & Isaac Record, *Justified belief in a digital age: on the epistemic implications of secret Internet technologies*, 10 EPISTEME 117–134 (2013).

[25] Hilary Kornblith, *Epistemology: Internalism and Externalism* (2001).

[26] Miller and Record, *supra* note 23.

[27] For a formal approach to the 'garbage in, garbage out' principle, see: CLAUDE E. SHANNON & WARREN WEAVER, THE MATHEMATICAL THEORY OF COMMUNICATION (1998).

impacts, which can be referred to as transformative effects, concern subtle shifts in how the world is conceptualised and organised.[28]

A final overarching concern addresses the need to specify common characteristics of AI systems and environmental conditions to ensure accountability and liability can be fairly apportioned across all actors and stakeholders involved in developing, deploying, and using AI systems:

- **Traceability** – AI systems often involve multiple agents which can include human developers and users, manufacturers and deploying organisations, and the systems and models themselves. AI systems can also interact directly, forming multi-agent networks characterised by rapid behaviours that avoid the oversight and comprehension of their human counterparts due to speed, scale, and complexity. As suggested in the original landscaping study by Mittelstadt et al., "algorithms are software-artefacts used in data-processing, and as such inherit the ethical challenges associated with the design and availability of new technologies and those associated with the manipulation of large volumes of personal and other data."[29] All of these factors mean it is difficult to detect harms, find their cause, and assign blame when AI systems behave in unexpected ways. Challenges arising through any of the aforementioned five types of concerns can thus raise a related challenge concerning traceability, wherein both the cause and responsibility for bad behaviours need to be established.[30]

As detailed in Figure 1, these types of concerns with decision-making algorithms and AI systems can be traced to widely discussed ethical challenges and concepts. In brief, according to this approach the following are some of the key ethical challenges arising from operational characteristics of decision-making algorithms and the six types of concerns described above[31]:

- **Unjustified actions** – Much algorithmic decision-making and data mining relies on inductive knowledge and correlations identified within a dataset. Correlations based on a 'sufficient' volume of data are often seen as sufficiently credible to direct action without first establishing causality.[32] Acting on correlations can be

[28] LUCIANO FLORIDI, THE FOURTH REVOLUTION: HOW THE INFOSPHERE IS RESHAPING HUMAN REALITY (2014).
[29] Mittelstadt et al., *supra* note 17.
[30] G. O. Mohler et al., *Self-Exciting Point Process Modeling of Crime*, 106 JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 100–108 (2011); Luciano Floridi, *Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions*, 374 PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A: MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES 20160112 (2016).
[31] Note: this list is adapted from a literature review conducted by the author and reported in the following: Mittelstadt et al., *supra* note 17.
[32] Mireille Hildebrandt, *Who Needs Stories if You Can Get the Data? ISPs in the Era of Big Number Crunching*, 24 PHILOS. TECHNOL. 371–390 (2011); Mireille Hildebrandt & Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73 THE MODERN LAW REVIEW 428–460 (2010); VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK AND THINK (2013); Tal Zarsky, *The Trouble with Algorithmic Decisions An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 118–132 (2016).

doubly problematic. Spurious correlations may be discovered rather than genuine causal knowledge. Even if strong correlations or causal knowledge are found, this knowledge may only concern populations while actions with significant personal impact are directed towards individuals.[33]

- **Opacity** – This is the 'black box' problem with AI: the logic behind turning inputs into outputs may not be known to observers or affected parties or may be fundamentally inscrutable or unintelligible. Opacity in machine learning algorithms is a product of the high dimensionality of data, complex code and changeable decision-making logic.[34] Transparency and comprehensibility are generally desired because algorithms that are poorly predictable or interpretable are difficult to control, monitor and correct.[35] Transparency is often naively treated as a panacea for ethical issues arising from new technologies.[36]

Information about the functionality of algorithms is often intentionally poorly accessible.[37] Besides being accessible, information must be comprehensible to be considered transparent.[38] Efforts to make algorithms transparent face a significant challenge to render complex decision-making processes both accessible and comprehensible. The longstanding problem of interpretability in machine learning algorithms indicates the challenge of opacity in algorithms.[39] In the context of medicine, the World Health Organization (WHO) has recognized the critical importance of combatting opacity through provisions to ensure transparency, 'explainability', and intelligibility in the design and usage of AI in healthcare.[40]

- **Bias** – The automation of human decision-making is often justified by an alleged lack of bias in AI and algorithms.[41] This belief is unsustainable; AI

---

[33] PHYLLIS MCKAY ILLARI & FEDERICA RUSSO, CAUSALITY : PHILOSOPHICAL THEORY MEETS SCIENTIFIC PRACTICE (2014).
[34] Jenna Burrell, *How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOCIETY (2016).
[35] ANDREW TUTT, *An FDA for Algorithms* (2016), http://papers.ssrn.com/abstract=2747994 (last visited Apr 13, 2016).
[36] Anjanette Raymond, *The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics*, NORTHWESTERN JOURNAL OF INTERNATIONAL LAW & BUSINESS, FORTHCOMING (2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469291 (last visited Jul 22, 2015); Kate Crawford, *Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 77–92 (2016); Daniel Neyland, *Bearing Account-able Witness to the Ethical Algorithmic System*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 50–76 (2016).
[37] Tasha Glenn & Scott Monteith, *New Measures of Mental State and Behavior Based on Data Collected From Sensors, Smartphones, and the Internet*, 16 CURR PSYCHIATRY REP 1–10 (2014); Meredith Stark & Joseph J. Fins, *Engineering Medical Decisions*, 22 CAMBRIDGE QUARTERLY OF HEALTHCARE ETHICS 373–381 (2013); Rob Kitchin, *Thinking critically about and researching algorithms*, INFORMATION, COMMUNICATION & SOCIETY 1–16 (2016); Matthias Leese, *The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union*, 45 SECURITY DIALOGUE 494–511 (2014).
[38] Matteo Turilli & Luciano Floridi, *The ethics of information transparency*, 11 ETHICS INF TECHNOL 105–112 (2009).
[39] Hildebrandt, *supra* note 31; Leese, *supra* note 36; Burrell, *supra* note 33; TUTT, *supra* note 34.
[40] World Health Organization, *supra* note 1 at xiii.
[41] Engin Bozdag, *Bias in algorithmic filtering and personalization*, 15 ETHICS INF TECHNOL 209–227 (2013); Gauri Naik & Sanika S. Bhide, *Will the future of knowledge work automation transform personalized medicine?*, 3 APPLIED & TRANSLATIONAL GENOMICS 50–53 (2014).

systems unavoidably make biased decisions.[42] A system's design and functionality reflects the values of its designer and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option. Development is not a neutral, linear path.[43] As a result, "the values of the author, wittingly or not, are frozen into the code, effectively institutionalising those values."[44] Inclusiveness and equity in both the design and usage of AI is thus key to combat implicit biases.[45] Friedman and Nissenbaum clarify that bias arise from (1) pre-existing social values found in the "social institutions, practices and attitudes" from which the technology emerges, (2) technical constraints and (3) emergent aspects of a context of use.[46]

- **Discrimination** – Discrimination against individuals and groups can arise from biases in AI systems. Discriminatory analytics can contribute to self-fulfilling prophecies and stigmatisation in targeted groups, undermining their autonomy and participation in society.[47] While a single definition of discrimination does not exist, legal frameworks internationally have a long history of jurisprudence discussing types of discrimination (e.g., direct and indirect), goals of equality law (e.g., formal and substantive equality), and appropriate thresholds for distribution of outcomes across groups. In this context, embedding considerations of non-discrimination and fairness into AI systems is particularly difficult.[48] It may be possible to direct algorithms not to consider sensitive attributes that contribute to discrimination,[49] such as gender or ethnicity,[50] based

---

[42] Kevin Macnish, *Unblinking eyes: the ethics of automating surveillance*, 14 ETHICS INF TECHNOL 151–167 (2012); Sue Newell & Marco Marabelli, *Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'*, 24 THE JOURNAL OF STRATEGIC INFORMATION SYSTEMS 3–14, 6 (2015); Bozdag, *supra* note 40; Batya Friedman & Helen Nissenbaum, *Bias in computer systems*, 14 ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS) 330–347 (1996); Omer Tene & Jules Polonetsky, *Big data for all: Privacy and user control in the age of analytics* (2013), http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (last visited Oct 2, 2014); Felicitas Kraemer, Kees van Overveld & Martin Peterson, *Is there an ethics of algorithms?*, 13 ETHICS AND INFORMATION TECHNOLOGY 251–260 (2011).

[43] JEFFREY ALAN JOHNSON, *Technology and Pragmatism: From Value Neutrality to Value Criticality* (2006), http://papers.ssrn.com/abstract=2154654 (last visited Aug 24, 2015).

[44] Macnish, *supra* note 41 at 158.

[45] World Health Organization, *supra* note 1 at xiii.

[46] Friedman and Nissenbaum, *supra* note 41.

[47] Macnish, *supra* note 41; Leese, *supra* note 36; Solon Barocas, *Data Mining and the Discourse on Discrimination* (2014), https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf (last visited Dec 20, 2015).

[48] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI*, 41 COMPUTER LAW & SECURITY REVIEW 105567 (2021); Sandra Wachter, Brent Mittelstadt & Chris Russell, *Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law*, 123 W. VA. L. REV. 735 (2020).

[49] SOLON BAROCAS & ANDREW D. SELBST, *Big Data's Disparate Impact* (2015), http://papers.ssrn.com/abstract=2477899 (last visited Oct 16, 2015).

[50] Toon Calders, Faisal Kamiran & Mykola Pechenizkiy, *Building classifiers with independency constraints*, *in* DATA MINING WORKSHOPS, 2009. ICDMW'09. IEEE INTERNATIONAL CONFERENCE ON 13–18 (2009); Faisal Kamiran & Toon Calders, *Classification with no discrimination by preferential sampling*, *in* PROC. 19TH MACHINE LEARNING CONF. BELGIUM AND THE NETHERLANDS (2010),

upon the emergence of discrimination in a particular context. However, proxies for protected attributes are not easy to predict or detect,[51] particularly when algorithms access linked datasets.[52]

- **Autonomy** – Value-laden decisions made by algorithms can also pose a threat to autonomy. Personalisation of content by AI systems, such as recommender systems, is particularly challenging in this regard. Personalisation can be understood as the construction of choice architectures which are not the same across a sample.[53] AI can nudge the behaviour of data subjects and human decision-makers by filtering information.[54] Different information, prices, and other content can be offered to profiling groups or audiences within a population defined by one or more attributes, for example the ability to pay, which can itself lead to discrimination. Personalisation reduces the diversity of information users encounter by excluding content deemed irrelevant or contradictory to the user's beliefs or desires.[55] This is problematic insofar as information diversity can be considered an enabling condition for autonomy.[56] The subject's autonomy in decision-making is disrespected when the desired choice reflects third-party interests above the individual's.[57]

A related challenge for autonomy concerns the intelligibility or comprehensibility of algorithmic systems and their outputs. Health professionals incorporating AI-based recommendations into their clinical care routines, for example, may experience a loss of autonomy if the basis for the recommendations is not well understood. Likewise, patients face a similar challenge when making informed decisions about their care based on AI recommendations. Recognising these risks, the WHO recognises "protecting human autonomy" as a key ethical principle for the design, usage, and governance of AI in healthcare due to the risk of decision-making power being transferred from humans to machines.[58]

- **Informational privacy and group privacy** – Algorithms also transform notions of privacy. Responses to discrimination, personalisation, and the inhibition of

---

http://wwwis.win.tue.nl/~tcalders/pubs/benelearn2010 (last visited Aug 24, 2015); Schermer, *supra* note 14.

[51] Zarsky, *supra* note 31; Andrea Romei & Salvatore Ruggieri, *A multidisciplinary survey on discrimination analysis*, 29 THE KNOWLEDGE ENGINEERING REVIEW 582–638 (2014).

[52] BAROCAS AND SELBST, *supra* note 48.

[53] Omer Tene & Jules Polonetsky, *Big data for all: Privacy and user control in the age of analytics*, NW.J. TECH. & INTELL. PROP. (2013), http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (last visited Oct 2, 2014).

[54] Mike Ananny, *Toward an Ethics of Algorithms Convening, Observation, Probability, and Timeliness*, 41 SCIENCE TECHNOLOGY HUMAN VALUES 93–117 (2016).

[55] ELI PARISER, THE FILTER BUBBLE : WHAT THE INTERNET IS HIDING FROM YOU (2011); Belinda A. Barnet, *Idiomedia: The rise of personalized, aggregated content*, 23 CONTINUUM 93–99 (2009).

[56] Jeroen van den Hoven & Emma Rooksby, *Distributive justice and the value of information: A (broadly) Rawlsian approach*, 376 INFORMATION TECHNOLOGY AND MORAL PHILOSOPHY (2008).

[57] Stark and Fins, *supra* note 36; Sally A. Applin & Michael D. Fischer, *New technologies and mixed-use convergence: How humans and algorithms are adapting to each other*, *in* 2015 IEEE INTERNATIONAL SYMPOSIUM ON TECHNOLOGY AND SOCIETY (ISTAS) 1–6 (2015).

[58] World Health Organization, *supra* note 1 at xii.

autonomy due to opacity often appeal to informational privacy,[59] or the right of data subjects to "shield personal data from third parties." Informational privacy concerns the capacity of an individual to control information about herself,[60] and the effort required by third parties to obtain this information. A right to identity derived from informational privacy suggests that opaque or secretive profiling is problematic when carried out by a third party. In a healthcare setting this could include insurers, remote care providers (e.g., chatbot and triage service providers), consumer technology companies, and others. Opaque decision-making inhibits oversight and informed decision-making concerning data sharing.[61] Data subjects cannot define privacy norms to govern all types of data generically because the value or insightfulness of data is only established through processing.[62]

Privacy protections based upon identifiability are poorly suited to limit external management of identity via analytics. Current regulatory protections struggle to address the informational privacy risks of analytics owing to the definition of 'personal data' being linked to an identified or identifiable individual; identifying a user is often unnecessary for purposes of algorithmic profiling and decision-making. Rather, knowledge is generated about algorithmically curated groups rather than uniquely identifiable individuals. Existing regulatory frameworks for privacy and data protection do not reflect the importance of profiling and groups to modern data analytics and automated decision-making.[63]

- **Moral responsibility and distributed responsibility** – When a technology fails, blame and sanctions must be apportioned.[64] Blame can only be justifiably attributed when the actor has some degree of control and intentionality in carrying out the action.[65] Traditionally, developers and software engineers have had "control of the behaviour of the machine in every detail" insofar as they can explain its overall design and function to a third party.[66] This traditional conception of responsibility in software design assumes the developer can reflect on the technology's likely effects and potential for malfunctioning,[67] and

[59] Schermer, *supra* note 14.
[60] L. Van Wel & L. Royakkers, *Ethical issues in web data mining*, 6 ETHICS AND INFORMATION TECHNOLOGY 129–140 (2004).
[61] Hojung Kim, Joseph Giacomin & Robert Macredie, *A Qualitative Study of Stakeholders' Perspectives on the Social Network Service Environment*, 30 INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION 965–976 (2014).
[62] Van Wel and Royakkers, *supra* note 59; Hildebrandt, *supra* note 31.
[63] Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30 PHILOSOPHY & TECHNOLOGY 475–494 (2017); 126 LINNET TAYLOR, LUCIANO FLORIDI & BART VAN DER SLOOT, GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES (2017), http://link.springer.com/book/10.1007/978-3-319-46608-8 (last visited Jan 18, 2017).
[64] Kraemer, van Overveld, and Peterson, *supra* note 41 at 251.
[65] Matthias, *supra* note 15.
[66] *Id.*
[67] Luciano Floridi, Nir Fresco & Giuseppe Primiero, *On malfunctioning software*, 192 SYNTHESE 1199–1220 (2014).

make design choices to choose the most desirable outcomes according to the functional specification.[68]

Justified allocation of moral responsibility is difficult for algorithms and AI systems with learning capacities. The traditional model for allocating responsibility in computing requires the system to be well-defined, comprehensible and predictable; complex and fluid systems (i.e., one with countless decision-making rules and lines of code) inhibit holistic oversight of decision-making pathways and dependencies. Machine learning algorithms are particularly challenging in this respect,[69] seen for instance in genetic algorithms that program themselves. The traditional model of responsibility fails because "nobody has enough control over the machine's actions to be able to assume the responsibility for them."[70] Distributed responsibility is thus a particular challenge for AI systems but could be addressed through application of strict liability or similar faultless responsibility schemes.

- **Automation bias** – A related problem concerns the diffusion of feelings of responsibility and accountability for users of AI systems, and the related tendency to trust the outputs of systems on the basis of their perceived objectivity, accuracy, or complexity.[71] Delegating decision-making to AI can shift responsibility away from human decision-makers. Similar effects can be observed in mixed networks of human and information systems as already studied in bureaucracies, characterised by reduced feelings of personal responsibility and the execution of otherwise unjustifiable actions.[72] Algorithms involving stakeholders from multiple disciplines can, for instance, lead to each party assuming others will shoulder ethical responsibility for the algorithm's actions.[73] Machine learning adds an additional layer of complexity between designers and actions driven by the algorithm, which may justifiably weaken blame placed upon the former.

- **Safety and resilience** – The need to apportion responsibility is acutely felt when algorithms malfunction. Unethical algorithms can be thought of as malfunctioning software artefacts that do not operate as intended. Useful distinctions exist between errors of design (types) and errors of operation (tokens), and between the failure to operate as intended (dysfunction) and the presence of unintended side-effects (misfunction).[74] Misfunctioning is distinguished from mere negative side effects by 'avoidability', or the extent to which comparable types of systems or artefacts accomplish the intended function without the effects in question. These distinctions clarify ethical aspects of AI systems that are strictly related to their functioning, either in the abstract

[68] Matthias, *supra* note 15.
[69] Burrell, *supra* note 33; Matthias, *supra* note 15; Zarsky, *supra* note 31.
[70] Matthias, *supra* note 15 at 177.
[71] Zarsky, *supra* note 31 at 121.
[72] HANNAH ARENDT, EICHMANN IN JERUSALEM: A REPORT ON THE BANALITY OF EVIL (1971).
[73] Michael Davis, Andrew Kumiega & Ben Van Vliet, *Ethics, Finance, and Automation: A Preliminary Survey of Problems in High Frequency Trading*, 19 SCIENCE AND ENGINEERING ETHICS 851–874 (2013).
[74] Floridi, Fresco, and Primiero, *supra* note 66.

(for instance when we look at raw performance), or as part of a larger decision-making system, and reveals the multifaceted interaction between intended and actual behaviour. Machine learning in particular raises unique challenges, because achieving the intended or "correct" behaviour does not imply the absence of errors or harmful actions and feedback loops.[75]

Both types of malfunctioning imply distinct responsibilities for algorithm and software developers, users and artefacts. Fair apportionment of responsibility for dysfunctioning and misfunctioning across large development teams and complex contexts of use is a difficult challenge. Requirements for resilience to malfunctioning as an ethical ideal in algorithm design need to be specified to ensure AI systems are both safe and resilient against dysfunctions and misfunctions. This reflects the ethical importance of human well-being and how it can be impacted by AI. Reflecting this, the WHO has explicitly recognized the importance of protecting human well-being and safety by enshrining it as a key ethical principle for usage of AI in healthcare.[76]

- **Ethical auditing** – How best to operationalise and set standards for testing of these ethical challenges remains an open question, particularly for machine learning. Merely rendering the code of an algorithm transparent is insufficient to ensure ethical behaviour. One possible path to achieve interpretability, fairness, and other ethical goals in AI systems is via auditing carried out by data processors,[77] external regulators,[78] or empirical researchers,[79] using ex post audit studies,[80] reflexive ethnographic studies in development and testing,[81] or reporting mechanisms designed into the algorithm itself.[82] For all types of AI, auditing is a necessary precondition to verify correct functioning. For systems with foreseeable human impact, auditing can create an ex post procedural record of complex automated decision-making to unpack problematic or inaccurate decisions, or to detect discrimination or similar harms.

[75] Except for trivial cases, the presence of false positives and false negatives in the work of algorithms, particularly machine learning, is unavoidable.
[76] World Health Organization, *supra* note 1 at xiii.
[77] Zarsky, *supra* note 31.
[78] TUTT, *supra* note 34; Zarsky, *supra* note 31; FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).
[79] Neyland, *supra* note 35; Kitchin, *supra* note 36.
[80] Christian Sandvig et al., *Auditing algorithms: Research methods for detecting discrimination on internet platforms*, DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY (2014), http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf (last visited Feb 13, 2016); Philip Adler et al., *Auditing Black-box Models by Obscuring Features*, ARXIV:1602.07043 [CS, STAT] (2016), http://arxiv.org/abs/1602.07043 (last visited Mar 5, 2016); Romei and Ruggieri, *supra* note 50; Kitchin, *supra* note 36; Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic investigation of computational power structures*, 3 DIGITAL JOURNALISM 398–415 (2015).
[81] Neyland, *supra* note 35.
[82] Alfredo Vellido, José David Martín-Guerrero & Paulo JG Lisboa, *Making machine learning models interpretable.*, 12 *in* ESANN 163–172 (2012).

The Oviedo Convention and human rights principles regarding health

The European Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine (ETS No. 164) of 1997, or the "Oviedo Convention," promotes the protection of human rights in biomedicine at a transnational level. The Oviedo Convention is a framework instrument meaning it contains general principles intended to be translated into domestic law by signatories. The Oviedo Convention contains many novel principles and requirements built on principles and rights contained in "previous international human rights treaties, such as the International Covenant on Civil and Political Rights of 1966 and the European Convention on Human Rights (ECHR) of 1950 (e.g. the rights to life, to physical integrity and to privacy, the prohibition of inhuman or degrading treatment and of any form of discrimination)."[83] The Oviedo Convention is inspired by and grounded in the rights to life, physical integrity and privacy, and prohibition of discrimination enacted through the ECHR. For the European Court of Human Rights, the Oviedo Convention has been used as an interpretative framework to elucidate and better understand the scope and significance of these rights in the context of biomedicine.[84]

The significance of these constituent human rights for the Oviedo Convention cannot be overstated. As a whole the Convention is designed to "protect the dignity and identity of all human beings and guarantee everyone, without discrimination, respect for their integrity and other rights and fundamental freedoms with regard to the application of biology and medicine" (Article 1). Across the Convention certain values and ends are explicitly upheld and protected, while others can be inferred through specific requirements. Above all, human dignity and the primacy of the patient are key to the Convention:

> "The notion of human dignity is clearly the bedrock of the Oviedo Convention. According to the Explanatory Report, "the concept of human dignity (...) constitutes the essential value to be upheld. It is at the basis of most of the values emphasised in the Convention." Recalling the history of the instrument, one of the members of the drafting group recognizes that "it was soon decided that the concept of dignity, identity and integrity of human beings/individuals should be both the basis and the umbrella for all other principles and notions that were to be included in the Convention.""[85]

Reference is made to other values and rights across the Oviedo Convention, such as the rights to life, physical integrity and privacy, and the prohibition of discrimination. For example, Article 10 reaffirms the right to privacy introduced in Article 8 of the

[83] Roberto Andorno, *The Oviedo Convention: A European Legal Framework at the Intersection of Human Rights and Health Law*, 2 133–143, 133 (2005).
[84] Francesco Seatzu & Simona Fanni, *The Experience of the European Court of Human Rights with the European Convention on Human Rights and Biomedicine*, 31 UTRECHT J. INT'L & EUR. L. 5–16 (2015).
[85] Andorno, *supra* note 82.

ECHR and the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data:

> 1. "Everyone has the right to respect for private life in relation to information about his or her health.
>
> 2. Everyone is entitled to know any information collected about his or her health. However, the wishes of individuals not to be so informed shall be observed."

Following the transparency requirements implied by the right to privacy in Article 10, Article 5 of the Oviedo Convention affirms the well-established requirement for informed consent in medicine:

> "An intervention in the health field may only be carried out after the person concerned has given free and informed consent to it.
>
> This person shall beforehand be given appropriate information as to the purpose and nature of the intervention as well as on its consequences and risks.
>
> The person concerned may freely withdraw consent at any time."

According to the Explanatory Report, the requirement for consent "makes clear patients' autonomy in their relationship with health care professionals and restrains the paternalist approaches which might ignore the wish of the patient." Paragraphs 35 and 36 of the Report provide further details on the specific requirements for consent to be considered free and informed including constraints on the doctor's influence on a patient's decision and requirements concerning the quality, breadth, and clarity of information provided:

> "35. The patient's consent is considered to be free and informed if it is given on the basis of objective information from the responsible health care professional as to the nature and the potential consequences of the planned intervention or of its alternatives, in the absence of any pressure from anyone. Article 5, paragraph 2, mentions the most important aspects of the information which should precede the intervention but it is not an exhaustive list: informed consent may imply, according to the circumstances, additional elements. In order for their consent to be valid the persons in question must have been informed about the relevant facts regarding the intervention being contemplated. This information must include the purpose, nature and consequences of the intervention and the risks involved. Information on the risks involved in the intervention or in alternative courses of action must cover not only the risks

> inherent in the type of intervention contemplated, but also any risks related to the individual characteristics of each patient, such as age or the existence of other pathologies. Requests for additional information made by patients must be adequately answered.
>
> 36. Moreover, this information must be sufficiently clear and suitably worded for the person who is to undergo the intervention. The patient must be put in a position, through the use of terms he or she can understand, to weigh up the necessity or usefulness of the aim and methods of the intervention against its risks and the discomfort or pain it will cause."

Article 10 provides both a "right to know" and a "right not to know" about their health status and any information collected about their health. These rights are core elements of the doctor-patient relationship envisioned in the Oviedo Convention. If patients are entitled to make an informed decision about their care, it follows that they are entitled to receive adequate information to make that decision in an informed manner.[86]

Concerning discrimination, Article 11 explicitly prohibits discrimination on the grounds of genetic heritage. Likewise, Article 3 provides for equitable access to healthcare of an appropriate quality:

> "Parties, taking into account health needs and available resources, shall take appropriate measures with a view to providing, within their jurisdiction, equitable access to health care of appropriate quality."

Inequality in access to care or standards of care could be considered a violation of the prohibition on discrimination contained in Article 14 of the ECHR, in particular in relation to discrimination in "association with a national minority, property, birth or other status" (see section entitled "Inequality in access to high quality healthcare"). Similarly, Article 4 addresses quality of care and professional standards in healthcare and research:

> "Any intervention in the health field, including research, must be carried out in accordance with relevant professional obligations and standards."

The Oviedo Convention understandably does not specify quality standards to be met in healthcare and research, but rather leaves the determination of standards to professional bodies and domestic law of signatories of the Convention according to local health needs and available resources. With that said, as the Convention prescribes a minimum standard for human rights protections, member states can

[86] *Id.*

choose to enact higher standards in their translation of the Convention into domestic law. With regards to quality of care standards, this can be done in relation to Articles 3 and 4. Paragraph 30 of the Explanatory Report clarifies the parties envisioned as setting these professional obligations and standards:

> "30. All interventions must be performed in accordance with the law in general, as supplemented and developed by professional rules. In some countries these rules take the form of professional codes of ethics (drawn up by the State or by the profession), in others codes of medical conduct, health legislation, medical ethics or any other means of guaranteeing the rights and interests of the patient, and which may take account of any right of conscientious objection by health care professionals."

Paragraphs 31 and 32 elaborate on the nature of medicine as a profession, variation in standards across countries, the commitment of doctors to uphold ethical and legal standards, and the content and development of standards over time:

> "31. The content of professional standards, obligations and rules of conduct is not identical in all countries. The same medical duties may vary slightly from one society to another. However, the fundamental principles of the practice of medicine apply in all countries. Doctors and, in general, all professionals who participate in a medical act are subject to legal and ethical imperatives. They must act with care and competence, and pay careful attention to the needs of each patient.
>
> 32. It is the essential task of the doctor not only to heal patients but also to take the proper steps to promote health and relieve pain, taking into account the psychological well-being of the patient. Competence must be determined primarily in relation to the scientific knowledge and clinical experience appropriate to a profession or speciality at a given time. The current state of the art determines the professional standard and skill to be expected of health care professionals in the performance of their work. In following the progress of medicine, it changes with new developments and eliminates methods which do not reflect the state of the art. Nevertheless, it is accepted that professional standards do not necessarily prescribe one line of action as being the only one possible: recognised medical practice may, indeed, allow several possible forms of intervention, thus leaving some freedom of choice as to methods or techniques."

Following this, Paragraph 33 of the Explanatory Report provides a brief indication of the ideal model for the doctor-patient relationship with respect to choosing interventions:

> “33. Further, a particular course of action must be judged in the light of the specific health problem raised by a given patient. In particular, an intervention must meet criteria of relevance and proportionality between the aim pursued and the means employed. Another important factor in the success of medical treatment is the patient's confidence in his or her doctor. This confidence also determines the duties of the doctor towards the patient. An important element of these duties is the respect of the rights of the patient. The latter creates and increases mutual trust. The therapeutic alliance will be strengthened if the rights of the patient are fully respected.”

The Oviedo Convention thus specifies a number of rights and requirements relating to or derived from human rights protected in other contexts. Key values and interests can be derived from the topics addressed throughout the Convention. These values embedded in human rights principles regarding health can guide the development of a theoretical framework for the doctor-patient relationship. Specifically, the Oviedo Convention prescribes and discusses the following values:

- **Human dignity**
- **Primacy of patient interests over societal and scientific interests**
- **Right to life**
- **Physical integrity**
- **Privacy and identity**
- **Informed consent**
- **Right to know and right not to know**
- **Prohibition of discrimination and inequality in access to healthcare**
- **Quality of care standards**

In the section entitled “Theoretical framework of the doctor-patient relationship”, these values will be discussed in the context of the goals of medicine as a profession and societal good and used as the basis to develop a theoretical framework for the doctor-patient relationship. This framework, and the values underpinning it derived from the Convention, suggests that certain goods must be met in the doctor-patient relationship. Likewise, different models for clinical encounters and the doctor-patient relationship will align better or worse with these values. These issues will be picked up in the aforementioned section following a brief overview of AI systems in medicine.

To situate this report in ongoing policy work by the Council of Europe, it is important to briefly note recent reports that have addressed other areas of work relevant to the impact of AI in healthcare. The “Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 223)” was opened in October 2018 and is set to be ratified in October 2023. The Protocol amends Convention ETS No. 108. Of particular relevance to AI in medicine is its

revision of Article 8 (now Article 9) of the Convention to grant individuals a variety of data protection rights:

> 1. "Every individual shall have a right:
>
>   a. Not to be subject to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration;
>   b. to obtain, on request, at reasonable intervals and without excessive delay or expense, confirmation of the processing of personal data relating to him or her, the communication in an intelligible form of the data processed, all available information on their origin, on the preservation period as well as any other information that the controller is required to provide in order to ensure the transparency of processing in accordance with Article 8, paragraph 1;
>   c. to obtain, on request, knowledge of the reasoning underlying data processing where the results of such processing are applied to him or her;
>   d. to object at any time, on grounds relating to his or her situation, to the processing of personal data concerning him or her unless the controller demonstrates legitimate grounds for the processing which override his or her interests or rights and fundamental freedoms;
>   e. to obtain, on request, free of charge and without excessive delay, rectification or erasure, as the case may be, of such data if these are being, or have been, processed contrary to the provisions of this Convention;
>   f. to have a remedy under Article 12 where his or her rights under this Convention have been violated;
>   g. to benefit, whatever his or her nationality or residence, from the assistance of a supervisory authority within the meaning of Article 15, in exercising his or her rights under this Convention."

Many of these rights mirror protections in the General Data Protection Regulation (GDPR), a data protection framework implemented by the European Commission in 2018, including a limited right not to be subject to an automated decision, a right to obtain information on data processing, and rights to request rectification and erasure of personal data.[87] These rights may come provide an important backbone to protect

[87] Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INTERNATIONAL DATA PRIVACY LAW 76–99 (2017); Sandra Wachter & B. D. Mittelstadt, *A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI*, 2019 COLUMBIA BUSINESS LAW REVIEW (2019).

the ideal of informed consent in medical applications of AI by providing access to information about the scope and nature of automated processing.

The October 2020 report "Artificial intelligence in health care: medical, legal and ethical challenges ahead," published by the Parliamentary Assembly of the Council of Europe and drafted by its Committee on Social Affairs, Health and Sustainable Development, proposed a draft recommendation responding to the growing impact of AI in healthcare.[88] The report's explanatory memorandum discusses in great detail the various medical, legal, and ethical impacts envisioned for AI, which include:

- **Need for ethical review in biomedical research and limitations on competences and capacities of ethics review bodies to assess the unique risks and opportunities of AI**
- **Liability of AI providers in medicine and healthcare**
- **Protection of personal data in the context of harmonising data systems and supporting AI innovation and research in Europe, in particular**
- **Ensuring lawfulness, fairness, purpose specification, proportionality, privacy-by-design and default, responsibility, compliance, transparency, data security, and risk management**
- **Challenges of guaranteeing meaningful control and informed consent for patients and other data subjects**
- **Positive obligations for states to protect life and health via national reporting mechanisms**
- **Navigating the tension between "freedom to innovate" and meaningful protection of human rights**

Rather than being discussed in detail here, these and other points raised in prior reports from the Council of Europe are reflected in the discussion of potential impacts on the doctor-patient relationship in the section entitled "Potential impact of AI on the doctor-patient relationship".

[88] COUNCIL OF EUROPE, *supra* note 2.

4 OVERVIEW OF AI TECHNOLOGIES IN MEDICINE

As described in the section entitled "Background and context", a broad array of technologies can be described as AI. With high-level definitions of relevant concepts including artificial intelligence, algorithms, and machine learning are defined, it is necessary to explore in more detail the potential types of medical AI applications. As this report focuses on the impact of AI on the doctor-patient relationship, not all potential medical applications will be considered. As a first step, we can distinguish between three types of AI according to their intended users:

- **AI for biomedical researchers**
- **AI for patients**
- **AI for health professionals**

> Of these categories, AI for patients and health professionals are most relevant for the purposes of this report given the focus on the doctor-patient relationship.

Other taxonomies are of course possible; a recent report by the WHO, for example, distinguishes between AI applications for use in:

- **Health care**
- **Health research and drug development**
- **Health systems management and planning**
- **Public health and public health surveillance**

The taxonomy deployed here focuses on the intended users of AI systems because appropriate solutions to ethical challenges introduced by these systems typically vary according to the interests, level of expertise, and requirements of different stakeholder groups.

Although not directly relevant to the doctor-patient relationship, it is worth reviewing a few examples of AI used for medical research. One of the most common applications in biomedical research is drug discovery. For example, a recent discovery by computer scientists and cancer specialists at the Institute of Cancer Research and Royal Marsden NHS Foundation Trust of a new drug regime for a rare form of brain cancer in children (diffuse intrinsic pontine glioma).[89] Deepmind's recent advances on protein folding via AlphaFold likewise indicate the promise of AI for fundamental research.[90]

[89] Andrew Gregory, *Scientists use AI to create drug regime for rare form of brain cancer in children*, THE GUARDIAN, September 22, 2021, https://www.theguardian.com/science/2021/sep/23/scientists-use-ai-to-create-drug-regime-for-rare-form-of-brain-cancer-in-children (last visited Sep 26, 2021); Carvalho et al., *supra* note 7.
[90] Jumper et al., *supra* note 7.

AI can also be used for structuring, labelling, and searching unorganized or heterogenous medical datasets; image classifiers, for example, can process huge volumes of medical imaging data much faster than manual labellers. Such systems can also be useful for administrative and operational purposes as discussed below.

One noteworthy usage of AI that blurs the boundaries between research and clinical care is that of polygenic embryo screening, in which an algorithm summarizes "the estimated effect of hundreds or thousands of genetic variants associated with an individual's risk of having a particular condition or trait." This practice raises the spectre of eugenics by potentially allowing parents to select embryos both for health advantages, but also for socially desirable non-disease-related traits.[91]

Many AI applications are in development to be used directly by patients, often in collaboration with a health professional or artificial agent. These include telemedicine applications used for remote observation, clinical encounters, and video-observed therapy; virtual assistants and chat bots for information or triage; applications for managing chronic illnesses such as cardiovascular disease or hypertension; health and well-being 'apps'; personal health monitoring systems including wearables with built-in analytics and behavioural recommendations; and remote monitoring systems for facial recognition, gait detection, biometrics, and health-related behaviours.[92]

One purported benefit of AI systems aimed at patients it to "empower patients and communities to assume control of their own health care and better understand their evolving needs."[93] Health monitoring and telemedicine systems could, for example, assist patients in self-management of chronic conditions like diabetes, hypertension, or cardiovascular disease.[94] Therapeutic "chat bots" may also be able to assist in management of mental health conditions.[95] It has been predicted, for example, that the GPT-3 natural language application could eventually be used as the basis for conversational agents working directly with patients, for example as an initial point of contact or (more controversially) for triaging non-critical patients.[96] These applications seem highly likely given the existing deployment of 'virtual GP' chat bots which direct service enquiries and provide information to patients[97]; it should be noted, however, that such applications have been the subject of significant debate over their ethical

[91] Sheetal Soni & Julian Savulescu, *Polygenic Embryo Screening: Ethical and Legal Considerations*, THE HASTINGS CENTER (2021), https://www.thehastingscenter.org/polygenic-embryo-screening-ethical-and-legal-considerations/ (last visited Nov 23, 2021).
[92] Mittelstadt et al., *supra* note 3.
[93] World Health Organization, *supra* note 1.
[94] Mittelstadt et al., *supra* note 3; SECRETARY OF STATE FOR HEALTH AND SOCIAL CARE, *The Topol Review: Preparing the healthcare workforce to deliver the digital future* (2019), https://topol.hee.nhs.uk/.
[95] SECRETARY OF STATE FOR HEALTH AND SOCIAL CARE, *supra* note 93.
[96] Diane M. Korngiebel & Sean D. Mooney, *Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery*, 4 NPJ DIGITAL MEDICINE 1–3 (2021).
[97] Weiyu Wang & Keng Siau, *Trust in health chatbots* (2018); Claire Woodcock et al., *The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study*, 23 JOURNAL OF MEDICAL INTERNET RESEARCH e29386 (2021).

acceptability and regulation.[98] Likewise, they may lead to reduced access to human care.[99]

Finally, a wide variety of applications are aimed at health professionals. Three broad categories can be distinguished:

- **Applications designed for diagnostics, therapeutics, and other forms of clinical care**
- **Applications designed for operational or administrative uses**
- **Applications designed for public health surveillance**

The distinction between these categories is not always clear, as will be discussed below. To limit the focus of this report to the potential impact of AI on the doctor-patient relationship, only the first two categories will be surveyed. Public health surveillance could also be conceived as an extension of the clinical experience or doctor-patient relationship, insofar as patients may be contacted proactively by public health officials for clinical follow-up. Nonetheless, this report is concerned principally with the immediate clinical experience and relationship between individual health professionals and their patients.

AI systems aimed at clinical care are designed to fulfil a broad range of tasks, including diagnosis recommendations, optimization of treatment plans, and various other forms of decision-support.

According to the WHO:

> "AI is being evaluated for use in radiological diagnosis in oncology (thoracic imaging, abdominal and pelvic imaging, colonoscopy, mammography, brain imaging and dose optimization for radiological treatment), in non-radiological applications (dermatology, pathology), in diagnosis of diabetic retinopathy, in ophthalmology and for RNA and DNA sequencing to guide immunotherapy."[100]

Future applications currently in development (but not yet deployed clinically) include systems to detect "stroke, pneumonia, breast cancer by imaging,[101] coronary heart

[98] GARETH IACOBUCCI, ROW OVER BABYLON'S CHATBOT SHOWS LACK OF REGULATION (2020); Wang and Siau, *supra* note 96.
[99] World Health Organization, *supra* note 1.
[100] Wenya Linda Bi et al., *Artificial intelligence in cancer imaging: clinical challenges and applications*, 69 CA: A CANCER JOURNAL FOR CLINICIANS 127–157 (2019); World Health Organization, *supra* note 1.
[101] Pranav Rajpurkar et al., *Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists*, 15 PLOS MEDICINE e1002686 (2018); Babak Ehteshami Bejnordi et al., *Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer*, 318 JAMA 2199–2210 (2017).

disease by echocardiography[102] and detection of cervical cancer,"[103] including systems designed specifically for use in low- and middle-income countries (LMIC).[104] Systems are being designed to predict the risk of lifestyle diseases including cardiovascular disease[105] and diabetes.[106]

Development of medical image classification systems has been highly prevalent in recent years. Prior work, for example, has shown that neural networks can achieve consistently higher sensitivity for pathological findings in radiology.[107] Image classification systems can also be used to support detection of tuberculosis,[108] COVID-19, and other conditions through interpreting staining images[109] or X-rays.[110] Another emerging phenomenon is that of "digital twins," which are systems that simulate individual organs or multi-organ systems of individual patients for purposes of disease modelling and prediction.[111]

Generally speaking, the deployment of AI in clinical care remains nascent. Clinical efficacy has been established for relatively few systems when compared to the significant research activity in healthcare applications of AI. Research, development, and pilot testing often do not translate into proven clinical efficacy, commercialization, or widespread deployment. The generalization of performance from trials to clinical practice generally remains unproven.[112]

A 2019 meta-analysis of deep-learning image classifiers in healthcare found that despite claims of equivalent accuracy between AI systems and human healthcare professionals:

---

[102] Maryam Alsharqi et al., *Artificial intelligence and echocardiography*, 5 ECHO RESEARCH AND PRACTICE R115–R125 (2018).
[103] Using Artificial Intelligence to Detect Cervical Cancer, , NIH DIRECTOR'S BLOG (2019), https://directorsblog.nih.gov/2019/01/17/using-artificial-intelligence-to-detect-cervical-cancer/ (last visited Dec 1, 2021).
[104] World Health Organization, *supra* note 1; Innovative, affordable screening and treatment to prevent cervical cancer, , UNITAID , https://unitaid.org/project/innovative-affordable-screening-and-treatment-to-prevent-cervical-cancer/ (last visited Dec 1, 2021).
[105] Rui Fan et al., *AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus*, 10 SCIENTIFIC REPORTS 1–8 (2020); Yang Yan et al., *The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine?*, 16 JOURNAL OF GERIATRIC CARDIOLOGY: JGC 585 (2019).
[106] Jyotismita Chaki et al., *Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review*, JOURNAL OF KING SAUD UNIVERSITY-COMPUTER AND INFORMATION SCIENCES (2020).
[107] Ohad Oren, Bernard J Gersh & Deepak L Bhatt, *Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints*, 2 THE LANCET DIGITAL HEALTH e486–e488 (2020).
[108] Yan Xiong et al., *Automatic detection of mycobacterium tuberculosis using artificial intelligence*, 10 JOURNAL OF THORACIC DISEASE 1936 (2018).
[109] *Id.*
[110] Apoorva Mandavilli, *These Algorithms Could Bring an End to the World's Deadliest Killer*, THE NEW YORK TIMES, November 20, 2020, https://www.nytimes.com/2020/11/20/health/tuberculosis-ai-apps.html (last visited Dec 1, 2021).
[111] Matthias Braun, *Represent me: please! Towards an ethics of digital twins in medicine*, J MED ETHICS (2021).
[112] World Health Organization, *supra* note 1 at 6.

> "Few studies present externally validated results or compare the performance of deep learning models and health-care professionals using the same sample." Likewise, "poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy."[113]

The evidence base for clinical efficacy of deep learning systems may have improved in subsequent years, but broad adoption will seemingly hinge on standardised reporting of accuracy to enable assessment of clinical efficacy by medical regulators and clinical care excellence bodies.

A near term challenge for image classifiers is to build systems which can assess multiple image or scan types, such as X-rays and CT scans, which are often considered in combination by human radiologists while AI systems typically can only interpret one or the other. A similar challenge exists for detection of multiple conditions or pathologies, with existing classifiers often trained to only detect a single type of abnormality.[114]

Finally, many AI systems are also designed for administrative or operational purposes. AI systems can help with several aspects of hospital administration and operational evaluations. Discharge planning tools, for instance, can estimate discharge dates and barriers for hospitalized patients and flag up those that are clinically (nearly) ready to be discharged to clinicians, along with a list of necessary steps to take prior to discharge. Some systems can even schedule necessary follow-up appointments and care.[115] Natural language processing systems could be used for automation of routine or labour-intensive tasks, such as searching and navigation of electronic health record (EHR) systems or automated preparation of medical documentation and orders.[116] According to the WHO, "Clinicians might use AI to integrate patient records during consultations, identify patients at risk and vulnerable groups, as an aid in difficult treatment decisions and to catch clinical errors. In LMIC, for example, AI could be used in the management of antiretroviral therapy by predicting resistance to HIV drugs and disease progression, to help physicians optimize therapy."[117]

Distinguishing between uses of AI for clinical care and research versus those used for operational and quality improvement purposes by hospitals and health systems is often difficult. Many such systems are designed to identify at-risk patients. The UCLA Health network, for example, uses a tool that identified patients in primary care that are at high risk of being hospitalized or making frequent visits to an emergency room in the coming year. Similarly, Oregon Health and Science University use a regression

---

[113] Xiaoxuan Liu et al., *A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis*, 1 THE LANCET DIGITAL HEALTH e271–e297 (2019).
[114] Stephanie Price, *Technological innovations of AI in medical diagnostics*, HEALTH EUROPA (2020), https://www.healtheuropa.eu/technological-innovations-of-ai-in-medical-diagnostics/103457/ (last visited Sep 6, 2021).
[115] Robbins and Brodwin, *supra* note 5.
[116] Korngiebel and Mooney, *supra* note 95.
[117] World Health Organization, *supra* note 1; Jerome Amir Singh, *Artificial Intelligence and global health: opportunities and challenges*, 3 EMERGING TOPICS IN LIFE SCIENCES 741–746 (2019).

algorithm to monitor patients across the hospital for signs of sepsis.[118] Both are treated as a type of operational tool for monitoring and prioritising quality of care, and not as part of clinical care or research.

[118] Robbins and Brodwin, *supra* note 5.

5 THEORETICAL FRAMEWORK OF THE DOCTOR-PATIENT RELATIONSHIP

Health is a fundamental good valued across many contexts, including personal, social and economic life, related to the maintenance and well-being of the whole person. Without health personal plans cannot be made, projects pursued, or identities created without restrictions imposed by a physical, mental or social ailment.[119] Health is therefore a prerequisite for the realisation of other human goods.

Broadly speaking, the end of medicine is to guarantee the health of a society and individuals within it.[120] Despite the difficulties of defining health and illness as concepts, medicine is broadly recognised as a practice to promote health, thereby working towards a fundamental good.[121] A lack of agreement on a 'correct' definition of health, reflected in debate on the topic, does not undermine the fundamental value of health to human life.[122] The ends of medicine are achieved through 'good' medical encounters with individual patients.[123] In pursuing these ends in the doctor-patient relationship, moral and technical capacities must work together in the interests of the patient because medical activity affects individuals with moral worth and interests.

As discussed in the section entitled "The Oviedo Convention and human rights principles regarding health", the Oviedo Convention prescribes the following values:

- **Human dignity**
- **Primacy of patient interests over societal and scientific interests**
- **Right to life**
- **Physical integrity**
- **Privacy and identity**
- **Informed consent**

[119] Andrew Edgar, *The expert patient: Illness as practice*, 8 MEDICINE, HEALTH CARE AND PHILOSOPHY 165–171 (2005).
[120] WORLD HEALTH ORGANIZATION, *Preamble to the Constitution of the World Health Organization* (1948); KENNETH WILLIAM MUSGRAVE FULFORD, MORAL THEORY AND MEDICAL PRACTICE (1989).
[121] FULFORD, *supra* note 119; EDMUND D PELLEGRINO & DAVID C THOMASMA, THE VIRTUES IN MEDICAL PRACTICE (1993); Paul Schotsmans, Bernadette Dierckx de Casterle & Chris Gastmans, *Nursing considered as moral practice: a philosophical-ethical interpretation of nursing*, 8 KENNEDY INSTITUTE OF ETHICS JOURNAL 43–69 (1998).
[122] FULFORD, *supra* note 119; Alan Petersen, *Risk, governance and the new public health*, *in* FOUCAULT: HEALTH AND MEDICINE 189–206 (Alan Petersen & Robin Bunton eds., 1997); Adele E. Clarke et al., *Biomedicalization: Technoscientific transformations of health, illness, and U.S. biomedicine*, 68 AMERICAN SOCIOLOGICAL REVIEW 161–194 (2003).
[123] ALASDAIR MACINTYRE, AFTER VIRTUE: A STUDY IN MORAL THEORY (3rd Revised edition ed. 2007); PELLEGRINO AND THOMASMA, *supra* note 120; GENERAL MEDICAL COUNCIL, *Good Medical Practice* (2013), http://www.gmc-uk.org/static/documents/content/GMP_2013.pdf_51447599.pdf.

- **Right to know and right not to know**
- **Prohibition of discrimination and inequality in access to healthcare**
- **Quality of care standards**

These values, and the different goals of medicine as a practice, can be realised through different types of doctor-patient relationships. Models of the (ideal) doctor-patient relationship have adapted over time in recognition of the growing importance of patient autonomy and its appropriate balance with other ethical obligations of the doctor towards beneficence, non-maleficence, and justice.[124] An influential paper from Emanuel and Emanuel (1992) proposed four models for the doctor-patient relationship:

- **Paternalistic Model** – This model vests the vast majority of decision-making power in the doctor. It assumes the existence of shared, objective values or criteria to define the best course of action to promote the patient's health and well-being. The doctor's role is expert, skilled practitioner tasked with "promoting the patient's well-being independent of the patient's current preferences." The doctor acts as "the patient's guardian, articulating and implementing what is best for the patient." Autonomy is realised only through patient assent to the doctor's determination of the best course of action.
- **Informative Model** – In contrast, this model vests the vast majority of decision-making power in the patient. The objective of clinical interactions "is for the doctor to provide the patient with all relevant information, for the patient to select the medical interventions he or she wants, and for the doctor to executive the selected interventions." Objectives values are not assumed; rather, the patient's values and interests are taken as known or fixed to the patient but not the doctor. The doctor's role is to provide facts to facilitate the patient making a decision that bests matches their interests.
- **Interpretive Model** – This model closely follows the informative model but provides a greater role for the doctor to assist the patient in understanding her values and interests, and the possible impact of different interventions in these terms. The doctor acts as an advisor to help the patient "elucidate and make coherent" their values but does not pass judgement on these values or attempt to prioritize them on behalf of the patient. The ultimate choice of intervention still rests with the patient in the interpretive model, but the doctor plays a more active role in shaping this choice than the informative model.
- **Deliberative Model** – This model closely follows the interpretive model but gives the doctor a greater role in judging and prioritizing patient values. It is the doctor's role to "elucidate the types of values embodied in the available

[124] TOM L. BEAUCHAMP & JAMES F. CHILDRESS, PRINCIPLES OF BIOMEDICAL ETHICS (2009); E. J. Emanuel & L. L. Emanuel, *Four models of the physician-patient relationship*, 267 JAMA: THE JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION 2221–2226 (1992).

> options…suggesting why certain health-related values are more worthy and should be aspired to." Deliberation between the doctor and patient remains limited to "health-related values, that is, values that affect or are affected by the patient's disease and treatments; he or she recognizes that many elements of morality are unrelated to the patient's disease or treatment and beyond the scope of their professional relationship." The aim of the deliberation is moral persuasion, but not coercion, with the patient ultimately deciding on the appropriate validity and priority of these values in their life. Whereas the doctor is an advisor or counsellor in the interpretive model, in the deliberative model they serve as "a teacher or friend, engaging the patient in dialogue on what course of action would be best." The doctor indicates both what the patient could do and, in the context of their understanding of the patient's life and values, what he thinks the patient should do in terms of choice of intervention. The final decision still remains with the patient but is subject to greater persuasion and normative argumentation on the part of the doctor. This model conceives of patient autonomy as a tool for moral self-development; "the patient is empowered not simply to follow unexamined preferences or examined values, but to consider, through dialogue, alternative health-related values, their worthiness, and their implications for treatment."

A fifth model is mentioned in Emanuel and Emanuel's treatment of the doctor-patient relationship, the 'instrumental model', but quickly discarded on moral grounds. In the instrumental model the patient's values are given no importance; rather, the doctor takes a decision or convinces the patient to choose a particular course of treatment on the basis of external values such as social or scientific good. While rightly condemned on moral grounds, it should be noted that this model remains potentially relevant as a warning for the deployment of AI. In cases where AI is pursued not for the good of the patient, but rather for the sake of efficiency or cost savings, one could argue the doctor-patient relationship is instrumentalized. The influence of such external values on the doctor-patient relationship are elaborated below.

Each of these models of the doctor-patient relationship show varying degrees of respect to patient autonomy and moral self-development. The rights and values embedded in the Oviedo Convention provide some indication of the general acceptability of these models of the doctor-patient relationship. A paternalistic model would appear prone to violating the informed consent requirement set out in Article 5. A deliberative model would likewise appear to violate a specific aspect of the consent requirement expanded on in the Convention's Explanatory Report: a patient's consent should be based on "objective information" provided "in the absence of any pressure from anyone." The difficulty of providing objective information will be picked up again in the section entitled "Potential impact of AI on the doctor-patient relationship" in discussing transparency in AI-mediated clinical care.

Professional ethics in medicine

The Oviedo Convention explicitly calls for quality standards to be set by member states and professional societies in Article 4. But how does medicine as a profession set its own standards for clinical care and the doctor-patient relationship, and according to which goals or values? To this end, this section proposes a theoretical framework for understanding medicine as a self-governing profession. This framework aligns with many of the values prescribed in the Oviedo Convention; this aspect is further discussed in the section entitled "Potential impact of AI on the doctor-patient relationship".

An influential approach which prescribes ideal ends (and thus norms and internal goods) of medicine based upon virtue ethics has been advanced by Pellegrino and Thomasma.[125] Within this approach, based upon Alisdair MacIntyre's virtue ethics,[126] medicine can be considered a "moral practice"[127] with virtues describing character traits required of doctors in addition to the "medical scientific knowledge, practical skills and experience that ensures that the doctor does the right things with the right attitude in order to reach the goals of medicine."[128] Medicine is a moral practice by MacIntyre's definition because as a profession it self-governs, defines, and upholds internal standards of good medical care and accreditation processes to uphold these standards.[129]

The telos of a practice can be understood through critical examination of its internal goods or norms of evaluation; for medicine, these norms can be found in the doctor-patient relationship.[130] As seen in this relationship, "the ends of medicine are...the restoration or improvement of health and, more proximately, to heal, that is, to cure illness and disease or, when this is not possible, to care for and help the patient to live with residual pain, discomfort or disability."[131] The doctor-patient relationship, understood as a type of "healing relationship," is the primary mechanism through which these ends are realised.

Treating medicine as a moral practice with norms of good practice realised through a healing relationship is not to adapt an antiquated view of medicine as a paternalistic patient-provider relationship. Rather, the healing relationship involves both clinical interventions and information or services provided to patients for the sake of knowledge, empowerment or self-care. Even in modern clinical encounters with patients 'empowered' with democratised access to medical information, personal

[125] PELLEGRINO AND THOMASMA, *supra* note 120 at 52.
[126] MACINTYRE, *supra* note 122.
[127] PELLEGRINO AND THOMASMA, *supra* note 120.
[128] Petra Gelhaus, *The desired moral attitude of the physician: (I) empathy*, 15 MEDICINE, HEALTH CARE AND PHILOSOPHY 103–113, 104 (2012).
[129] PELLEGRINO AND THOMASMA, *supra* note 120; PAUL STARR, THE SOCIAL TRANSFORMATION OF AMERICAN MEDICINE (REVISED EDITION): THE RISE OF A SOVEREIGN PROFESSION AND THE MAKING OF A VAST INDUSTRY (2nd Revised ed. edition ed. 2017); General Medical Council, *Consent Guidance* (2008), http://www.gmc-uk.org/guidance/ethical_guidance/consent_guidance_index.asp; GENERAL MEDICAL COUNCIL, *supra* note 122.
[130] PELLEGRINO AND THOMASMA, *supra* note 120 at 52.
[131] *Id.* at 52–3.

values and lived experience with disease,[132] the doctor as an ideal-type 'role' requiring certain technical expertise and professional training is beyond question—the point of contention is rather whether this expertise should be deferred to without challenge.

Fiduciary duties and the healing relationship

Human rights principles regarding health and supportive rights enacted through policies such as the Charter of Fundamental Rights of the European Union reflect the moral and fiduciary duties of medicine as a profession. As discussed above, these obligations can be traced to the core aims or ends of medicine as a practice, and can be traced to many possible theoretical foundations, including human rights, care ethics and feminist ethics, and virtue ethics.

The remainder of this section focuses on an account of the healing relationship and medicine's fiduciary duties developed in the context of virtue ethics. A virtue-based approach emphasises the importance of treating the patient as a whole and promoting the patient's well-being through good practice. Standards are defined against goods such as compassion that "safeguards that the patient is not only seen as a number,"[133] contextual understanding of the patient's values, history and concerns, an "interest in the inner processes of the patient…an adequate skill in responding non-verbally and by skilful and sensitive dialogue,"[134] alongside technical skill in 'fixing' the patient's disorder or managing a persistent condition. With that said, these core aims are shared by many other approaches outside of virtue ethics. For example, approaches to care ethics and feminist ethics focus on related goods such as the caring role of the health professional, relationships and care responsibilities (in contrast to a focus on justice and rights),[135] tacit knowledge and context-sensitive care that responds to the interests and needs of patients as unique, socially embedded individuals, and power imbalances and coercion owing to the vulnerable position of the patient.

Several characteristics of the healing relationship create moral obligations on practitioners to protect the interests of patients.[136] Specifically, the relationship can be characterised by the following traits:

- **Vulnerability and Inequality** – Patients experience a loss of control to define and pursue personal goals, and may experience emotional stress, fear, worry, and anxiousness.[137] The immediate goal of life becomes the restoration of health and well-being by relieving or curing symptoms. An imbalanced

[132] Emanuel and Emanuel, *supra* note 123; Edgar, *supra* note 118.
[133] Petra Gelhaus, *The desired moral attitude of the physician: (II) compassion*, 15 MEDICINE, HEALTH CARE AND PHILOSOPHY 397–410, 405 (2012).
[134] Gelhaus, *supra* note 127 at 108.
[135] CAROL GILLIGAN, IN A DIFFERENT VOICE: PSYCHOLOGICAL THEORY AND WOMEN'S DEVELOPMENT (1993).
[136] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4; Schotsmans, Dierckx de Casterle, and Gastmans, *supra* note 120.
[137] PELLEGRINO AND THOMASMA, *supra* note 120; David B. Morris, *About suffering: Voice, genre, and moral community*, 125 DAEDALUS 25–45 (1996); Keith Bauer, *Cybermedicine and the moral integrity of the physician–patient relationship*, 6 ETHICS AND INFORMATION TECHNOLOGY 83–91 (2004); Deborah Lupton, *The digitally engaged patient: self-monitoring and self-care in the digital health era*, 11 SOCIAL THEORY & HEALTH 256–270, 263 (2013).

relationship is created in which the patient is compelled to seek the help of an individual with privileged medical expertise in the pursuit of a return to health. Doctors have an obligation to not use their expertise or privileged position of power to exploit the "vulnerable" patient.[138]

- **Fiduciary Nature** – The patient explicitly or tacitly places trust in a chosen doctor and reveal aspects of himself and his life to allow diagnosis and healing, surrendering some privacy in allowing "others access to personal information or [their] bodies."[139] Doctors have a moral obligation to make use of the information and access provided by the patient in a trusting relationship in the patient's best interests, and not for self-interest.[140]

- **Nature of Medical Decisions** – Medical decisions are a combination of technical and moral features. The doctor's diagnosis and treatment of the patient must be technically accurate to promote physical health.[141] However, decisions should also support the patient's moral well-being or autonomy as an entity with moral value, in the sense that the decision should match with the patient's values.[142]

- **Characteristics of Medical Knowledge** – Medical knowledge is non-proprietary. To ensure a sufficient quantity of health professionals, societies provide doctors with privileged knowledge and access to human bodies necessary to gain medical expertise and may limit recognition of practitioners of medicine to individuals thus trained. Doctors have a moral obligation to act as stewards to this knowledge, ensuring it is readily available to others, used ethically in the treatment of patients, and not purely for self-interest.[143]

- **Moral Complicity** – The doctor is the channel through which medical interventions flow to the patient, in the sense that the doctor must agree to each intervention carried out. In this position the doctor has a moral obligation to act as a gatekeeper, safeguarding the patient's well-being and acknowledging his complicity in any interventions carried out.[144]

These characteristics are not beyond question; for instance, the experience of illness as vulnerability and inequality can be criticised in that it only seems to apply to acute problems with potential cures.[145] Although the 'healing relationship' approach

---

[138] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6; GILLIGAN, *supra* note 134.
[139] BEAUCHAMP AND CHILDRESS, *supra* note 123 at 298.
[140] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4; Bauer, *supra* note 136; John Heritage et al., *Problems and Prospects in the Study of Physician-Patient Interaction: 30 Years of Research*, 32 ANNUAL REVIEW OF SOCIOLOGY 351–374, 355 (2006); O. Karnieli-Miller & Z. Eisikovits, *Physician as partner or salesman? Shared decision-making in real-time encounters*, 69 SOCIAL SCIENCE & MEDICINE 1–8, 2 (2009).
[141] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4.
[142] BEAUCHAMP AND CHILDRESS, *supra* note 123; Karnieli-Miller and Eisikovits, *supra* note 139.
[143] PELLEGRINO AND THOMASMA, *supra* note 120 at 35–6, 42–4.
[144] *Id.* at 35–6, 42–4.
[145] MARTHA C. NUSSBAUM, FRONTIERS OF JUSTICE DISABILITY, NATIONALITY, SPECIES MEMBERSHIP (OIP): DISABILITY, NATIONALITY, SPECIES MEMBERSHIP (TANNER LECTURES ON HUMAN VALUES) (New Ed ed.

describes an idealistic model of the doctor-patient relationship (and thus, medicine itself), the underlying notion that being a doctor includes moral obligations to the patient is widely accepted.[146] The fundamental character of the medical relationship as one in which a patient in need seeks medical knowledge, expertise, or treatment is beyond question. In seeking out professional help, the patient is tacitly agreeing to reveal herself and private aspects of her life to the doctor with medical expertise in the pursuit of health. The relationship is an exchange of sensitive goods for improvements in quality of life which the patient is coerced through illness to engage in if a return to health is desired. Doctors are consulted not merely as 'encyclopaedias of knowledge', but rather as 'trusted' experts capable of subjective evaluation and understanding the patient as a socially embodied person with a history and values.[147]

Being a medical professional, or belonging to medicine understood as a formal profession, requires committing oneself to the moral obligations of the healing relationship.[148] Medicine can be considered a 'moral practice' in this context because its members form a community which shares a common goals and moral obligations,[149] meaning they are "guided by some shared source of morality—some fundamental rules, principles, or character traits that will define a moral life consistent with the ends, goals, and purposes of medicine".[150] Critically, this account contrasts the norms and obligations of individual practitioners with those of the institutions through which care is provided. Whereas the individual health professional's first obligation is to the patient, institutions have other (legitimate) interests concerning resourcing and quality of care across the institution as a whole. From a virtue ethics perspective, medical virtues and internal norms of good practice can help ensure the ends of medicine, and ultimately the obligations to individual patients incurred through the healing relationship, are met over time and resist erosion due to the corrupting influence of institutions and external goods.[151] For a discussion of specific virtues of good medical practice, see the Appendix.

Emergent challenges in the doctor-patient relationship

It could be argued that the healing relationship model is outdated, as "the notion of patients placing themselves under the care of a doctor and seeking their expert advice has moved to the concept of patients as producing health knowledges and as acquiring expert knowledge so as to manage their illness themselves."[152] This

2007); Barbara Page-Hanify, *Intellectual Handicap - Achievement of Potential*, 27 AUSTRALIAN OCCUPATIONAL THERAPY JOURNAL 53–60 (1980).

[146] BEAUCHAMP AND CHILDRESS, *supra* note 123; Andrew Edgar & Stephen Pattison, *Integrity and the moral complexity of professional practice*, 12 NURSING PHILOSOPHY 94–106 (2011); Gelhaus, *supra* note 127; Y. M. Barilan & M. Brusa, *Deliberation at the hub of medical education: beyond virtue ethics and codes of practice*, 16 MEDICINE, HEALTH CARE AND PHILOSOPHY 3–12 (2013).

[147] Emanuel and Emanuel, *supra* note 123 at 2225; Gelhaus, *supra* note 127 at 110.

[148] STARR, *supra* note 128.

[149] PELLEGRINO AND THOMASMA, *supra* note 120 at 3; Morris, *supra* note 136; Schotsmans, Dierckx de Casterle, and Gastmans, *supra* note 120.

[150] PELLEGRINO AND THOMASMA, *supra* note 120 at 3.

[151] *Id.* at 32.; MACINTYRE, *supra* note 122.

[152] Deborah Lupton, *M-health and health promotion: The digital cyborg and surveillance society*, 10 SOCIAL THEORY & HEALTH 229–244, 233 (2012).

characterisation of medicine suggests that the doctor-patient relationship has evolved and can seamlessly incorporate AI without altering the character of medical care.

As the practice of medicine changes in the face of emerging technologies, "something of the past is inevitably lost, not always for the worse."[153] Medicine has long been affected by advances in technology that disrupt the traditional one-to-one, face-to-face model of clinical care between doctor and patient. The Internet, for example, has empowered patients with greater access to medical information, but introduced risks owing to misleading or inaccurate information. Introducing new stakeholders into care relationships is not self-evidently problematic, but must be measured in terms of impact on the healing relationship and the ends of medicine; in other words, in the impact on patient care.

The healing relationship must be understood as an idealistic framework of the relationship between 'expert' doctors and 'vulnerable' patients. As an ideal, the model is not reflective of the 'empowered patient' model of care that has emerged in parallel over the past several decades.[154] Assuming modern medicine is characterised by 'empowered' patients eroding the privileged position of doctors as 'experts', trust cannot be assumed to exist whenever healing occurs.

However, the healing relationship describes the motivations of patients to seek professional care, or knowledge and technologies for self-care. Whether addressed through professional or self-directed care, the vulnerability of the patient is not eliminated. Similarly, the fiduciary duties created by this vulnerability do not change when diffused to different sources of expertise, be they medical professionals, databases of medical knowledge and advice, or other technologies and systems supporting self-care such as telemedicine or readily available medical information on the Internet.

Finding new ways to live up to the fiduciary duties of medicine in practice takes on renewed importance in this context and in the future deployment of AI in medicine. Pertinent questions have been asked, for example, about the validity and efficacy of medical knowledge available through internet portals. Furthermore, although medical information is increasingly available through other mediums, the role of expertise as an indication of fidelity to trust does not change.[155] Providers of low-quality medical advice, information or care can be criticised, regardless of format.

On this basis, the healing relationship model can be understood as a description of the moral character and obligations of medical practice, traditionally embodied by health practitioners but increasingly diffused across various platforms and persons, including web portals, consumer device developers, providers of wellness services, and others. Even if modern medicine has moved beyond the single doctor-patient model described in the healing relationship, the obligations of this relationship have not disappeared. Rather, the diffusion and displacement of these obligations by new technological actors in medicine is a cause for concern in considering how best to

[153] PELLEGRINO AND THOMASMA, *supra* note 120 at 32.
[154] Emanuel and Emanuel, *supra* note 123.
[155] *Id.*

govern the introduction of AI in medicine. Our notion of the healing relationship could, of course, be revised to give primacy to patient autonomy above all else. However, doing so risks reducing the doctor to a mere service-provider, incapable of exercising the full range of medical virtues and practice-internal norms.

When evaluating the impact of AI and algorithmic technologies on the doctor-patient relationship, choice of metric is key. If measured solely in terms of cost-benefits, or utility, the justification for AI mediation and augmentation of care is straightforward. However, while algorithmic technologies may allow for a greater number of patients to be treated more efficiently or at lower cost, their usage can simultaneously undermine non-mechanical dimensions of care. A distinction can be drawn between those effects of algorithmic systems (and components of utility) which contribute to the good of the patient or medicine as a practice governed by well-established internal norms and codes of conduct, and those which contribute to the good of medical institutions and healthcare services.

The moral complicity that characterises the doctor-patient relationship, wherein treatment is ideally guided by the professional's contextually and historically aware assessment of a patient's condition, cannot be easily replicated in interactions with AI systems. The role of the patient, the factors that lead people to seek medical attention, and the patient's vulnerability are not changed by the introduction of AI as a mediator or augmenter of medical care. Rather, what changes is the means of care delivery, how it can be provided, and by whom. The shift of expertise and care responsibilities to AI systems can be disruptive in many ways, which are explored in the section entitled "Potential impact of AI on the doctor-patient relationship".

6 POTENTIAL IMPACT OF AI ON THE DOCTOR-PATIENT RELATIONSHIP

AI promises a variety of opportunities, benefits, and risks for the practice of medicine. Drawing on the framework of ethical challenges facing AI and policy context developed in the sections entitled “Background and context”, “Overview of AI applications in medicine”, and “Theoretical framework of the doctor-patient relationship”, this section identifies six potential impacts of AI on the doctor-patient relationship.

Inequality in access to high quality healthcare

As an emerging technology the deployment of AI systems will not be immediate or universal across all member states or healthcare systems. Deployment across institutions and regions will inevitably be inconsistent in terms of scale, speed, and prioritisation. Telemedicine systems, for instance, are well suited to providing access to care in remote or inaccessible places, or where shortages exist in healthcare workers or specialists.[156] This promises to fill gaps in healthcare coverage but not necessarily with care of equivalent quality to traditional face-to-face care. Impact on the doctor-patient relationship in the near term may therefore be much greater in areas suffering from existing staffing shortages or new shortages owing to the COVID-19 pandemic. The quality and degree of this impact remains to be seen.

The unavoidable variability in deployment of AI raises the possibility that geographical bias in performance and inequalities in access to high quality care will be created through the usage of AI systems. This cuts both ways. If AI systems raise the quality of care, for example by providing more accurate or efficient diagnosis, expanded access to care, or through the development of new pharmaceutical and therapeutic interventions, then patients served by ‘early adopter’ regions or health institutions will benefit before others. AI systems may also be used to free up clinicians from menial, labour intensive tasks such as data entry and thus provide more time with patients than was previously possible.[157]

However, these benefits are not foregone conclusions. The impact of AI on clinical care and the doctor-patient relationship remains uncertain and will certainly vary by application and use case. AI systems may prove to be more efficient than human care, but also provide lower quality care featuring fewer face-to-face interactions. In many areas AI is seen as a promising means to cut costs, reduce waiting times, or fill existing gaps in coverage where access to health professionals and institutions is limited.[158] Patients in early adopter areas will at a minimum receive a different type of care which

[156] World Health Organization, *supra* note 1.
[157] *Id.* at 8.
[158] Department of Health, *Innovation Health and Wealth: Accelerating Adoption and Diffusion in the NHS* (2011); DEPARTMENT OF HEALTH, EQUITY AND EXCELLENCE: LIBERATING THE NHS. (2010).

may not be of the same quality as traditional care provided by human health professionals.

The inconsistent rollout of AI systems with uncertain impacts on access and care quality poses a risk of creating new health inequalities in member states. It may prove to be the case that regions that have historically faced unequal access or lower quality care are seen as key test beds for AI-mediated care. Patients in these areas may have better access to AI systems, such as chatbots or telemedicine, but continue to face limited access to human care or face-to-face clinical encounters. The likelihood of this risk depends largely on the strategic role given to AI systems. If they are treated as a potential replacement for face-to-face care, rather than as a means to free up clinicians' time greater inequality in access to human care seems inevitable.

Article 4 of the Oviedo Convention addresses care provided by healthcare professionals bound by professional standards. It remains unclear whether developers, manufacturers, and service providers for AI systems will be bound by the same professional standards. The Convention's Explanatory Report raises this question indirectly, noting that "from the term 'professional standards' it follows that [Article 4] does not concern persons other than health care professionals called upon to perform medical acts, for example in an emergency." Can a chatbot designed for initial triage of patients be considered a "person" performing a "medical act"?[159] If not, how can the involvement of an appropriately bounded healthcare professional be guaranteed?

Any reduction in oversight or clinical care by health professions caused by the rollout of AI systems could thus potentially be viewed as a violation of Article 4. In particular, care models that incorporate chat bots or other artificial agents designed to provide care or support directly to patients would seem to pose this risk. Careful consideration must be given to the role played by healthcare professions bound by professional standards when incorporating AI systems that interact directly with patients.

Transparency to health professionals and patients

AI challenges our notions of accountability in both familiar and new ways. Systems increasingly trusted to help make life-changing decisions and recommendations have their foundation in our technological past, but they are digital, distributed, and often imperceptible. When important decisions are taken which affect the livelihood and well-being of people, one expects that their rationale or reasons can be understood.

This expectation is reflected in Article 5 of the Oviedo Convention which reaffirms the right to informed consent for patients prior to being subject to medical interventions or research. As detailed above, the Convention's Explanatory Report specifies a non-comprehensive list of information to be provided. An overarching requirement is that the information must be provided to patients in an easily understandable way to ensure it can meaningfully inform their decisions. Traditionally, this would impose requirements on how health professionals explain their decisions and

[159] Korngiebel and Mooney, *supra* note 95 at 3.

recommendations to patients. In cases where AI systems provide some form of clinical expertise, for example by recommending a particular diagnosis or interpreting scans, this requirement to explain one's decision-making would seemingly be transferred from doctor to AI system, or at least to manufacturer of AI system.

The difficulty of explaining how AI systems turn inputs into outputs poses a fundamental epistemological challenge for informed consent. Aside from the patient's capacity to understand the functionality of AI systems, in many cases patients simply do not have sufficient levels awareness to make free and informed consent possible. AI systems use unprecedented volumes of data to make their decisions, and interpret these data using complex statistical techniques, both of which add to increase the difficulty and effort required to remain aware of the full scope of data processing informing one's diagnosis and treatment.[160]

In practice, transparency requirements in the service of informed consent can be borne out in several ways. Assuming doctors remain as the primary point of care for patients, the doctor can be seen as a mediator between the patient and the AI system. In this mediation model, the doctor can be the recipient of an explanation from the AI system and then act as a 'translator' for the patient, translating the system's explanation into a meaningful and easily understandable format. Where doctors do not act as mediators, for example where chatbots provide diagnosis or triage directly to patients, AI systems may then be expected to explain their decision-making directly to patients.

Both models pose challenges in explaining complex 'black box' behaviours to expert or non-expert users. At a minimum, AI systems interacting directly with patients should self-identify as an artificial system. Whether any usage of AI systems in care should be disclosed to patients by clinicians and healthcare institutions is a more difficult question.[161]

A commonly cited concern with AI used for operational purposes by hospitals, including risk stratification and discharge, planning tools is a failure to inform patients about the usage of AI in their care.[162]

On the one hand, health professionals routinely consult many sources of information in diagnosing and treating patients, such as models, charts, X-rays, etc., that they would not disclose or proactively discuss as part of informed consent. On the other hand, AI systems which effectively provide artificial clinical expertise, for instance by interpreting scans and recommending a classification of abnormalities, may be a qualitatively different type of information than sources that traditionally factor into clinical decision-making.

Nonetheless, in practice AI systems used to support clinical care and stratify risk among patients are often treated as purely operational rather than clinical applications. According to many health institutions they are used to improve the quality and

[160] COUNCIL OF EUROPE, *supra* note 2.
[161] I. Glenn Cohen, *Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? Symposium: Law and the Nation's Health*, 108 GEO. L.J. 1425–1470 (2019); Robbins and Brodwin, *supra* note 5.
[162] Cohen, *supra* note 160; Robbins and Brodwin, *supra* note 5.

efficiency of care, not to inform clinical decision-making. In this regard, they can be considered equivalent to other administrative systems used in hospitals that handle patient data but not for their immediate care.[163] Of course, not all health institutions treat AI risk prediction systems as purely operational; in some cases, patients are asked to explicitly consent to the usage of an AI system designed to identify patients at risk of death in the next 48 hours.[164] Recommendations concerning disclosure of the usage of AI systems will be returned to in the Section entitled "Public register of medical AI systems for transparency".

Independent of the question of whether particular AI applications should be classified as clinical or operational/administrative, there are pertinent questions concerning the intelligibility of 'black box' systems at a more fundamental level. Compared to human and organisational decision-making, AI poses a unique challenge. The internal state of a trained machine learning model can consist of millions of features connected in a complex web of dependent behaviours. Conveying this internal state and dependencies in a human comprehensible way is extremely challenging.[165] How AI systems make decisions may thus be too complex for human beings to thoroughly understand their full decision-making criteria or rationale.

Assuming the transparency requirement underlying informed consent is a key value in the AI-mediated doctor-patient relationship, the challenge of opacity raises a question: how should AI systems explain themselves to doctors and patients? We can begin to unpack this question by examining the different types of questions, notably we may ask about AI systems to make them understandable:

- **How does an AI system or model function? How was a specific output produced by an AI system?** These are questions of interpretability. Questions of interpretability address the internal functionality and external behaviour of an AI system. A fully interpretable model is one which is human comprehensible, meaning a human can understand the full set of causes of a given output.[166] Poorly interpretable models 'are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs'.[167] Interpretability can also be defined in terms of the predictability of the model; a model is interpretable if a well-informed person could consistently predict its outputs and behaviours.[168] Questions of model behaviour

[163] Robbins and Brodwin, *supra* note 5.
[164] *Id.*
[165] Jenna Burrell, *How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOCIETY (2016); Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV:1606.03490 [CS, STAT] (2016), http://arxiv.org/abs/1606.03490 (last visited Oct 15, 2016).
[166] Paulo JG Lisboa, *Interpretability in Machine Learning–Principles and Practice*, *in* FUZZY LOGIC AND APPLICATIONS 15–21 (2013), http://link.springer.com/chapter/10.1007/978-3-319-03200-9_2 (last visited Dec 19, 2015); Tim Miller, *Explanation in artificial intelligence: Insights from the social sciences*, 267 ARTIFICIAL INTELLIGENCE 1–38 (2019).
[167] Burrell, *supra* note 164 at 1.
[168] Been Kim, Rajiv Khanna & Oluwasanmi O. Koyejo, *Examples are not enough, learn to criticize! criticism for interpretability*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 2280–2288 (2016).

narrowly address how a particular output or behaviour of the model occurred.[169] However, model behaviour can also be broadly interpreted to include effects on reliant institutions and users and their AI-influenced decisions, for example how a doctor's diagnosis was influenced by an expert system's recommendation, are also relevant.[170]

- **How was an AI system designed and tested? How is it governed?** These are questions of transparency. Unlike interpretability, transparency does not address the functionality or behaviour of the AI system itself, but rather the processes involved in its design, development, testing, deployment, and regulation. Transparency principally requires information about the institutions and people that create and use AI systems, as well as the regulatory and governance structures that control both the institutions and systems. Here, interpretability play a supplementary but supportive role. Interpretable models or explanations of specific decisions taken by a system may, for example, be needed for regulators to effectively audit AI and ensure regulatory requirements are being met in each context of use.
- **What information is required to investigate the behaviour of AI systems?** This is a question of traceability. To audit the behaviour of AI systems, certain evidence is needed, which can include 'data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used'.[171] This data needs to be consistently recorded as the system operates for effective governance to be feasible. Traceability is thus a fundamental requirement for post hoc auditing and explanations of model behaviour; without the right data, explanations cannot be computed after a model has produced a decision or other output.[172]

Answers to each of these questions may be necessary to achieve informed consent in AI-mediated care. This is not to say both patients and health professions require answers to each question; rather, it may be the case that certain questions are better directed towards one or the other. For example, patients may be most immediately interested in questions concerning how their specific case was decided, or a diagnosis or recommendation reached.[173] Questions concerning how AI systems have been designed and tested, and how they are secured and validated over time, may be more immediately relevant to health professionals and administrators who must assess a system's trustworthiness in terms of integrating it into existing clinical and operational

[169] The degree to which the reasons for specific model behaviours can be explained is sometimes referred to as the *explainability* of a model. Here it is treated as one component of *interpretability* alongside intrinsic model comprehensibility.
[170] HIGH LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, *Ethics Guidelines for Trustworthy AI* (2019).
[171] *Id.*
[172] Mittelstadt et al., *supra* note 17.
[173] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 3 HARVARD JOURNAL OF LAW & TECHNOLOGY 841–887 (2018).

decision-making pathways.[174] As suggested in the section entitled "Theoretical framework of the doctor-patient relationship", the informed consent ideal is one component of the doctor-patient relationship requiring discussion between patients and health professionals of possible treatment options, values, and the like. Directing explanation types to the parties best equipped to understand them, or most immediately interested in them, need not undermine ideals of transparency or informed consent, but rather can be seen as a facilitator of meaningful dialogue between patient and doctor about options in AI-mediated care.

Risk of social bias in AI systems

As discussed in the section entitled "Common ethical challenges in AI", AI systems are inevitably biased in some respect. Many biases arise due to technical reasons, such as a mismatch between training and testing environments.[175] System developers and manufacturers inevitably design systems that reflect their values or relevant regulatory requirements; this can also be treated as a type of bias which will vary between manufacturers and member states.[176] However, in AI systems biased and unfair decision-making often occurs not for technical or regulatory reasons, but rather reflect underlying social biases and inequalities.[177]

These types of social biases are concerning for several reasons.

- First, they may undermine the accuracy of models across different populations or demographic groups. Many biases can be traced to datasets that are not representative of the population targeted by a system. In medicine, there are crucial data gaps that can be filled but to date are not due to limitations on resources, access, or motivation. Clinical trials and health studies are predominantly undertaken on white male subjects meaning results are less likely to apply to women and people of colour.[178] A serious and dangerous data gap exists because many clinical models treat women as "little men"[179] and thus do not account for biological differences.[180] For example, different percentage of body fat, thinner skin, different hormone levels and compositions, changing hormone levels throughout the menstrual cycle, changing hormone levels

[174] COUNCIL OF EUROPE, *supra* note 2.
[175] Friedman and Nissenbaum, *supra* note 41; Wachter, Mittelstadt, and Russell, *supra* note 47.
[176] COUNCIL OF EUROPE, *supra* note 2.
[177] *Id.*; Wachter, Mittelstadt, and Russell, *supra* note 47.
[178] CAROLINE CRIADO PEREZ, INVISIBLE WOMEN: EXPOSING DATA BIAS IN A WORLD DESIGNED FOR MEN 115–116 (2019); on how to adress bias in the medical setting see Timo Minssen et al., *Regulatory responses to medical machine learning*, JOURNAL OF LAW AND THE BIOSCIENCES (2020); and Mirjam Pot, Wanda Spahl & Barbara Prainsack, *The Gender of Biomedical Data: Challenges for Personalised and Precision Medicine*, 9 SOMATECHNICS 170–187 (2019).
[179] ANGELA SAINI, INFERIOR: HOW SCIENCE GOT WOMEN WRONG AND THE NEW RESEARCH THAT'S REWRITING THE STORY 59 (2017).
[180] PEREZ, *supra* note 177 at 116 One of the reasons why this is not done is because it is more complex (e.g. fluctuating hormone levels during the menstrual cycle), risky (e.g. female participants could be pregnant), time and resource intensive to study women. SAINI, *supra* note 178 at 58.

before puberty and after menopause are factors that affect how well drugs work or how much we are affected by toxins or environmental impacts.[181]

- Second, social biases can lead to unequal distribution of outcomes across populations or protected demographic groups. Inequality of this type is particularly severe in the context of medicine which affects fundamental goods: "any bias in the functioning of an algorithm could lead to inadequate prescriptions of treatment and subject entire population groups to unwarranted risks that may threaten not only rights but also lives."[182] Large segments of Western societies currently face significant prejudice and inequality which are captured in historical decisions and can influence the training of future systems. Historical trends in decision-making have led to diminished and unequal access to opportunities and outcomes among certain groups.[183] Without intervention, these pre-existing patterns in access to opportunities and resources in society will be learned and reinforced by AI systems.

As discussed, Article 14 of the ECHR prohibits discrimination. Equality is a key value underlying human rights. However, achieving substantive equality or a 'level playing field' in practice is extremely difficult. With regards to AI, dataset bias and feedback loops are key challenges to ensure systems do not exacerbate existing inequalities and create new forms of discrimination that would run counter to Article 14. The Parliamentary Assembly of the Council of Europe has recognised the risk of bias in this respect, noting that "Council of Europe member states should participate more actively in the development of AI applications for health care services, or at least provide some sort of sovereign screening and authorisations for their deployment. States' involvement would also help to ensure that such applications are fed with sufficient, unbiased and well protected data."[184]

Concerning dataset bias, conceiving of bias solely as a property of datasets is insufficient to achieve substantive equality in practice.[185] Assuming it is possible to create a dataset that perfectly captures existing biases and inequalities in society, training a model with this dataset would do nothing to correct the inequalities captured by it. Rather, such assurances can only be provided by also examining, testing for, and perhaps correcting biases in the trained AI system and its outputs.

With regards to feedback loops, reinforcing existing biases in society that have been learned by an AI system can make matters substantively worse for already disadvantaged groups. However, simply avoiding reinforcement of existing biases and inequalities, or ensuring AI systems do not make the status quo worse, does not achieve substantive equality in practice.[186] Rather, this requires critically examining the acceptability of existing inequalities and taking steps to positively improve the situation of disadvantaged groups. Likewise, AI systems can create novel forms of

[181] SAINI, *supra* note 178 at 62; PEREZ, *supra* note 177 at 116.
[182] COUNCIL OF EUROPE, *supra* note 2.
[183] See for example ANGELA Y. DAVIS, WOMEN, RACE, & CLASS (2011).
[184] COUNCIL OF EUROPE, *supra* note 2.
[185] Wachter, Mittelstadt, and Russell, *supra* note 47.
[186] *Id.*

discrimination rather than simply reinforcing existing forms of bias and inequality.[187] Both the need for critical positive action and the possibility of novel forms of discrimination fuelled by AI need to be accounted for in deploying AI in medicine.

Detecting biases in AI systems is not straightforward. Biased decision-making rules can be hidden in 'black box' models. Other biases can be detected by examining the outputs of AI systems for unequal distributions across demographic groups or relevant populations. However, accessing the full range of decisions or outputs of a system is not necessarily straightforward, at a minimum due to data protection standards; "certain restrictions on the use of personal health data may disable essential data linkages and induce distortions, if not errors, in AI-driven analysis."[188] At a minimum, this suggests that simply anonymising health data may not be an adequate solution to mitigate biases or correct their downstream effects. Even where decision sets are accessible, demographic data may not exist for the relevant populations meaning bias testing cannot measure distribution across relevant legally protected groups.[189]

These various challenges of social bias, discrimination, and inequality suggest health professionals and institutions face a difficult task in ensuring their usage of AI systems does not further existing inequalities and create new forms of discrimination. Combatting social bias is a multifaceted challenge which must include robust bias detection and testing standards, high quality collection and curation standards for training and testing datasets, and individual-level testing to ensure patient outcomes and recommendations are not predominantly determined by legally protected characteristics.[190] Failing to implement robust bias testing standards risks further exacerbating inequalities in AI-driven care and undermining the trustworthiness of AI-mediated care. These risks are particularly acute in the context of existing inequalities in access to high-quality care where the deployment of AI may be accelerated for the sake of efficiency and resource allocation rather than purely clinical considerations.

Dilution of the patient's account of well-being

Traditionally, clinical care and the doctor-patient relationship are ideally informed by the doctor's contextual, historically aware assessment of a patient's condition. This type of care cannot be easily replicated in technologically-mediated care. Data representations of the patient necessarily restrict the doctor's understanding of the patient's case to measured features. This can present a problem when clinical assessments increasingly rely on data representations, constructed for example by remote monitoring technologies, or other data not collected in face-to-face encounters. Data representations of patients can come to be seen as an 'objective' measure of health and well-being, reducing the importance of contextual factors of health or the view of the patient as a socially embodied person. Data representations can create a

[187] Wachter, Mittelstadt, and Russell, *supra* note 47.
[188] COUNCIL OF EUROPE, *supra* note 2.
[189] Wachter, Mittelstadt, and Russell, *supra* note 47; Sandvig et al., *supra* note 79; Brent Mittelstadt, *Automation, Algorithms, and Politics*| *Auditing for Transparency in Content Personalization Systems*, 10 INTERNATIONAL JOURNAL OF COMMUNICATION 12 (2016).
[190] Wachter, Mittelstadt, and Russell, *supra* note 47; Wachter, Mittelstadt, and Russell, *supra* note 47; Matt J. Kusner et al., *Counterfactual Fairness* (2017).

'veneer of certainty', in which 'objective' monitoring data is taken to represent a true representation of the patient's situation, losing sight of the patient's interpersonal context and other tacit knowledge.[191]

Medical professionals face this difficulty when attempting to incorporate AI systems into care routines. The amount and complexity of data and technologically derived recommendations about a patient's condition makes it difficult to identify when important contextual information is missing. Reliance upon data collected by 'health apps' or monitoring technologies (e.g., smart watches) as a primary source of information about a patient's health, for example, can result in ignorance of aspects of the patient's health that cannot easily be monitored. This includes essential elements of mental health and well-being such as the patient's social, mental, and emotional states. 'Decontextualisation' of the patient's condition can occur as a result, wherein the patient loses some control over how her condition is presented and understood by clinicians and carers.[192]

All of these possibilities suggest the encounters through which the basic trust necessary for a doctor-patient relationship is traditionally developed may be inhibited by technological mediation. Technologies which inhibit communication of "psychological signals and emotions" can impede the doctor's knowledge of the patient's condition, undermining "the establishment of a trusting and healing doctor-patient relationship."[193] Care providers may be less able to demonstrate understanding, compassion, and other desirable traits found within 'good' medical interactions in addition to applying their knowledge of medicine to the patient's case. As a mediator placed between the doctor and patient, AI systems change the dependencies between clinicians and patients by turning some degree of the patient's ongoing care over to a technological system. This can increase the distance between health professionals and patients thereby suggesting a loss of opportunities to develop tacit understanding of the patient's health and well-being.[194]

Risk of automation bias, de-skilling, and displaced liability

As discussed in the section entitled "Common ethical challenges in AI", the introduction of AI systems into clinical care poses a risk of automation bias, according to which clinicians may trust the outputs or recommendations of AI systems not due to proven clinical efficacy, but rather on the basis of their perceived objectivity, accuracy, or complexity.[195] Any deployment of AI systems designed to augment human decision-making with recommendations, warnings, or similar interventions runs the risk of introducing automation bias. Empirical work on the phenomenon is somewhat nascent, but one recent study showed how even expert decision-makers can be prone to automation bias over time for problematic reasons (e.g., the cost of an AI system

[191] Mark Coeckelbergh, *E-care as craftsmanship: virtuous work, skilled engagement, and information technology in health care*, 16 MEDICINE, HEALTH CARE AND PHILOSOPHY 807–816 (2013).
[192] Mittelstadt et al., *supra* note 3.
[193] Bauer, *supra* note 136 at 84.
[194] Coeckelbergh, *supra* note 190.
[195] Zarsky, *supra* note 31 at 121.

as a proxy for accuracy or equality).[196] The Council of Europe has clearly recognised the risk of automation bias in calling for guarantees that "AI-driven health applications do not replace human judgement completely and that thus enabled decisions in professional health care are always validated by adequately trained health professionals."[197]

Reliance on AI systems as clinical care providers or expert diagnostic systems can inhibit the development of skills, professional communities, norms of 'good practice' within medicine. This phenomenon is referred to as 'de-skilling',[198] and runs counter to what the WHO has referred to as 'human-centred AI' which supports and augments human expertise and skill development, rather than undermining or replacing them.[199] Medical professionals develop virtues or norms of good practice through their experiences of practicing medicine. To define norms, practitioners can draw on practical wisdom developed through their experience. Members of the medical profession form a community which shares common goals and moral obligations.[200] The virtues or internal norms of a practice help ensure its ends are met over time by combating the influence of institutions and external goods. The development, maintenance, and application of these norms can be displaced through technological mediation of care.

It follows that the development, maintenance, and application of internal norms necessary to meet moral obligations to patients can be undermined when care is technologically mediated, and thus provided in part by non-professional individuals and institutions. A potential exists for algorithmic systems to displace responsibilities traditionally fulfilled by medical professionals, while providing more efficient or 'better' care measured solely in terms of cost-benefit. To prevent the erosion of holistically good, not merely technically 'efficient', medical care, these moral obligations to benefit and respect patients in the first instance need to be taken seriously by new care and services providers that are not part of traditional medical communities. In other words, a gap in professional skills and accountability can be created by AI-mediated care.

De-skilling and automation bias also pose risks directly to patients. One function of human clinical expertise is to protect the interests and safety of patients. Risks to safety come from a variety of sources, including "malicious attacks on software, unethical system design or unintended system failure, loss of human control and the "exercise of digital power without responsibility" that can lead to tangible harm to human health, property and the environment."[201]

If this human expertise is eroded through de-skilling or displaced through automation bias, testing and evidence of clinical efficacy must fill the gap to ensure patient safety. A similar trade-off exists in relation to opacity and accuracy; some scholars have

---

[196] Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853 (2019).
[197] COUNCIL OF EUROPE, *supra* note 2.
[198] *Id.*; Coeckelbergh, *supra* note 190.
[199] World Health Organization, *supra* note 1.
[200] MACINTYRE, *supra* note 122.
[201] COUNCIL OF EUROPE, *supra* note 2; COUNCIL OF EUROPE, *Responsibility and AI* (2019), https://rm.coe.int/responsability-and-ai-en/168097d9c5.

argued that medical AI systems do not necessarily need to be explainable if their accuracy and clinical efficacy can be reliably validated.[202] In both cases the protection of vital patient interests, or the fiduciary obligations typically shouldered by health professionals, are transferred to providers of AI systems or the systems themselves.

As a result, to continue to ensure patient safety and replace the protection offered by human clinical expertise, robust testing and validation standards should be an essential pre-deployment requirement for AI systems in clinical care contexts. These standards should also address complementary non-clinical aspects of safety such as cybersecurity, malfunctioning and resilience.[203] While a seemingly obvious conclusion, the existence of such requirements and evidence meeting them cannot be taken for granted. As discussed in the section entitled "Overview of AI technologies in medicine", evidence of clinical efficacy does not yet exist for many AI applications in healthcare, which has justifiably proven a barrier to widespread deployment.

A related but equally important topic concerns liability for malfunctioning and other harmful effects of AI. As discussed in the section entitled "Overview of AI technologies in medicine", distributed responsibility is both a morally and legally difficult challenge. The Parliamentary Assembly of the Council of Europe has recognised the need to clarify the liability of stakeholders in AI including "developers to regulatory authorities, intermediaries and users (including public authorities, health-care professionals, patients and the general public)." Member states of the Council of Europe are called on to "elaborate a legal framework for clarifying the liability of stakeholders for the design, deployment, maintenance and use of health-related AI applications (including implantable and wearable medical devices) in the national and pan-European context, redefine stakeholder responsibility for risks and harms from such applications and ensure that governance structures and law enforcement mechanisms are in place to guarantee the implementation of this legal framework."[204] A 2019 report from the Council of Europe Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) explored the specific challenges of liability and responsibility gaps in AI in much greater detail than is possible here.[205]

Impact on the right to privacy

AI poses several unique challenges to the human right to privacy and complementary data protection regulations. As discussed in the section entitled "The Oviedo Convention and human rights principles regarding health", the Council of Europe is currently in the processing of ratifying amendments to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108 and CETS No. 223). These additional rights seek to provide individuals with greater transparency and control over automated forms of data processing. These

[202] Boris Babic et al., *Beware explanations from AI in health care*, 373 SCIENCE 284–286 (2021).
[203] COUNCIL OF EUROPE, *supra* note 2.
[204] *Id.*
[205] COUNCIL OF EUROPE, *supra* note 200.

rights will undoubtedly provide valuable protection for patients across a variety of use cases of medical AI.

One distinct challenge unique to AI worth further consideration concerns the usage of patient data for training and testing AI systems. Confidentiality in the doctor-patient relationship is a key value to protect the human right to privacy. At the same time, greater development, deployment, and reliance on AI systems in care may create a greater need to create or curate high-quality real-world patient datasets to train and test systems. Innovation can threaten privacy and confidentiality in two ways. First, there may be a greater pressure to re-purpose and grant third party access to (deidentified) patient data and electronic health records to test and develop AI systems.

Second, clinicians may be encouraged to prescribe additional tests and analysis not for their clinical value but rather due to their utility for training or testing AI systems. This has implications both in terms of rising costs for healthcare but also exposure of patients to unnecessary risks of data leakage or other breaches of privacy. The Oviedo Convention sets out a specific application of the right to privacy (Article 8 ECHR) which recognises the particularly sensitive nature of personal health information and sets out a duty of confidentiality for health care professionals. Any generation of data with questionable clinical value or clearly motivated by its utility solely for the testing or development of AI systems would seemingly violate the Convention's specification of the right to privacy.

As this suggests, where a legitimate need exists for real-world data to test and train AI systems, interests in innovation and care efficiency or quality must be balanced with the patient's individual interests in privacy and confidentiality. Failing to strike this balance risks undermining trust between patients and care providers. Trust would be lost not owing to a failure to use AI appropriately in individual clinical encounters, but rather due to an institutional failure to protect patient interests in privacy and confidentiality at an institutional level. At a minimum, any re-purposing of patient health records for training and testing AI systems should be subject to sufficient deidentification and privacy enhancing techniques such as differential privacy (which introduces noise to prevent identification of a particular person in the dataset).[206]

[206] Cynthia Dwork, *Differential Privacy*, *in* AUTOMATA, LANGUAGES AND PROGRAMMING 1–12 (Michele Bugliesi et al. eds., 2006), http://link.springer.com/chapter/10.1007/11787006_1 (last visited Apr 4, 2016); Paul Ohm, *Broken promises of privacy: Responding to the surprising failure of anonymization*, 57 UCLA LAW REVIEW 1701 (2010).

7 RECOMMENDATIONS FOR COMMON ETHICAL STANDARDS FOR TRUSTWORTHY AI

The preceding discussion in the section entitled “Potential impact of AI on the doctor-patient relationship” concluded that ethical standards need to be developed around transparency, bias, confidentiality, and clinical efficacy to protect patient interests in informed consent, equality, privacy, and safety. Together, such standards could serve as the basis for deployments of AI in healthcare that help rather than hinder the trusting relationship between doctors and patients. These standards can address both how systems are designed and tested prior to deployment, as well as how they are implemented in clinical care routines and institutional decision-making processes.

The Oviedo Convention acts as a minimum standard for the protection of human rights which requires translation into domestic law. On this basis, there is an opportunity to make specific, positive recommendations concerning the standard of care to be met in AI-mediated healthcare. These recommendations must not interfere with the exercise of national sovereignty in standard setting through domestic law and professional bodies as detailed in Article 4 of the Oviedo Convention. However, it is also possible to set standards which do not interfere with Article 4 and can be considered directly enforceable. Specifically, as noted by Andorno:

> “The common standards set up by the Council of Europe will mainly operate through the intermediation of States. This does not exclude of course that some norms contained in the Convention may have self-executing effect in the internal law of the States having ratified it. This is the case, for instance, of some norms concerning individual rights such as the right to information, the requirement of informed consent, and the right not to be discriminated on grounds of genetic features. Prohibition norms can also be considered to have immediate efficacy, but in the absence of legal sanctions, whose determination corresponds to each State (Article 25), their efficacy is restricted to civil and administrative remedies.”

Where AI can be observed to have a clear impact on rights and protections set out in the Oviedo Convention, it is appropriate for the Council of Europe to introduce binding recommendations and requirements for signatories concerning how AI is deployed and governed. Recommendations should focus on a higher positive standard of care with regards to the doctor-patient relationship to ensure it is not unduly disrupted or by the introduction of AI in care settings. Of course, such standards should be supportive to a degree of local interpretation around key normative issues like acceptable degrees of automation bias, acceptable trade-offs between outcomes between patient groups, and similar areas influenced by local norms.

The following example recommendations detail possible essential requirements and recommendations for an intelligibility standard that aims to protect informed consent in AI-mediated care, a transparency standard for public intelligibility, and a standard for collection of sensitive data for purposes of bias testing. Each should be treated as an example of the type of recommendation that can be drawn from the preceding discussion of the potential ethical impacts of AI on the doctor-patient relationship.

Intelligibility requirements for informed consent

According to the Explanatory Report, Article 5 of the Oviedo Convention contains an incomplete list of information that should be shared as part of an informed consent process. As this list is incomplete, the Council of Europe could set standards for what and how information about the recommendation of an AI system concerning a patient's diagnosis and treatment should be communicated to the patient. Given the traditional role of the doctor in sharing and discussing this type of information in clinical encounters, these standards should likewise address the doctor's role in explaining AI recommendations to patients and how AI systems can be designed to support the doctor in this role.

Several concepts are common across the questions and goods that motivate interpretability in AI. Interpretability methods seek to explain the functionality or behaviour of the 'black box' machine learning models that are a key component of AI decision-making systems. Trained machine learning models are 'black boxes' when they are not comprehensible to human observers because their internals and rationale are unknown or inaccessible to the observer, or known but uninterpretable due to their complexity.[207] Interpretability in the narrow sense used here refers to the capacity to understand the functionality and meaning of a given phenomenon, in this case a trained machine learning model and its outputs, and to explain it in human understandable terms.[208]

'Explanation' is likewise a key concept in AI interpretability. Generically, explanations in AI relate 'the feature values of an instance to its model prediction in a humanly understandable way'.[209] This rough definition hides significant nuance. The term captures a multitude of ways of exchanging information about a phenomenon, in this case the functionality of a model or the rationale and criteria for a decision, to different stakeholders.[210]

To understand how 'explanation' can be operationalised in medicine, two key distinctions are relevant:

---

[207] Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUT. SURV. 93:1-93:42 (2018); INFORMATION COMMISSIONER'S OFFICE & THE ALAN TURING INSTITUTE, *Explaining decisions made with AI* (2020), https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/.

[208] Finale Doshi-Velez & Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, ARXIV:1702.08608 [CS, STAT] (2017), http://arxiv.org/abs/1702.08608 (last visited Sep 22, 2017).

[209] CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING 31 (2020), https://christophm.github.io/interpretable-ml-book/ (last visited Jan 31, 2019).

[210] Lipton, *supra* note 164; Miller, *supra* note 165.

- First, methods can be distinguished in terms of what it is they seek to explain. Explanations of model functionality address the general logic the model follows in producing outputs from input data. Explanations of model behaviour, in contrast, seek to explain how or why a particular behaviour exhibited by the model occurred, for example how or why a particular output was produced from a particular input. Explanations of model functionality aim to explain what is going on inside the model, whereas explanations of model behaviour aim to explain what led to a specific behaviour or output by referencing essential attributes or influencers on that behaviour. It is not strictly necessary to understand the full set of relationships, dependencies, and weights of features within the model to explain model behaviour.
- Second, interpretability methods can be distinguished in how they conceptualise 'explanation'. Many methods conceptualise explanations as approximation models, which are a type of simpler, human interpretable model that is created to reliably approximate the functionality of a more complex 'black box' model. The approximation model itself is often and confusingly referred to as an explanation of the 'black box' model. This approach contrasts with the treatment of 'explanation' in philosophy of science and epistemology in which the term typically refers to explanatory statements that explain the causes of a given phenomenon.[211]

The usage of 'explanation' in this fashion can be confusing. Approximation models are best thought of as tools from which explanatory statements about the original model can be derived.[212] Explanatory statements themselves can be textual, quantitative, or visual, and report on several aspects of the model and its behaviours.

Further distinctions help classify different types of explanations and interpretability methods. A basic distinction in interpretability can be drawn between global and local interpretability. This distinction refers to the scope of the model or outputs a given interpretability or explanatory method aims to make human comprehensible. Global methods aim to explain the functionality of a model as a whole or across a particular set of outputs in terms of the significance of features, their dependencies or interactions, and their effect on outputs. In contrast, local methods can address, for example, the influence of specific areas of the input space or specific variables on one or more specific outputs of the model.

Models can be globally interpretable at a holistic or modular level.[213] Holistic global interpretability refers to models which are comprehensible to a human observer in the sense that the observer can follow the entire logic or functional steps taken by the model which lead to all possible outcomes of the model.[214] It should be possible for a

[211] Brent Mittelstadt, Chris Russell & Sandra Wachter, *Explaining Explanations in AI*, PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY - FAT* '19 279–288 (2019).
[212] *Id.*
[213] MOLNAR, *supra* note 208.
[214] Guidotti et al., *supra* note 206.

single person to comprehend holistically interpretable models in their entirety.[215] An observer would have 'a holistic view of its features and each of the learned components such as weights, other parameters, and structures'.[216]

Given the limitations of human comprehension and short-term memory, global holistic interpretability is currently only practically achievable on relatively simple models with few features, interactions, or rules, or strong linearity and monotonicity.[217] For more complex models, global interpretability at a modular level may be feasible. This type of interpretability involves understanding a particular characteristic or segment of the model, for example the weights in a linear model, or the splits and leaf node predictions in a decision tree.[218]

With regards to local interpretability, a single output can be considered interpretable if the steps that led to it can be explained. Local interpretability does not strictly require that the entire series of steps be explained; rather, it can be sufficient to explain one or more aspects of the model that led to the output, such as a critically influential feature value.[219] A group of outputs is considered locally interpretable if the same methods to produce explanations of individual outputs can be applied to the group. Groups can also be explained by methods that produce global interpretability at a modular level.[220]

These distinctions lead to some initial conclusions about how AI can best explain itself to doctors and patients. At the point of adoption global explanations of model functionality seem appropriate to ensure a reliable fit between the intended use of the AI system in a given healthcare context, and the actual performance of the system. For explaining specific outputs or recommendations to patients, explanations of model behaviour formed as explanatory statements appear to strike the best fit between explaining the decision-making logic of the system while remaining comprehensible to expert and non-expert users alike. In this context methods such as 'counterfactual explanations' may be preferable as they facilitate debugging and testing of system performance by expert users while remaining comprehensible on an individual explanation level to non-expert patients.[221] To summarise, to make AI systems intelligible to patients, simple, local, contrastive explanations are preferable to global approximation explanations which can be difficult to understand and interpret.

An alternative but complementary approach is to use only intrinsically interpretable models in clinical care to enable health professionals to holistically understand systems and better explain them to their patients.[222] Implementing this approach would, however, create additional requirements for technical expertise in computer

[215] Lipton, *supra* note 164.
[216] MOLNAR, *supra* note 208 at 27.
[217] Guidotti et al., *supra* note 206.
[218] MOLNAR, *supra* note 208.
[219] *Id.*; Wachter, Mittelstadt, and Russell, *supra* note 172.
[220] MOLNAR, *supra* note 208.
[221] Wachter, Mittelstadt, and Russell, *supra* note 172.
[222] Cynthia Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, 1 NAT MACH INTELL 206–215 (2019).

science, statistics, and machine learning among health professionals which could be very difficult and perhaps unreasonable to meet in practice.

Public register of medical AI systems for transparency

As regards the issue of disclosure to patients of the usage of AI systems for operational and clinical purposes discussed in the section entitled "Transparency to health professionals and patients", the Parliamentary Assembly of the Council of Europe has recognised the importance of raising population awareness of uses of AI in healthcare to build trust with patients and ensure informed consent is possible in AI-mediated care. Specifically, their October 2020 report suggests that transparency of AI systems in healthcare "may require the establishment of a national health-data governance framework which could build on proposals from the international institutions. The latter include the Recommendation "Unboxing Artificial Intelligence: 10 steps to protect Human Rights" by the Council of Europe Commissioner for Human Rights (May 2019), the Ethics Guidelines for Trustworthy AI put forward by the European Union (April 2019), the OECD Recommendation and Principles on AI (May 2019) and the G20 Principles on Human-centred Artificial Intelligence (June 2019)."[223]

> Following these proposals and recommendations, a public database is seen as a key element to improve "algorithmic literacy" among the general public which is a fundamental precursor for exercising many human and legal rights.[224]

Insofar as the proposed framework is designed to increase population awareness of AI systems in healthcare, it can best be thought of as a type of public register for AI systems in healthcare. Registries are public lists of systems currently in use containing a standardised description of each system. Information included on registries varies but can include things like the intended usage or purpose of the system; its manufacturer or supplier; the underlying method(s) (e.g., deep learning, regression); any testing undergone both in terms of accuracy but also biases and other ethical and legal dimensions; a description of training and testing datasets; and an explanation of how predictions or outputs of the system are utilized by human decision-makers or otherwise integrated in existing services and decision-making processes.[225] Registries

[223] COUNCIL OF EUROPE, *supra* note 2.
[224] *Id.*
[225] Corinne Cath & Fieke Jansen, *Dutch Comfort: The limits of AI governance through municipal registers*, ARXIV PREPRINT ARXIV:2109.02944 (2021); Luciano Floridi, *Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki*, 33 PHILOSOPHY & TECHNOLOGY 541–546 (2020); Timnit Gebru et al., *Datasheets for Datasets* (2018), https://arxiv.org/abs/1803.09010 (last visited Oct 1, 2018); Margaret Mitchell et al., *Model Cards for Model Reporting*, PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY - FAT* '19 220–229 (2019); Sarah Holland et al., *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*, ARXIV:1805.03677 [CS] (2018), http://arxiv.org/abs/1805.03677 (last visited Oct 1, 2018).

also often have a feedback function to allow citizens to provide input on current and proposed uses of AI by public bodies and services.[226]

There are several examples of existing registries from municipal, national, and international public bodies. In 2020, Amsterdam and Helsinki launched public registries for AI and algorithmic systems used to deliver municipal services.[227] In November 2021, the UK Cabinet Office's Central Digital and Data Office launched a national algorithmic transparency standard which will effectively function as a type of public register.[228] Internationally, the recently proposed Artificial Intelligence Act contains a provision to create a public EU-wide database in which standalone high-risk AI applications must be registered.[229] The Council of Europe has an opportunity to complement these emerging transparency standards by introducing a public AI register for medical AI in member states which is aimed at patients to raise awareness of AI systems currently in use by their public health services.

Collection of sensitive data for bias and fairness auditing

Biases in AI systems linked to gaps in training and testing data could foreseeably motivate greater collection of sensitive data about legally protected groups for purposes of bias and fairness testing. It is a generally accepted fact, that in order to prevent discriminatory or biased outcomes, data on sensitive groups must be collected. Failure to collect this data will not prevent discrimination against protected groups, but arguably make it more difficult to detect.[230] Sensitive data is needed to test whether automated decision-making discriminated against groups based on protected attributes (e.g., data on race, disability, sexual orientation).[231] On the other hand, collecting such data has significant privacy implications. This is a legitimate concern and closely related to troubling historical experiences that significantly harmed specific groups in society.[232] For example, data collected for research and public purposes

[226] Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI, , VENTUREBEAT (2020), https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/ (last visited Dec 1, 2021).
[227] *Id.*
[228] UK government publishes pioneering standard for algorithmic transparency, , GOV.UK , https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency (last visited Dec 1, 2021).
[229] EUROPEAN COMMISSION, *supra* note 16 at Art. 51 and 60.
[230] SANDRA WACHTER, BRENT MITTELSTADT & CHRIS RUSSELL, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI* 34–35 (2020), https://papers.ssrn.com/abstract=3547922 (last visited Apr 19, 2020); Cynthia Dwork & Deirdre K. Mulligan, *It's not privacy, and it's not fair*, 66 STAN. L. REV. ONLINE 35 (2013); Cynthia Dwork et al., *Fairness Through Awareness*, ARXIV:1104.3913 [CS] (2011), http://arxiv.org/abs/1104.3913 (last visited Feb 15, 2016); Anupam Datta et al., *Proxy Non-Discrimination in Data-Driven Systems*, ARXIV:1707.08120 [CS] (2017), http://arxiv.org/abs/1707.08120 (last visited Jan 9, 2021); Kusner et al., *supra* note 189.
[231] Kusner et al., *supra* note 189; Chris Russell et al., *When worlds collide: integrating different counterfactual assumptions in fairness*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 6396–6405 (2017).
[232] MAYER-SCHÖNBERGER AND CUKIER, *supra* note 31; For a US and EU comparison see Joris Van Hoboken, *From collection to use in privacy regulation? A forward looking comparison of European and US frameworks for personal data processing*, 231 EXPLORING THE BOUNDARIES OF BIG DATA (2016); For an international view 63 LEE A. BYGRAVE, DATA PRIVACY LAW: AN INTERNATIONAL PERSPECTIVE (2014); For

have contributed to eugenics in Europe, the UK[233] and the US,[234] genocide during WWII, racist immigration practices and the denial of basic human rights in the US,[235] justification of slavery,[236] forced sterilisation in the UK,[237] US, Germany and Puerto Rico from the early to the mid-20th Century,[238] punishment, castration and imprisonment of LGBT members,[239] and denial to women of equal rights and protection (e.g. sexual violence).[240] Clearly, privacy interests must be taken seriously when considering collection of sensitive personal data for purposes of bias testing.[241]

Setting these concerns aside for a moment, one could be tempted to think that the bias problems will naturally be solved by collecting more (sensitive) data and closing gaps in representation in training and testing datasets. However, fair and equal outcomes will not automatically result when representation gaps and other data biases are closed. Awareness of inequalities is not the same as rectifying them.[242] Rather, the persistence of social biases across Western societies suggest that significant political, social, and legal effort is needed to overcome them, rather than simply more data collection and testing.

Countering inequalities requires intentional and often cost intensive changes to decision processes, business models, and policies. To justify further collection and usage of sensitive data, it is necessary to first demonstrate serious commitment and political will to rectifying inequality. From a standard setting perspective, these observations suggest that any proposed collection of sensitive category data for the sake of testing medical AI systems form biases must have clear purpose limitations and confidentiality guarantees in place alongside a commitment to rectify social inequalities underlying biases discovered through testing. Operationalizing these commitments is not straightforward. The EU Artificial Intelligence Act, for example, proposes the creation of "regulatory sandboxes" in which AI providers can test their

---

an European view Sandra Wachter, *Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR*, 34 COMPUTER LAW & SECURITY REVIEW 436–449 (2018); Sandra Wachter, *The GDPR and the Internet of Things: a three-step transparency model*, 10 LAW, INNOVATION AND TECHNOLOGY 266–294 (2018); for a EU and German view see Mario Martini, Wiebke Fröhlich & Saskia Fritzsche, *Algorithmen als Herausforderung fu r die Rechtsordnung* (2017); for empirical evidence of mobile data collection see Reuben Binns et al., *Third party tracking in the mobile ecosystem*, *in* PROCEEDINGS OF THE 10TH ACM CONFERENCE ON WEB SCIENCE 23–31 (2018); on online harms see Woods Lorna & Perrin William, *An updated proposal by Professor Lorna Woods and William Perrin*, https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/01/29121025/Internet-Harm-Reduction-final.pdf (last visited May 11, 2019).

[233] This happened until the 1930's, see RENI EDDO-LODGE, WHY I'M NO LONGER TALKING TO WHITE PEOPLE ABOUT RACE 20–21 (2020).

[234] JEAN HALLEY, AMY ESHLEMAN & RAMYA MAHADEVAN VIJAYA, SEEING WHITE: AN INTRODUCTION TO WHITE PRIVILEGE AND RACE 36 (2011).

[235] *Id.* at 25.

[236] *Id.* at 36–37.

[237] EDDO-LODGE, *supra* note 232 at 20–21.

[238] HALLEY, ESHLEMAN, AND VIJAYA, *supra* note 233 at 36–38.

[239] JEAN HALLEY & AMY ESHLEMAN, SEEING STRAIGHT: AN INTRODUCTION TO GENDER AND SEXUAL PRIVILEGE 15–17 (2016).

[240] SAINI, *supra* note 178 at 233–235.

[241] For surveillance and chilling effects, see JON PENNEY, *Chilling Effects: Online Surveillance and Wikipedia Use* (2016), https://papers.ssrn.com/abstract=2769645 (last visited Dec 27, 2017).

[242] EDDO-LODGE, *supra* note 232 at 208.

systems for bias using special category data collected explicitly for testing purposes.[243] This proposal lacks the essential element of a commitment to rectify discovered inequalities.

[243] EUROPEAN COMMISSION, *supra* note 16 at Art. 53.

8 CONCLUDING REMARKS

Medical care is increasingly diffused across a variety of institutions, personnel, and technologies. The doctor-patient relationship has always adapted over time to advances in medicine, biomedical research, and care practices. At the same time, the capacity of AI to replace or augment human clinical expertise utilising highly complex analytics and unprecedented volumes and varieties of data suggests the impact of the technology on the doctor-patient relationship may be unprecedented.

The adoption of AI need not be a fundamental barrier to good doctor-patient relationships. AI has the potential to alter care relationships and displace responsibilities traditionally fulfilled by medical professionals, but this is not a foregone conclusion. The degree to which AI systems inhibit 'good' medical practice hinges upon the model of service. If AI is used solely to complement the expertise of health professionals bound by the fiduciary obligations of the doctor-patient relationship, the impact of AI on the trustworthiness and human quality of clinical encounters may prove to be minimal.

At the same time, if AI is used to heavily augment or replace human clinical expertise, its impact on the caring relationship is more difficult to predict. It is entirely possible that new, broadly accepted norms 'good' care will emerge through greater reliance on AI systems, with clinicians spending more time face-to-face with patients and relying heavily on automated recommendations.

The impact of AI on the doctor-patient relationship remains highly uncertain. We are unlikely to see a radical reconfiguration of care in the next five years in the sense of human expertise being replaced by artificial intelligence. With that said, developments like the COVID-19 pandemic and the increased pressures it has placed on health services may transform the mode of delivery of care if not the expertise behind it. Remote delivery of care, for example, may become increasingly commonplace even if diagnosis and treatment remain firmly in the hands of human health professionals.

A radical reconfiguration of the doctor-patient relationship of the type imagined by some commentators, in which artificial systems diagnose and treat patients directly with minimal interference from human clinicians, continues to seem far in the distance. Movement in this direction continues to hinge on proof of clinical efficacy which, as noted above, continues to prove a barrier to commercialisation and widespread adoption.[244] Likewise, new modes of clinical care would need to be derived that utilise the best aspects of human clinicians and artificial systems, implement appropriate safety and resilience checks, and minimise the weaknesses and implicit biases of both agents. Without due consideration of the implications of AI for medical practice, the "moral integrity of the doctor-patient relationship" may come to be dominated by institutional and external interests, with patient experiences of care suffering as a result.[245]

[244] Liu et al., *supra* note 112; Robbins and Brodwin, *supra* note 5.
[245] Bauer, *supra* note 136 at 90.

As AI is adopted across different healthcare systems and jurisdictions, it is important to remember that the moral obligations of the doctor-patient relationship are always affected and perhaps displaced by the introduction of new care providers. While technology continues to develop at a rapid pace, the patient's experience of illness (e.g., vulnerability, dependency) and expectations of the healing relationship do not radically or quickly change. The doctor-patient relationship is a keystone of 'good' medical practice, and yet it is seemingly being transformed into a doctor-patient-AI relationship. The challenge facing AI providers, regulators, and policymakers is to set robust standards and requirements for this new type of healing relationship to ensure patients' interests and the moral integrity of medicine as a profession are not fundamentally damaged by the introduction of disruptive emerging technologies.

APPENDIX: MEDICAL VIRTUES

Virtues are defined against the ends of the practice which they are meant to serve. For medicine, these ends are providing adequate care for a society, consisting of individual patients, in terms of physical and mental health and well-being. These ends are realised through the healing relationship, the nature of which introduces certain moral obligations.

As with all practices, phronesis or prudence is a central virtue in medicine, without which other virtues cannot be incorporated into behaviour through virtuous acts.[246] Justice, truthfulness and courage are also necessary to protect medicine from the corrupting power of medical institutions, including hospitals, paying organisations and government departments.[247] These three core virtues are necessary for continuous revision of standards of excellence and internal goods by practitioners, which requires critical self-reflection on the relationship between one's actions and the norms of the practice, or the institutional influence on the definition and realisation of norms.[248]

Justice is defined broadly as "the strict habit of rendering what is due to others,"[249] or "the virtue of rewarding desert and of repairing failures in rewarding desert within an already constituted community."[250] To be just, standards for treating people in a community must be "uniform and impersonal," meaning it is unjust to favour personal acquaintances. In social or national healthcare systems, justice can be applied to the distribution of medical resources (e.g., pharmaceuticals, treatments, clinical encounters) in a manner fair to all stakeholders. Justice is not merely a quantitative notion, by which all stakeholders receive an equal share, but instead requires matching resources to the needs of the patient and making judgments between the relative importance of different needs.

Fidelity to trust and beneficence can also be understood as core virtues unique to medicine because of the need for trust in healing relationships.[251] A trusting relationship needs to develop over time between the virtuous doctor and patient, in which the values, expectations and thoughts on illness and appropriate medical care are shared. The patient must at a minimum believe the doctor is acting beneficently, or in his interests and well-being, to some degree for trust to exist.[252]

---

[246] MACINTYRE, *supra* note 122 at 154; G. Widdershoven & Lieke Van der Scheer, *Theory and methodology of empirical ethics: a pragmatic hermeneutic perspective*, *in* EMPIRICAL ETHICS IN PSYCHIATRY 23–36 (2008), http://books.google.co.uk/books?hl=en&lr=&id=Lvq0lkDyEBQC&oi=fnd&pg=PA23&dq=Theory+and+methodology+of+empirical+ethics:+a+pragmatic+hermeneutic+perspective&ots=IXt3OC6Obh&sig=EU-idi92-6EzBI6uTp8UNReq4AY#v=onepage&q&f=false; PELLEGRINO AND THOMASMA, *supra* note 120.
[247] MACINTYRE, *supra* note 122 at 192.
[248] *Id.* at 191.
[249] PELLEGRINO AND THOMASMA, *supra* note 120 at 92.
[250] MACINTYRE, *supra* note 122 at 156.
[251] PELLEGRINO AND THOMASMA, *supra* note 120 at 71, 156.
[252] *Id.* at 156.

Other virtues include compassion, fortitude, integrity and temperance. Compassion is the trait of a doctor which allows him to 'enter the perspective' of the patient, to understand how the patient's values, expectations of care, social, emotional and physical well-being affect his experience of illness, and to customise his care and recommendations to the needs of each patient as a unique individual.[253] Compassion may also necessitate the promotion of health-related values and deliberation with the patient to convince him of the best intervention in terms of fit between health outcomes as perceived by the doctor and the patient's values.[254]

Fortitude is a form of moral courage, by which an individual is willing to "suffer personal harm for the sake of a moral good" such as a doctor refusing to act in accordance with institutional rules which would be detrimental to his patient's well-being, risking harm to his career and professional membership.[255] Fortitude can create an obligation for doctors to speak out against the potential harms of new institutional policies, technologies or treatments for their patients. Temperance is the restriction of behaviour in a practice to meet the moral obligations of that practice. It can be used synonymously with virtue itself but is distinct as a character trait of the virtuous doctor who suppresses self-interest in treating patients. Without such restraint other virtues cannot be practiced.[256]

Integrity is the possession of all virtues combined with the ability to discern between moral principles in choosing appropriate actions conducive to the good of medicine in different situations.[257] It is the core virtue of the narrative quest for the good life, and can be seen in a life of virtuous behaviour.[258] Integrity can be exercised when a doctor promotes the patient's interests and welfare in the face of institutional pressure, for example by not sending a patient home early from hospital.[259] Edgar and Pattison define integrity as "the capacity to deliberate and reflect usefully in the light of context, knowledge, experience and information (that of self and other) on complex and conflicting factors bearing on action or potential action."[260] Integrity is therefore perhaps indistinguishable from phronesis, temperance and fortitude.

[253] *Id.* at 79, 81.
[254] Emanuel and Emanuel, *supra* note 123 at 2226.
[255] PELLEGRINO AND THOMASMA, *supra* note 120 at 109.
[256] *Id.* at 117.
[257] *Id.* at 127.; Edgar and Pattison, *supra* note 145 at 102.
[258] MACINTYRE, *supra* note 122.
[259] Edgar and Pattison, *supra* note 145 at 94.
[260] *Id.* at 102.