Charlotte Högberg & Stefan Larsson

AI and Patients' Rights: Transparency and Information Flows as Situated Principles in Public Health Care

Abstract

The development of artificial intelligence (AI) for medicine and health care is rapidly evolving. However, the automation, scale and data dependency of AI-driven decision-making and decision-support calls for a reassessment of principal ethical and legal norms of transparency, in the light of these novel methodologies. The quality of AI-driven health care, we argue, is depending on it. In this chapter, we provide an overview of novelties that AI in health care bring about, in order to identify key aspects potentially affecting current legal and normative (medical ethical) principles related to transparency and explainability. We develop a conceptual framework on transparency in general and explainability in particular, in relation to AI in health care. Further, we analyse principal and normative legal frameworks of patients' rights relating to transparency and explainability – e.g., right to information, autonomy and privacy – within Sweden and the EU. Doing so, we outline main challenges in the implementation of AI in, primarily public, health care. We argue that there is an interdependency between health care quality and transparency. As transparency is not a binary state, but something that is *situated* in information practices, it is important to consider what kind of transparency is needed to safeguard the best possible health care. We find that meaningful and contextual transparency and explainability of AI-systems and methodologies is necessary to adhere to the basic principles of normative and legal frameworks of Swedish health care, including

patient autonomy. In addition, meaningful and contextual transparency is also a prerequisite for assessing if the best possible care is given to the one most in need.

1 Introduction

According to the modernized version of the Hippocratic oath, The Declaration of Geneva, a physician's main priority should be the health and well-being of the patient.[1] This ideal and other moral values – such as to respect human life, the integrity and autonomy of the patient, the patient's right to information, and to conduct care in an ethical manner and use medical knowledge for good – are considered principles for medicine and health care. These can be found in a wide array of policies, legal frameworks and guidelines. Technological innovations in drug discovery, treatments, diagnostic tools and so forth, have contributed to the fulfilment of these values and to improved chances for health and longevity. Still, the implementation of new technologies can pose ethical and legal challenges to ideals of medicine and healthcare. Technology also tends to develop faster than regulations adapt, causing *the pacing problem* – as pointed out in socio-legal studies.[2] Many of the latest technological innovations in medicine and health care are based in Artificial Intelligence (AI) and machine learning in particular. AI consists of a broad collection of technologies and methods. As described by Dignum:

> [AI] deals not only with how to represent and use complex and incomplete information logically but also with questions of how to see (vision), move (robotics), communicate (natural language, speech) and learn (memory, reasoning, classification).[3]

While AI is not new, it is now a fast-growing field due to increased access to data, computer power, and the creation of new and improved machine learning models. This is true also for AI within medicine and health care.

[1] 'Declaration of Geneva' (World Medical Association 2021) <https://www.wma.net/policies-post/wma-declaration-of-geneva/> accessed 2021-05-20.

[2] E.g., Stefan Larsson, 'AI in the EU: Ethical Guidelines as a Governance Tool', in Antonia Bakardjieva Engelbrekt, Karin Leijon, Anna Michalski & Lars Oxelheim (eds.) *The European Union and the Technology Shift*. (Palgrave Macmillan 2021).

[3] Virginia Dignum, 'Introduction' in Virginia Dignum (ed.), *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer International Publishing 2019) 3.

The purpose of this chapter is to outline main transparency challenges of implementing AI in the public health sector. In order to do this, we adress the following research question:

- What is the role of transparency and explainability of artificial intelligence in relation to patients' rights and information flows in Swedish health care?

Our main methodology is a qualitative analysis of patients' rights in health care information flows, which could be affected by the use of applications based on AI. In the remainder of this section, we discuss the role of AI in healthcare. In section 2 we present a theoretical discussion of transparency and explainability in relation to AI. This theoretical discussion forms the foundation for a socio-legal analysis (section 3) of documents representing medical ethics, as exemplified by the Declaration of Geneva,[4] and an array of relevant legal frameworks at both EU level as well as at the Swedish level. The most central regulations at the EU level are the General Data Protection Regulation, GDPR,[5] in force since May 2018, and the Medical Device Regulation, MDR,[6] that is fully applicable since May 2021. In Sweden, at the national level, the main regulatory instruments are the Health and Medical Service Act,[7] the Patient Act,[8] the Patient Safety Act[9] and the Patient Data Act.[10]

Clearly, however, this list is not exhaustive. The rights of patients and the obligations of health care providers and health professionals are regulated by a large number of national and international laws, including Swedish constitutional law, as well as by extra-legal norms, ideals, global policies, common standards and local guidelines. In different ways, they concern the legitimacy of information flows. Analysing all of these and

[4] 'Declaration of Geneva' (n. 1).

[5] Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Hereinafter cited as GDPR.

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

[7] Hälso- och sjukvårdslag (SFS 2017:30).

[8] Patientlag (SFS 2014:821).

[9] Patientsäkerhetslag (SFS 2010:659).

[10] Patientdatalag (SFS 2008:355).

their hierarchical validities and mutual relations goes beyond the scope of this chapter. Instead, the goal of this chapter is to pinpoint main informational principles within the legal and normative frameworks of medicine and health care, which may be affected by the implementation of AI. These informational principles include the right to (equal) health care (section 3.1), the right to privacy and integrity (section 3.2), the right to information (section 3.3), and the right to dignity and autonomy (section 3.4). In section 4 we discuss how an AI application with high predictive accuracy but with low transparency does not lead to the best possible care, if the lack of transparency means that the aforementioned informational principles are not adhered to. In section 5 we summarize the main argument and findings presented in this chapter.

1.1 Background: The promises and challenges of AI in health care

Medicine and health care are important fields of application for AI. The use of AI-systems and methodologies in healthcare could have a beneficial, or even vital, impact if it would result in improved predictions, diagnoses and prognoses of diseases. The broader effects could be better health, improved well-being and an increased amount of successful outcomes of treatments. A desired goal of AI implementation is also increased efficiency, especially as the health sector is also facing increased costs and administrative burdens, scarcity of practitioners and aging populations.[11] Another hope is the personalisation of medicine, as stated by experts in the field:

> Machine learning will become an indispensable tool for clinicians seeking to truly understand their patients. As patients' conditions and medical technologies become more complex, the role of machine learning will grow, and clinical medicine will be challenged to grow with it.[12]

[11] E.g., *Ethics and governance of artificial intelligence for health: WHO guidance* (World Health Organization 2021), Arash Shaban-Nejad, Martin Michalowski and David L. Buckeridge, 'Explainability and Interpretability: Keys to Deep Medicine' in Arash Shaban-Nejad, Martin Michalowski and David L. Buckeridge (eds.), *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (Springer International Publishing 2021).

[12] Ziad Obermeyer and Ezekiel J. Emanuel, 'Predicting the Future – Big Data, Machine Learning, and Clinical Medicine' (2016) 375 N Engl J Med 1216 1218.

One of the areas in the forefront is image analysis for radiology,[13] where AI-systems have, for example, been found to have an accuracy in cancer detection comparable to average breast radiologists.[14] AI is also used for other types of clinical decision support-tools, as well as in the form of conversational AI, offering more and faster ways of communication. It is furthermore used for administrative purposes, such as scheduling staff, allocating resources and making cost predictions.[15] Contributing to this development is the increase of digital health data and the datafication of health care.[16] The Scandinavian countries have a possible advantage due to large amounts of public health data, such as in national registers. The instated *Vision for e-health* declares that by the year 2025, Sweden should be world leading in using the opportunities provided by digitalization and e-health.[17] One factor identified as important to the realization of this vision is the implementation of AI.[18]

In brief, there is a large number of hurdles within medicine and health care that could potentially be overcome with the help of AI. However, alongside the great potential there are also significant social, ethical and legal challenges, especially with regards to the high stakes of life and death.[19] These challenges include risks for patient safety, treatment of outliers, concerns whether systems will be able to differentiate between

[13] J. Raymond Geis and others, 'Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement' (2019) 293 Radiology 436.

[14] Alejandro Rodriguez-Ruiz and others, 'Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists' (2019) 111 JNCI: Journal of the National Cancer Institute 916.

[15] E.g., Eric Racine, Wren Boehlen and Matthew Sample, 'Healthcare uses of artificial intelligence: Challenges and opportunities for growth' (2019) 32 Healthcare Management Forum 272, World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance* (n. 11), K. H. Yu and I. S. Kohane, 'Framing the challenges of artificial intelligence in medicine' (2019) 28 BMJ Qual Saf 238.

[16] E.g., Minna Ruckenstein and Natasha Dow Schüll, 'The Datafication of Health' (2017) 46 Annual Review of Anthropology 261.

[17] E-hälsa 2025, 'Om vision e-hälsa 2025' (2021) <https://ehalsa2025.se/visionen/> accessed 2021-05-20.

[18] E-hälsomyndigheten, *Fokusrapport – Artificiell intelligens och e-hälsa*, (2020).

[19] E.g., A. Blasimme and E. Vayena, 'The Ethics of AI in Biomedical Research, Patient Care, and Public Health' in S. Das, Pasquale, F. and Dubber, Markus D. (ed.), *The Oxford Handbook of Ethics of AI* (Oxford University Press 2020), Titti Mattsson and Vilhelm Persson, 'E-hälsa' in Kavot Zillén, Titti Mattsson and Santa Slokenberga (eds.), *Medicinsk rätt* (Nordstedts Juridik 2020).

correlation and causality, and risks of unfair treatment due to bias regarding such as gender, ethnicity and age. Other risks are *competence loss* (in the sense that human knowledge of certain medical skills may erode, due to AI-systems taking over or heavily assisting the task), *automation bias* (if health professionals are over reliant towards AI systems), *overtreatment and overmedicalization*, *risk of integrity breaches*, as well as concerns about the *effects on responsibility, liability and trust*.[20]

Currently, and drawn to their extremes, at least two different discourses are demonstrated simultaneously; an idea of solutionism wherein AI-systems and methodologies will be the answer to if not all, at least most, problems, as well as a dystopian view of maleficent biased autonomous systems. Both views are adhering to deterministic views of technology, but, as laid out by Bucher: "there is nothing inherently neutral about algorithms or biased about humans, these descriptive markers emerge from particular contexts and practices."[21]

There are more balanced hopes for AI in health care, yet, the large interest, great expectations and inflated hopes seem to be fueled by the stakes involved: adopting AI in health care can literally be a matter of life and death. It also represents a possible profitable area of application for product developers, making commercial interest a contributing factor as well. However, as emphasized by legal and medical researchers, the use of AI in healthcare might also undermine traditional principles of medical law and patients' rights.[22] Another question that is raised is if patients have the right to refuse being subject to AI-systems.[23] But what is it with AI, compared to previously implemented technologies, that constitutes grounds for concerns?

[20] E.g., Obermeyer and Emanuel, 'Predicting the Future – Big Data, Machine Learning, and Clinical Medicine' (n. 12), Ziad Obermeyer and others, 'Dissecting racial bias in an algorithm used to manage the health of populations' (2019) 366 Science 447, Jessica Morley and others, 'The ethics of AI in health care: A mapping review' (2020) 260 Social Science & Medicine 113172.

[21] Taina Bucher, *If…then: algorithmic power and politics* (Oxford University Press 2018) 56.

[22] Iñigo de Miguel, Begoña Sanz and Guillermo Lazcoz, 'Machine learning in the EU health care context: exploring the ethical, legal and social issues' (2020) 23 Information, Communication & Society 1139.

[23] T. Ploug and S. Holm, 'The right to refuse diagnostics and treatment planning by artificial intelligence' (2020) 23 Med Health Care Philos 107.

1.2 The novelty of AI

Neither technical complexity nor information technologies are novel to Swedish health care, so what is new with AI-systems and methodologies in the clinical setting? Health professionals do not know the inner workings of all non-AI medical equipment, but hopefully the main logic behind their results or have a *human-in-the-loop* who could provide such explanations if needed.[24] The following characteristics of novelty, and associated opportunities and risks, are gathered from both the growing medical AI literature, as well as the wider media and communications and STS literature on automation and *datafication* in contemporary society. In short, from a transparency-focused and sociotechnical approach, AI-systems and methodologies may contribute to:

1. An increased *automation* of decision-making processes, with the benefits of efficiency, speed and avoiding dependency on overworked medical staff, which of course is highly attractive to these domains, but also the risks of reproducing historical skewness without sufficient oversight or scrutiny (including the impact of automation bias on human decision-making).[25]
2. *Large-scale adoption* as a result of automation. While having similar benefits as automation, it also entails both the advantage of excellence not being limited to certain human actors, as well as the heightened risk of errors or subjective prejudice decisions being applied on a large-scale as built-in features affecting large populations.[26]
3. *Opacity* resulting from the "black-box" nature of some AI-systems, with the risk of lacking explainability in complex algorithmic models, or systemic lack of transparency as AI-systems are applied in proprietary settings, with a complex array of data-sharing entities.[27]
4. *Data-dependency*, with large-scale quantification and "datafication" of everyday activities, which at best contributes to insights-driven incen-

[24] Jens Christian Bjerring and Jacob Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (2021) 34 Philosophy & Technology 349 364 P.364.

[25] C.f. Stefan Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (2019) 103 Droit et société 573.

[26] C.f. studies by media sociologist Jonas Andersson Schwarz, 'Platform Logic: An Interdisciplinary Approach to the Platform-Based Economy' (2017) 9 Policy & Internet 374; or on the platformisation of data-driven platforms, Stefan Larsson, 'Putting trust into antitrust? Competition policy and data-driven platforms' (2021) European Journal of Communication 02673231211028358.

[27] This is extensively developed in the following section.

tives and evidence-based decision-making, but with the risk of being at odds with established and regulated ideas of privacy, data-minimisation and rights to be forgotten, and that could result in the creation of what could be termed medical surveillance.[28]

5. *Obscured causability*. While medicine has always been considering multiple factors simultaneously (anamnesis, blood samples, measurements, etc.), automated medical decisions and classifications could propose hardships of deciphering which variables have led to a decision, and whether co-occurrences are wrongfully treated as causes.[29]
6. A *personalisation* of medicine, with the possibility of tailored drugs and treatments, as well as the risk of privacy breaches, challenged autonomy, mistreatment and discrimination.[30]
7. An increased *private-public complexity*, which is of particular relevance in Sweden, as Swedish health care is to a large extent public, while at the same time reliant on corporate service developers. The complex intertwinement of private and public dimensions are also relevant in terms of how this complexity can be handled from a regulatory perspective. It may be problematic from the public scrutiny point-of-view, if it hinders transparency of public sector organisations (see point 3 above,) or from the challenge of balancing the benefits of publicly collected data being used to train private AI-systems sold on markets.[31]

[28] E.g., Ruckenstein and Schüll, 'The Datafication of Health' (n. 16).

[29] E.g., Andreas Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (2019) 9 WIREs Data Mining and Knowledge Discovery e1312, Sendhil Mullainathan and Ziad Obermeyer, 'On the Inequity of Predicting A While Hoping for B' (2021) 111 AEA Papers and Proceedings 37.

[30] For an overview of both promises and pitfalls, see the editorial for a special issue on the subject, T. Feiler and others, 'Personalised Medicine: The Promise, the Hype and the Pitfalls' (2017) 23 New Bioeth 1.

[31] For example, pointed to as a challenge in studies on smart cities, Robert Brauneis & Ellen P. Goodman, 'Algorithmic transparency for the smart city' (2018) 20 *Yale JL & Tech.* 103, as well as pointed to in terms of the importance of improved procurement by the High-Level Expert Group on AI. See also Mattsson and Persson, 'E-hälsa' (n. 19) and Obermeyer and Emanuel, 'Predicting the Future – Big Data, Machine Learning, and Clinical Medicine' (n. 20).

2 Framing AI-transparency in health care

This section outlines main theoretical notions of AI-transparency in general, and in relation to medicine and health care in particular, in order to facilitate an analysis of its role for patients' rights. We argue that transparency is a wide concept, encompassing for example the more computer-scientific notion of explainability of AI-systems.[32] As mentioned, the lack of interpretability of AI is commonly described as "black-box".[33] The term can refer to a model that is too complicated to be interpretable, with only poor insights in how the training based on large data-sets reached a particular functionality or precision, or a model that is proprietary and hidden from external review.[34] It can also be both. This has led to a call for *transparency* of AI. Transparency is considered a key prerequisite for trustworthy AI, as stated by the EU high-level expert group and also mirrored in the current EU proposal for an AI regulation (AIA), published in April 2021.[35] But what does this idea of transparency entail?

2.1 AI transparency

Transparency is a multifaceted concept, as stressed by Larsson and Heintz,[36] often caught in a trade-off between different types of interests.[37] The term can be seen as a metaphor, where the material physical state of transparency – the see-through nature of an object – is deployed to describe cognitive, social, organizational phenomena and relations, and is

[32] E.g., Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

[33] Frank Pasquale, *The black box society: the secret algorithms that control money and information* (Harvard University Press 2015), Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI' (2019) Proceedings of the Conference on Fairness, Accountability, and Transparency 279.

[34] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' (2019) 1 Nature Machine Intelligence 206.

[35] High-Level Expert Group on Artificial Intelligence (HLEG), *Ethics guidelines for trustworthy AI*, 2019, hereinafter cited as HLEG 2019, 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 COM/2021/206 final. Hereinafter cited as AIA.

[36] Stefan Larsson and Fredrik Heintz, 'Transparency in artificial intelligence' (2020) 9 Internet Policy Review.

[37] For a discussion of seven different aspects, such as proprietary claims versus explainability versus human literacy, see Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

based on the idea that what can be seen can be known (seeing is knowing).[38] Transparency can be considered a normative socio-legal ideal. In general, transparency is a broad concept, and this is also true regarding how the term is used in relation to AI. It can be used aiming at data underlying or used by AI models, algorithms and their logic, governance of AI-models, and so forth. Algorithmic transparency is a commonly used concept as well, but it could be misleading as AI in use is more than the function of algorithms,[39] and hence meaningful transparency need to encompass more than that.[40]

As pointed out above, there are several ways in which AI-systems can be opaque. Accordingly, three different forms of opacity are identified by Burrell: (1) intentional corporate or state opacity due to secrecy reasons, (2) opacity due to technical illiteracy and (3) opacity due to scale of operation of algorithms.[41] Another set of distinctions is made by Ferretti et al.: lack of disclosure, epistemic opacity and explanatory opacity.[42] Further, transparency in public decision-making can be described as information disclosure of different degrees, as described by de Fine Licht and de Fine Licht: informing about what the final decision (or recommendation or classification) is, about the process resulting in the decision (transparency in process) and about the reasons behind the decision (transparency in rational).[43]

Transparency is a vague concept, as pointed out by de Vries, stressing the need to ask: transparency of what, to whom, and when?[44] In addition, one must define what the problem is, if transparency is to be the answer. One issue, related to the vagueness, is the binary notion by which the concept of transparency is often used. Lee argues that we should not consider algorithms as binary, being either opaque or transparent. Instead,

[38] Larsson and Heintz, 'Transparency in artificial intelligence' (n. 36).

[39] Dignum, 'Introduction' (n. 3).

[40] Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

[41] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) 3 Big Data & Society 2053951715622512.

[42] Agata Ferretti, Manuel Schneider and Alessandro Blasimme, 'Machine Learning in Medicine: Opening the New Data Protection Black Box' (2018) 4 European Data Protection Law Review (EDPL) 320.

[43] de Fine Licht and de Fine Licht, 'Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy' (2020) 35 AI & Society 917 918.

[44] Katja de Vries, 'Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM' (2021) 8 Critical Analysis of Law 121 124.

we need to consider them contextually and in practice, to analyse how agency and power is constructed. There are different degrees of agency and opacity in different parts of *algorithmic assemblages*, which are always situated in practice:

> …algorithmic assemblages can be differently understood in different situations and are therefore neither completely opaque, nor completely transparent. Opacity is not only varying in degree or type, but also varies depending on the actor's situatedness.[45]

One important goal of transparency is to enable the assessment of, and demand for, fairness and accountability.[46] In the Swedish context, there is a far-reaching ideal of transparency of public administration in general. This encompasses publicly run health care and research institutions. Their decision making and procurement of technologies need to be to some degree interpretable, explainable and open for scrutiny. The patient's own access to electronic health data in journal systems is an example of transparency in practice, also pushed by legislation such as the GDPR.[47] The recently applied Medical Device Regulation promotes transparency within the health sector, for the purpose of medical safety.[48] In the AIA, transparency is also emphasized as a key component for safeguarding fundamental rights.[49]

2.2 Towards explainable AI

As a response to the call for AI-transparency, one part of the solution put forward is increased *explainability* (by some considered a concept included under the transparency 'umbrella').[50] The EU High-Level Expert Group on AI identify explainability as a core element of transparency, together with traceability and communication.[51] The urgency of transparency and explainability for AI in health care is echoed from a multi-

[45] Francis Lee, 'Enacting the Pandemic: Analyzing Agency, Opacity, and Power in Algorithmic Assemblages' (2021) 34 Science & Technology Studies 65 17.
[46] C.f. Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).
[47] GDPR (n. 5).
[48] MDR (n. 6).
[49] AIA, Explanatory Memorandum, 2.3 (n. 35).
[50] Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).
[51] HLEG 2019 (n. 35).

tude of perspectives.[52] Interpretability and explainability are considered crucial aspects to achieve trustworthiness, emphasized by for example the World Health Organization's guidance on ethics and governance of AI for health, according to which one of the key principles is to ensure transparency, explainability and intelligibility:

> AI technologies should be intelligible or understandable to developers, medical professionals, patients, users and regulators. Two broad approaches to intelligibility are to improve the transparency of AI technology and to make AI technology explainable.[53]

Explainable AI, also known as "xAI", can be defined as "a characteristic of an AI-driven system allowing a person to reconstruct why a certain AI came up with the presented predictions."[54] However, as with AI and transparency, there is not one agreed upon definition of what is included in the concept. Explainability has many facets and the terms transparency and interpretability are often used synonymously.[55] Lipton states that explainability is used to refer to some form of model interpretability, and distinguishes different forms that AI explanations can take: text (e.g., generated captions), visualizations (e.g., generated images), local explanations (e.g., gradient map masks, highlighting influential areas for classification of images) and explanation by example. The latter could be in the form of generated nearest neighbours and counterfactuals.[56] One could also distinguish between post-hoc explainable systems, providing local explanations on demand, and ante-hoc systems, built with "glass-

[52] E.g., Shaban-Nejad, Michalowski and Buckeridge, 'Explainability and Interpretability: Keys to Deep Medicine', Julia Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' (2020) 20 BMC Medical Informatics and Decision Making 310, Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

[53] World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance*. Xiii (n. 11).

[54] Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' 2 (n. 52).

[55] Ibid. (n. 52).

[56] Zachary C. Lipton, 'The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery' (2018) 16 Queue 31, de Vries, 'Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM' (n. 44), Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 Harvard Journal of Law & Technology (Harvard JOLT) 841.

box" approaches aiming to be interpretable by design.[57] Yet another distinction is whether the explanation should concern a system's general functionality, components of models, the training algorithm or rather of specific decisions.[58]

Insights on what constitutes a useful explanation can be found in diverse fields within the social sciences, as laid out by Miller, the main being that: why-questions are *contrastive* (responses to counterfactual cases), explanations are *selective*, causality is of greater importance than probabilities, and lastly, explanations are *social* as well as transfers of knowledge, in likeness to conversations or interactions.[59] Miller argues that all these factors converge around one single important point:

> [E]xplanations are not just the presentation of associations and causes (*causal attribution*), they are *contextual*. While an event may have many causes, often the explainee cares only about a small subset (relevant to the context), the explainer selects a subset of this subset (based on several different criteria), and explainer and explainee may interact and argue about this explanation.[60]

To be of use to involved parties, explanations should be "contrastive, selective, and social" rather than limited to models in science.[61] There is also an emphasis on "target audiences" in some of the literature on explainable AI,[62] that is, the awareness of that different types of addressees, such as medical staff, patients, and developers of AI-systems, will be having different types of need for what the explanations should hold.

On the other hand, the workings of AI-tools are also discussed as something that could deliberately be kept in the dark. To some extent this can be due to arguably valid reasons, such as protection of intellectual property rights or protection against maleficent gaming of systems or

[57] Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' 5 (n.29).

[58] Mittelstadt, Russell and Wachter, 'Explaining Explanations in AI'. (n 33).

[59] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences' (2019) 267 Artificial Intelligence 1.

[60] Ibid. p. 3.

[61] Mittelstadt, Russell and Wachter, 'Explaining Explanations in AI' (n. 33).

[62] E.g., Alejandro Barredo Arrieta and others, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI' (2020) 58 Information Fusion 82.

security breaches.[63] Even though actors share the main goal of wanting to improve health care and well-being of patients, there can be reasons, intentional or unintentional, for actors to oppose meaningful transparency of their products. Bucher discusses how algorithms can function as *strategic unknowns*, where the opaqueness is deliberately maintained and could also be used as an advantage (which could, for example, be an inherent incentive in the private-public complexity pointed to above). Bucher argues that there is a risk of misplaced focus to keep telling the tale of the "black box"-ness of algorithms as something static and unavoidable, since it could serve different functions and be used as excuse for letting them continue to be kept opaque.[64]

It has been pointed out that there is a trade-off between better performance and explainability, meaning that improved prediction and accuracy, by applying more complex techniques, comes at the cost of decreased possibilities to interpret the models. However, Rudin argues that such a trade-off is not always given; sometimes there are models that score high on both interpretability and accuracy. Instead of trying to make black-box models explainable post-hoc, it should be an ex-ante concern to choose inherently interpretable models if they are going to be used for high-stakes areas such as health care.[65] In the medical context, relevant results can be found from diverse sets of data, hence it needs to be possible for practitioners to understand how and why a decision was made, as noted in the point on obscured causality in the novelty characteristics. Holzinger et al. argue that we need to go further than *explainable* AI, we also need *causability* to reach actual explainable medicine, by providing "causes of observed phenomena in a comprehensible manner through a linguistic description of its logical and causal relationships."[66]

Moreover, opaqueness of black-box medicine is at conflict with ideals of patient-centered medicine, argues Bjerring and Busch.[67] If AI-supported systems are expected to perform better than physicians, this creates a situation of epistemic obligation, where the physician has to follow

[63] C.f. Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

[64] Bucher, *If…then: algorithmic power and politics* (n. 21).

[65] Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' (n. 34).

[66] Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (n. 29).

[67] Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

the system's recommendation. If not sufficiently explainable, they cannot explain how and why they give the recommendation or come to a certain conclusion. Hence, the patient cannot be adequately informed and not make an autonomous and rational decision.[68] What is needed is to make the AI *decision-making* explainable, which is required both by the patients and the physicians, for the latter not to be merely operators of non-comprehensible AI decisions.[69] If clinical decision support systems are omitting explainability, it threatens core ethical values of medicine, Amann et al. argue. From a legal perspective, they identify three core fields where xAI in health care is needed: "(1) Informed consent, (2) Certification and approval as medical devices (acc. to Food and Drug Administration/FDA and Medical Device Regulation/MDR) and (3) Liability."[70]

In sum, transparency and explainability can be considered important tools to even power-imbalances of the information (and knowledge) asymmetry at play in the health sector. Before an analysis of patient rights, and how they relate to AI, obligations of care providers and health professionals and information flows, we need to consider the state of rights and obligations in this context.

3 Rights, obligations and power dynamics in the context of health care

The following section outlines patients' rights of most relevance to transparency and explainability in relation to AI, primarily focusing on health care equality, issues of privacy and integrity, the right to information and patient autonomy, as well as what impact the implementation of AI could have on them.

There are national differences in how legal frameworks position the patient. For example, in Norway there is a patient's *rights* act,[71] while in

[68] Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

[69] Thomas Hoeren and Maurice Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (2018) 4 European Data Protection Law Review (EDPL) 308.

[70] Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' p. 3 (n. 52).

[71] E.g., E. M. Aasen and B. M. Dahl, 'Construction of patients' position in Norway's Patients' Rights Act' (2019) 26 Nurs Ethics 2278.

Sweden the legal framework that specifically concerns health care is not expressed as a regulation of patients' legal rights, but in terms of *obligations* of actors (state, region, health care providers, health professionals) towards patients (in the Health and Medical Service Act, Patient Act, Patient Safety Act, Patient Data Act).[72] In this way the patient's rights are only implicitly expressed and a patient has limited possibilities to legally challenge medical decisions by judicial proceeding, as few of them are considered administrative legal decisions, argues Johnsson.[73] Health care providers and health professionals have to provide medical care in accordance to the legislation and can be held accountable for any wrongdoings. Rights in relation to health care can also be considered to belong to the public at large, with transparency as tool for accountability of public administration in accordance with the public access to information principle in Swedish law.[74]

In health care there are many actors involved: there are those who provide, those who receive, and those who facilitate or steer care. The different roles come with a difference in power. The starting point of discussing patients' *rights* and caregivers' *obligations* is in itself telling of a structure of power dynamics. The patient is considered in need of rights towards caregivers and the health system, due to the fact that the patient is considered to be in a position of less power in comparison to the other actors. Simultaneously, health professionals have obligations to not abuse their position of power. This dynamic can be found in several different instances, such as the patient being exposed to physical examinations and procedures, possibly life-saving or life-threatening, but also in terms of knowledge and information. The patient is in the hands of health care systems and practitioners who in general know more about how the system works, the specific medical condition and the treatments, as well as the medical state and personal sensitive health information of the individual patient (although this could be argued to not always be the case). This constitutes an *information asymmetry*. In the context of AI in the health sector, additional actors in this equation are the developers of AI models, tools and systems; that is, computer scientists, statisticians,

[72] Hälso- och sjukvårdslag (SFS 2017:30), Patientlag (SFS 2014:821), Patientsäkerhetslag (SFS 2010:659), Patientdatalag (SFS 2008:355).

[73] Lars-Åke Johnsson, 'Patientens ställning i vården och personalens skyldigheter' in Kavot Zillén, Mattsson, Titti, Slokenberga, Santa (ed.), *Medicinsk rätt* (Nordstedts Juridik 2020) 73.

[74] Tryckfrihetsförordning (1949:105), ch. 2.

medical researchers and more, and also the commercial entities or public suppliers of end products for screening, data handling or various types of predictions, etc. These other actors are in possession of yet another set of knowledge and informational power that health professionals and patients lack.

3.1 The right to (equal) health care

While access to health care varies greatly, globally as well as within nations, the right to health care is included as a basic human right, by article 25 of the United Nations' Universal Declaration of Human Rights.[75] In Swedish law, The Health and Medical Service Act reads that medical care should be provided with respect to equal value of all humans and the dignity of each individual.[76] In addition, the Declaration of Geneva points out; "I WILL NOT PERMIT considerations of age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor to intervene between my duty and my patient."[77] Even though the said ideals and regulations exist, medicine and health care are not free from prejudicial and discriminatory practices. This is exemplified in reports by health professionals[78] as well as in research, such as by studies showing that women's expressions of pain are treated less serious than those of men,[79] and cases of racist interpretations leading to misdiagnoses or deprivation of treatment.[80] When AI-systems are trained on historical (or simply biased) data, mistreatment and discrimination could be reproduced and upscaled. This demands an awareness of what is built into the processes of data collection, labelling and interpretation, being the basis for learning algorithms.[81] Discrimi-

[75] Universal Declaration of Human Rights, (United Nations 2021), <https://www.un.org/en/about-us/universal-declaration-of-human-rights> Art. 25 accessed 2021-05-25.

[76] Hälso- och sjukvårdslag (SFS 2017:30).

[77] Declaration of Geneva (n. 1).

[78] Joakim Andersson, 'Läkare i stort upprop – vill se åtgärder mot rasism i vården' (2021) Läkartidningense.

[79] Anke Samulowitz and others, '"Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain' (2018) 2018 Pain Research and Management 6358624.

[80] Sarah Hamed and others, 'Racism in European Health Care: Structural Violence and Beyond' (2020) 30 Qualitative Health Research 1662.

[81] E.g., Wiegand, T. et al. (ITU), *Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health* (The International Telecommunication Union 2018) 3–4.

natory practices could be reproduced due to AI-systems learning from history (status quo bias). For example, if a system is trained on a set of previously given treatments, or costs of previous treatments, groups that have a history of more easily receiving treatment could be incorrectly classified as higher risk patients.[82] This exemplifies the danger of treating covariances as explanations.

Besides equal care, The Swedish Health and Medical Service Act states that the person *most in need* should be prioritized,[83] which is challenged on the "quasi-market" of online doctors.[84] In theory, AI can help to find an answer to the conundrum of distinguishing who is most in need, by the ability to analyse large sets of data in a short amount of time and taking more variables into account. Well-trained algorithms could provide better predictions, finding previously unknown patterns or risks, for example in x-ray screenings or by improving prioritizing during triage by more accurate predictions of risk of re-admission or even death. However, the principle of most in need entails the necessity to be able to motivate decisions within health care – why one person is prioritized or not (for example during a triage process).

Further, the Swedish Patient Safety Act states that medical care should be conducted in *accordance with science and proven experience*.[85] Also, the MDR demands that evidence for clinical performance is provided.[86] A challenge of these principles is that it is ill-defined what should constitute the determinants of accuracy. AI models can also be working well for the large majority, but be less sensitive for identifying outliers and atypical symptoms or rare diagnoses, proposing the risk of patients being discriminated or mistreated, even when models on paper reach a standard of accuracy.

Transparency and explainability of implemented AI-systems are needed to be able to assess *fairness*, the equality of care, and that the persons most in need are in fact given priority and are treated in accordance with science and proven experience. Without proper information, these factors cannot be assessed.

[82] Obermeyer and others, 'Dissecting racial bias in an algorithm used to manage the health of populations' (n. 20).

[83] Hälso- och sjukvårdslag (SFS 2017:30), 3.1.

[84] Peter Bergwall, *Exploring Paths of Justice in the Digital Healthcare: A Socio-Legal Study of Swedish Online Doctors*, 51 (Faculty of Social Sciences, Lund University, 2021).

[85] Patientsäkerhetslag (SFS 2010:659), 6.1.

[86] MDR (n. 6).

3.2 The right to privacy and integrity

A motor in the development of AI models is the access to large amounts of reliable, accurate and representative data in order to train models and test their validity. A challenge for health care and medicine is the sensitive nature of the data needed.

A person's right to privacy is declared in the Universal Declaration of Human Rights, in the European Convention for Human Rights and in the EU Charter.[87] The right to respect for private life is also emphasized in medical ethics. The Declaration of Geneva states "I WILL RESPECT the secrets that are confided in me, even after the patient has died."[88] In recent time, important legal advances to strengthen individuals' right to privacy and control of personal information have been enforced in the form of the GDPR. According to Art. 9 GDPR, health data is a sensitive category of personal data, together with biometric and genetic data (when used for purpose of identification). In principle the processing of health data is prohibited unless there is an applicable exception that is listed in Art. 9.2 GDPR. The most relevant exceptions for processing sensitive data in health care are:

- after explicit consent (Art. 9.2(a)),
- if necessary for the protection of vital interests of a data subject incapable of giving consent (Art. 9.2(c)), for example in the case of an unconscious patient that needs treatment,
- if necessary for the assessment of a medical diagnosis, provision of health care or management of health care systems and services, if the data are processed by an actor legally bound by professional secrecy and confidentiality (Art. 9.2(h)),
- if necessary for reasons of public interest in the area of public health (Art. 9.2(i)) or for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes under certain provisions (Art. 9.2(j)).[89]

[87] Universal Declaration of Human Rights, Art. 12 (no arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation), European Convention for Human Rights, Art. 8 (Respect for private and family life), and the Charter of Fundamental Rights of the EU (2000/C 364/01), Art. 7 (Respect for private and family life) and Art. 8 (Protection of personal data).

[88] Declaration of Geneva (n. 1).

[89] GDPR, Art. 9, Art. 9.2(a), Art. 9.2(c), Art. 9.2(h), (n. 5).

The last exception represents a wider interest than that of the individual patient, and is therefore also of interest in relation to the public-private complexity accounted for in section 1.2. How should the line be drawn for the public interest exception when publicly held patient data are used for the training of private companies proprietary AI-systems? How can the public interest be ensured when the data is utilised by private interests, albeit for systems used in public health care?

According to the Swedish Public Information and Secrecy Act, health professionals have a legal obligation of confidentiality regarding patients' health conditions and other personal information, if the information cannot be shared without the individual or their relatives suffering.[90] Also, The Patient Safety Act states that a person working (past or presently) within health care is not allowed to share any information regarding an individual's health or other personal conditions, obtained in course of the work.[91] The Swedish Patient Act and Patient Data Act state that personal information should be registered and further processed with respect to the integrity of patients and others.[92]

The purpose for which sensitive information can be processed within health care is broad. The Swedish Act of complementary provisions to the GDPR, states that processing of sensitive personal information in health care is allowed, if necessary, for reasons such as preventative health care and medicine, medical diagnosis, providing health care or treatment, administration of health services and systems.[93] Further, processing of sensitive information for statistical use is permitted when benefits clearly outweigh potential risks for the privacy of individuals.[94] The Nordic countries' national registers could function as "goldmines" of health data for training algorithms. According to the Swedish Patient Data Act, processing of personal (health) information is permitted for national and regional registers, if consent is given.[95] Data from different registers are also combined to perform research and improve health care and medical knowledge. To facilitate longitudal studies, data need to be able to tie to the same individual to follow how health evolves over time. This is also

[90] Offentlighets- och sekretesslag (SFS 2009:400), ch. 25.1.
[91] Patientsäkerhetslag (SFS 2010:659), ch. 6.12.
[92] Patientlag (SFS 2014:821), ch. 10 and Patientdatalag (SFS 2008:355), ch. 1.2.
[93] Lag (2018:218) med kompletterande bestämmelser till EU:s dataskyddsförordning, 3.5.
[94] Lag (2018:218) med kompletterande bestämmelser till EU:s dataskyddsförordning, 3.7.
[95] Patientdatalag (SFS 2008:355), ch. 7.

true for research on for example how social, economic and demographic factors affect health.

Access to large sets of health data is crucial to develop accurate and fair AI models.[96] By the use of natural language processing, patient journals can also be important information sources by which vital knowledge could be gained. However, one problematic aspect of using patient journals is that less structured and standardized data provide challenges in the control of what could be revealed in the process of data sharing and learning of algorithms.

Here we have identified an important double bind: while big sets of health data constitute a necessity for AI development, data sharing – for example between public entities and commercial product developers – can pose challenges. Even though stripped from personal identifications, there is a risk of back door identification by reconstruction of aggregated data and combining of data sources. The principles of privacy and doctor's confidentiality could be contested by the data hunger of AI development and the increasing diffusion of data flows.

3.3 The right to information

The High-Level Expert Group on AI states that explanations should be timely and adapted to the level of expertise of the receiver.[97] This is also in line with the demands of the Swedish Patient Act,[98] which specifies that the caregiver must provide the patient with information regarding, for example, their health condition, methods for examination, expected course of treatment, any risks for complications and side-effects and methods to prevent diseases or injuries. The same act stipulates that information needs be tailored to the receiver's age, maturity, language background, and other individual preconditions, and that the one providing the information should make sure, as far as possible, that the content and significance of it has been understood.[99] Complaints by patients should also be answered with the receiver's ability to obtain the information in mind, and the health provider is obliged to provide an *explanation* of the

[96] Wiegand, T. et al (ITU) *Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health* (n. 81) 3.
[97] HLEG 2019 (n. 35).
[98] Patientlag (SFS 2014:821), 3.1.
[99] Patientlag (SFS 2014:821), 3.6, 3.7.

course of events, and describe actions planned for a similar event not to occur again, as stated by the Patient Safety Act.[100]

In recital 43 of the Medical Device Regulation, transparency and access to information are emphasized as "essential in the public interest, to protect public health, to empower patients and health care professionals and to enable them to make informed decisions, to provide a sound basis for regulatory decision-making and to build confidence in the regulatory system."[101] It also stipulates that information should be "appropriately presented for the intended user."[102] In addition, the GDPR states that data subjects (in this context: the patients) have the right to access information on data treatment in a "concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child."[103] It also stipulates a requirement for ex ante notification that should contain information about the purpose of the processing, how long the data will be kept and by whom the data will be processed. In addition, data subjects have the right to access information that the data controller holds about them: what categories of information, as well as copies of data, purpose of processing and with whom it is shared.[104]

In general, re-use of data for another purpose is prohibited, unless the re-use is for a purpose that is compatible with the initial purpose, such as research or statistical analysis,[105] or if a data subject consents ("downstream consent") to further processing for a new, incompatible, purpose.[106] In either case of further processing, data subjects should be notified. However, if providing a notification directly to data subjects is considered impossible or a disproportionate effort, especially for uses for scientific or statistical purposes, there is the option of making information publicly available instead, providing a basis for the opt-out principle for research studies and register data use[107] The GDPR is also demand-

[100] Patientsäkerhetslag (SFS 2010:659). 3.8(b).
[101] MDR, Recital 43 (n. 6).
[102] MDR, Recital 43 (n. 6).
[103] GDPR, Art. 12 (n. 5).
[104] GDPR, Art. 15 (n. 5).
[105] GDPR, Art. 5.1(b) and GDPR Art. 6.4 (n. 5).
[106] GDPR Art. 6.1(a) and GDPR Art. 6.4 GDPR. (n. 5), Regarding downstream consent, see Article 29 Working Party *Opinion 15/2011 on the definition of consent*, 2011, WP 187, p. 19.
[107] GDPR, Art. 14.5 and Art. 89.1 (research exception) (n. 5).

ing notification of the existence of solely automated decision-making, including profiling, with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[108] In addition, Recital 71 states that in the case of automated decision-making or profiling, data subjects should have the right to: "obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."[109]

Legally, health care providers are obliged to provide information in a way that the patient can understand. This could present a great challenge even without opaque AI systems. Conversational AI could provide solutions and improvements regarding this, but also pose difficulties as the sensitive nature of information given could require empathic skills and experience. Regardless, as previously discussed, health professionals do have information obligations in contact with the patient, to fulfil the requirements stipulated in the Swedish health care legislation. The versatility in the need for information means that AI-supported systems in health care have to be explainable with diverse levels of specificity, in different stages of implementation, to be intelligible by different actors or audiences.[110] It could mean initially providing explanations suitable for the developers of the system themselves, then for health professionals and people responsible for certification and procurement, and further challenging, to all relevant patients. Medical ethics and legal framework call for meaningful information, by contextual transparency and explainability, for caregivers to be able to fulfil their obligations and cater to patients' rights to information.

3.4 The right to dignity and autonomy

A main principle of medical ethics is the autonomy of the patient, which in both the normative and legal frameworks is tied to the dignity of human beings. The Declaration of Geneva states "I WILL RESPECT the

[108] GDPR, Art. 13.2(f) (n. 5), which refers to Article 22, that also adds the provision that the decision-making has to have "legal effect". See also Section 3.4.

[109] GDPR, Recital 71 (n. 5).

[110] Patientlag (SFS 2014:821), 3.6, 3.7. Also see e.g., Barredo Arrieta and others, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI' (n. 62).

autonomy and dignity of my patient."[111] Part of this principle is that no treatment should be performed without informed consent, if not impossible to obtain due to a medical condition. Health care should as far as possible be planned and conducted in consultation with the patient, with consideration and respect, as stipulated by The Swedish Patient Safety Act.[112] The patient's autonomy and integrity ought to be respected and before consent, proper information must be provided, according to the Swedish Patient Act. When there are multiple treatment options (in line with science and proven experience), the patient should be able to make an informed choice.[113]

The principle of autonomy and choice is highly intertwined with the previously discussed principle of right to information, since real autonomy and choice could hardly be achieved if the patient has not had access to underlying information, and hence no possibility to interpret it. The idea of informed consent and individuals' control over their personal information is a main part of the GDPR. Article 22 regulates decision-making based on *solely* automatic processing, giving individuals the right not to be subject to such processing, requiring consent.[114] This regards automated processing for decisions that have a *legal effect* (or similar significant effect – which is to be interpreted as including all medical decisions). However, it is not established how the word *solely* should be understood in the context of health care.[115] If AI-systems analyse and prepare decisions, which are merely confirmed by a human doctor, should this be considered solely automatic processing ?[116]

Another aspect of the dignity and autonomy of patients, with regards to AI use, is in the situation of communication and information exchange. Is it a prerequisite for the dignity of a patient to have a human present, to talk to and provide information in an empathic manner? A hope for AI in health care is that the automation of certain tasks will free time for health professionals to be able to increase (or at least not reduce) the time spent with patients and on patient-close care. If realized, this is an opportunity to increase dignity in patient care and also autonomy,

[111] Declaration of Geneva (n. 1).
[112] Patientsäkerhetslag (2010:659), 6.1.
[113] Patientlag (SFS 2014:821), ch. 4, 5 and 7.
[114] GDPR, Art. 22 (n. 5).
[115] Hoeren and Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (n. 69).
[116] Ibid. (n. 69).

if additional time can be allocated to information exchanges with patients and gaining knowledge of patients' preferences. However, parts of this could be (and are already) subject to automatization. Chatbots on caregivers' websites can increase access. Some people could also prefer talking to a bot – or robot[117] – rather than a human-being on topics of sensitive nature, possibly lowering thresholds.[118] While having potential benefits, automation of information tasks in health care could contest the principles of dignity and self-determination of patients.

Apart from the legal requirement in Art. 22 GDPR that decisions should not be based *solely* on automated processing, current legal and normative frameworks do not yet specify the role of the human-in-the-loop, and the question is whether the patient has a right to a human doctor.[119] Patients rely on clinicians being able to convey explanations in an accurate and understandable manner, improving the patient's agency in terms of risk assessment and informed choice.[120] If AI is not explainable, it could pose a challenge for health professionals to provide enough information on the reasons for classifications and proposed treatments, for patients to exercise their right to autonomy. Further, the personalisation of medicine could enhance autonomy but also lessen the experienced control of individuals. It could also be considered intrusive in practice, depending on the development and information and room for action provided to patients.

[117] See for example Maria Kyrarini and others, 'A Survey of Robots in Healthcare' (2021) 9 Technologies 8, and Laetitia Tanqueray, Tobiaz Paulsson, Mengyu Zhong, Stefan Larsson and Ginevra Castellano, 'Gender Fairness in Social Robotics: Exploring a Future Care of Peripartum Depression' In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Association for Computing Machinery, ACM, 2022).

[118] E.g., Bergwall, *Exploring Paths of Justice in the Digital Healthcare: A Socio-Legal Study of Swedish Online Doctors* (n. 84).

[119] For discussions regarding this, see for example Hoeren and Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (n. 69), Fabrice Jotterand and Clara Bosco, 'Keeping the "Human in the Loop" in the Age of Artificial Intelligence' (2020) 26 Science and Engineering Ethics 2455 and Therese Enarsson, Lena Enqvist and Markus Naarttijärvi, 'Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts' (2021) Information & Communications Technology Law 1.

[120] As pointed out by Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' (n. 52) and Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

4 Discussion

The development of AI-supported tools for medicine and health care is rapidly evolving, and their broad adoption is imagined for the near future. However, the social, legal and ethical implications of AI use and automated decision-making/automated decision-support, could present severe challenges for successful and responsible implementation. In this chapter, we provide a brief overview of the novelties that AI in health care bring about in comparison to previous technologies, in order to point to key aspects of what this entails for current legal and normative (medical ethical) principles, especially with regards to transparency and explainability.

While acknowledging what could be gained by the adoption of AI, we must also consider what could be disrupted. This urges us to look for frictions between the basic principles of the normative and legal frameworks of health care and the implementation of AI. As stated by Obermeyer and Emanuel: "this challenge will create winners and losers in medicine. But we are optimistic that patients, whose lives and medical histories shape the algorithms, will emerge as the biggest winners as machine learning transforms clinical medicine."[121] If their optimism is to be realized, the rights of patients cannot be overlooked. Equality of care, privacy, access to information, autonomy and dignity are basic principles of the legal framework of patients' rights. When developing and implementing AI-systems and methodologies to be used in the context of health care, there is indeed a need to jointly address how they could be compliant with these basic principles, and hopefully even support and strengthen them.

We argue that transparency needs to be understood as *situated* in the *information practices* of health care, in line with Lee's notion of algorithms in practice,[122] and not as a binary state of full transparency or opaqueness. Data flows in health care are based on medical ethical ideals and are two-faced; both carefully protecting and carefully providing information, between patients and the healthcare systems, as well as developers, registers, and other actors and infrastructures in public and private sector. This becomes evident in the right to privacy and the right to information,

[121] Obermeyer and Emanuel, 'Predicting the Future - Big Data, Machine Learning, and Clinical Medicine' 1218 (n. 20).

[122] Lee, 'Enacting the Pandemic: Analyzing Agency, Opacity, and Power in Algorithmic Assemblages' (n. 45).

which also constitute a foundation for the right to autonomy and dignity. The data flows represent both opportunity and risk, further emphasized by the *private-public complexity*. As stated above, this may be detrimental to transparency-requirements in assessment of whether the specific AI-tools or services actually work in a fair and trustworthy way, given proprietary interests or the need for keeping business secrets on competitive markets. This also stresses another balancing question of principal interest. On the one hand, it is feasible that market-driven players may indeed be best suited for developing new AI-systems to be utilised as products or services procured by the public sector. But, on the other hand, if that development is dependent on data collected in the public sector, from its patients, protected by the GDPR but shared in the name of the public interest, there may indeed also be a need for more thought on how to ensure that these applications actually serve the public interest, while under clear private interest custody.

Furthermore, if we are to benefit the most of AI in medicine, it is not to be used in a one-size-fits-all-manner, but to make the best decision out of all available information about the individual patient. The level of risk, or best treatment option, may depend on aspects such as age, gender or ethnic origin. Not taking these factors into account when applicable, could lead to discriminatory results by ignoring known (or unknown) risk factors and posing disadvantages to vulnerable groups. This *personalisation* could enhance, as well as contest, values of equal care and most in need, while at the same time pose risks for discriminatory bias of systems, privacy breaches and weakened autonomy, especially pushed by *automation* and *large-scale application*.

Transparency and explainability constitute prerequisites for assessment of patient's rights, demanding fairness and accountability. However, this leads to vital questions connected to the theoretical foundation of what type of transparency is to be aspired to, when, and to whom. Only asking for general transparency and explainability might not be meaningful unless further specified, as pointed out by de Vries.[123] Explainability could function as a tool for patient autonomy as well as sound scepticism and scrutiny, mending over-reliance on algorithms and algorithmic bias. If AI-systems are not sufficiently explainable, they could end up not being used by health professionals due to lack of trust, maybe depriving the

[123] de Vries, 'Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM' (n. 44).

person most in need of being prioritized, i.e., affecting the most in need principle (discussed in section 3.1). Transparency and explainability are necessary tools to assess how well AI-systems perform and align with scientific knowledge in the specific context of its use, calling also for *causability*, as Holzinger et al. suggest.[124] If the instrumental goal of transparency is increased knowledge – for health professionals, patients, and the public – information and explanations need to be adjusted accordingly, and be specific enough to be meaningful in individual cases. This could promote the goals of fostering agency, accountability and fairness, and in the long-run; trust. For the patient wanting to know the reason for being sent home from the ER, one cannot refer merely to an algorithm suggesting that it was the right call to make. Transparency needs to be *contextually* understood and, as Miller and Mittelstadt et al. point out, explanations are executed as social, selective and contrastive functions[125] – what is the smallest difference that would have resulted in a different decision, that is, that the patient is *not* sent home from the ER?

Still, one could ask, what should lead the way forward – best possible care or most transparency? While this is perhaps in some cases a false dichotomy,[126] the concept of best possible care could also be argued to not consist only of the physical health outcome. It also encompasses the adherence to ethical core values of how health care should be conducted.[127] There is also an epistemic aspect: how would we know if it is the best possible care, or even an accurate improvement, unless it could be assessed by sufficient transparency? Medical ethics and the legal framework do not allow for AI-tools not being transparent and explainable *in meaningful ways*, with careful consideration of the needs of different addressees. Such opacity would hinder the possibilities for health practitioners to interpret and assess a recommendation, decision or classification, and by that limit the possibilities for patients to get the information they are entitled to, to

[124] Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (n. 29).

[125] Miller, 'Explanation in artificial intelligence: Insights from the social sciences', (n. 59), Mittelstadt, Russell and Wachter, 'Explaining Explanations in AI' (n. 33).

[126] See the discussion on possible trade-off between accuracy and explainability in section 2.2.

[127] In line with the reasoning of Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' (n 52), Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24) and Hoeren and Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (n. 69).

make a correct risk assessment. AI-systems cannot be allowed to function as *strategic unknowns*,[128] neither in the form of decision-support nor in automated decision-making, in the context of health care. In the implementation of new technology, health professionals need to be given a chance to live up to medical ideals and regulatory requirements, because, no matter how accurate AI-technologies ever get, and how well they ever learn to imitate human compassion, they will never feel the burden of the Hippocratic oath.

5 Conclusions

In this chapter, we ask what role transparency and explainability of AI could have, in relation to patients' rights and information flows in Swedish health care. We outlined a set of novelty characteristics associated with AI-systems in health, including: automation, scale, opacity, data-dependency, obscured causality, personalisation and a private-public complexity. By first setting a foundation of a conceptual framework on transparency in general and explainability in particular, in relation to AI in health care, we analyse the legal and normative regulatory framework of patients' rights. We address those rights that are most relevant to transparency and explainability in relation to AI; the right to equal health care, privacy, information, dignity and autonomy, with the purpose of pinpointing main challenges in the implementation of AI in, primarily public, health care.

We find that it is not possible to adhere to the basic principles of the normative and legal frameworks of Swedish health care, if meaningful and contextual transparency and explainability are not deployed in the implementation of AI. Instead of focusing the highest quality of health care as something which stands on opposite side of the requirement for transparency (by the accuracy versus explainability trade-off), we argue for the need to consider them as interdependent. The best possible health care cannot be achieved without transparency. As transparency is situated in information practices, and not a binary state, the way forward is to find what kind of transparency that is needed to safeguard best possible health care. Meaningful and contextual transparency and explainability are necessary for the provision of patient autonomy and as a means to assess if the best possible care is given to the ones most in need.

[128] Bucher, *If…then: algorithmic power and politics* (n. 21). See section 2.2.