Authors name: Kristina Astromskė, dr. Eimantas Peičius, dr. Paulius Astromskis

Title of paper: Ethical and Legal Challenges of Informed Consent Applying Artificial Intelligence in Medical Diagnostic Consultations

Affiliations: Lithuanian University of Health Sciences (Lithuania), Vytautas Magnus University (Lithuania).

Address: A. Mickevičiaus g. 9, LT 44307 Kaunas (Lithuania), K. Donelaičio g. 58, LT-44248 Kaunas (Lithuania)

E-mail: Kristina.Astromske@lsmuni.lt, Eimantas.Peicius@lsmuni.lt, paulius@astromskis.lt;
Ethical and Legal Challenges of Informed Consent Applying Artificial Intelligence in Medical Diagnostic Consultations

Abstract
This paper inquires into the complex issue of informed consent applying artificial intelligence in medical diagnostic consultations. The aim is to expose the main ethical and legal concerns of the New Health phenomenon, powered by intelligent machines. To achieve this objective, the first part of the paper analyzes ethical aspects of the alleged right to explanation, privacy, and informed consent, applying artificial intelligence in medical diagnostic consultations. This analysis is followed by a legal analysis of the limits and requirements for the explainability of artificial intelligence. Followed by this analysis, recommendations for action are given in the concluding remarks of the paper.

Keywords
Medical diagnostics, informed consent, opacity, trust, artificial intelligence, medical ethics, right to explanation

1. Introduction
More than 50 years ago, Isaac Asimov predicted, with the punctilio of accuracy, that robots are neither common, nor very good in 2014, but they are in existence (Asimov 1964). Indeed, artificial intelligence and robotics are not science fiction anymore as they are present in households and workplaces throughout the world. Technologies fundamentally alter the healthcare sector too, exponentially enhancing physicians’ abilities to deliver better and faster results more than ever before. Explosion in the amount of healthcare data, information technology development, democratization of access for healthcare, and willingness of the general public to be more active participants in their own health are identified as major trends that transform and define the “New Health” phenomenon, where robotics and artificial intelligence increasingly become a part of our healthcare ecosystem (PwC 2017).

The converging trends in artificial intelligence and robotics developments used to define the abovementioned “New Health” transformations, are visible everywhere – from surgery, drug discovery, patient care, and information management, to diagnostics and beyond (Feldman et al. 2019; PwC 2017). Artificial intelligence systems, accompanied by increasing computational powers, can analyze personal characteristics, medical records, large amounts of literature, and other medical data with high-accuracy and using just a fraction of time and cost required for the same task to be completed by a human counterpart. These intelligent machines are accompanied with high-incentives to support diagnostic decisions of the physicians, since decreased time, cost, and medical errors increase access and quality of healthcare (Sion 2019). It is predicted that eventually machines will have demonstrably better success rates than human physicians in providing the most accurate and higher-quality diagnoses or in making the best decisions when choosing the most beneficial therapy (Bachle 2019; Froomkin et al. 2019). Even though artificial intelligence tools bear an inherent risk of making incorrect determinations, they make fewer mistakes than humans, thus, reducing the number of medical errors (Bachle 2019; Meskó, Hetényi, and Györrfy 2018). Indeed, machines don’t get tired, don’t allow emotion to influence their judgment, make decisions faster and can be programmed to learn more readily than humans (AoRMC 2019).

Historically, one of the most common obstacles of precise and clear diagnosis (along with visually observed symptoms) was the lack of medical data. Nowadays, this gap has been immensely compensated by the usage of a plethora of powerful and data-driven devices. For instance, the Watson Oncology Advisor (IBM) uses almost 15 million pages of medical literature to advise oncologists on cancer diagnoses and chemotherapy plans, with an accuracy of 90% (Hoerenand Niehoff 2018). Additionally, artificial intelligence enables a review and translation of mammograms 30-times faster with 99% accuracy or correctly detected 92.4% of breast cancer tumors compared to the 73.2% detected correctly by human doctors. The recent study of McKinney et al (2020) revealed that the artificial intelligence system applied for breast cancer screening has outperformed all of the human readers: the area under the receiver operating characteristic curve (AUC-ROC) for the artificial intelligence system was greater than the AUC-ROC for average radiologist by an absolute margin of 11.5%; the artificial intelligence system reduced the workload of the second reader by 88% and resulted in an absolute reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives. Thus, AI could significantly reduce the need for unnecessary biopsies as well as decrease the uncertainty and stress of a misdiagnosis (Griffiths 2016; Liu et al. 2017; PwC 2017). IBM’s Watson for Health is working on cognitive technology that can process medical information exponentially faster than any human, providing findings and decisions free of cognitive biases or overconfidence, thus reducing misdiagnosis (PwC 2017). Google DeepMind learning algorithms can interpret quickly eye scans from routine clinical practice with unprecedented accuracy and recommend correctly where patients should be referred to for treatment of over 50 sight-threatening eye diseases as accurately as any world-leading expert doctor (De Fauw et al. 2018; Gulshan et al. 2016). Wearables like Cyrcadia’s iTBra for the detection of breast cancer or Cardio Diagnostics for cardiac monitoring...
and rhythm management (PwC 2017), chatbots like Wysa or Woebot for depression, and other mental health issues detection and treatment (Stiefel 2019) and many other technological developments in radiology, ophthalmology, dermatology, pathology, and other medical specialties are already saving lives, disrupts early diagnostic practices, and attracting the attention of policymakers and investors (see for e.g., Konnangath 2019; Liang et al. 2019; Yu et al, 2018; and others).

Although the goal with which artificial intelligence systems are applied in medicine is to assist humans to execute their tasks more efficiently, medical ethics has begun to highlight certain concerns regarding algorithmic bias, transparency, dehumanization of physician-patient relationships, decline of practical physician skills, and others (see for e.g., Char et al. 2018; Mittelstadt et al. 2016; Obermeyer and Emanuel 2016; Schiff, 2019). Schönberger (2019) has identified biased training data, inconclusive correlations, intelligibility, inaccuracy, unfair outcomes and other key concerns related to decision-making capacity of artificial intelligent systems. Indeed, machines could be poorly programmed, poorly trained, used inappropriately, contain incomplete or biased data, and could be misled or hacked (AoRMC 2019). However, analysis of Ienca et al. (2018) revealed that privacy and confidentiality are by far the dominant concern in the ethical domain, followed by informed consent, but the important issues of fairness, discrimination, trust, transparency, responsibility, and others compose a relatively small portion of the spectrum current ethical and legal issues discussed in the literature.

Thus, together with the “New Health” concept, new types of unprecedented risks emerge which need to be considered. Due to data dependency, there is a risk that patients may never be able to function holistically but as mere bearers of medically (ir)relevant data. This may lead to an increased de-personalization of individual patients by reducing the quantity and quality of human contact in doctor-patient relationships. It also entails much less visible consequences, such as a slow change in our understanding of what it means to be healthy and the general tendency to over-diagnose or over-treat (Bachle 2019). Therefore, informed consent requires a debate on the core purposes, content design, and form of it in the situations when intelligent machines are used in diagnostics or other interventions into health (Wolf et al., 2018).

Moreover, throughout the history of humanity, decision making was a fundamental human process, for assistance of which conventional tools were created. Never before humans have built machines that would be able to mimic a human decision process in ways their creators sometimes do not understand and cannot explain, nor have the machines been created that would deserve the call to explore the possibility of applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently (European Parliament 2017). This also raises the question whether existing legal systems and rules, originally intended and designed for human-to-human (in personam) and human-to-machine (in rem) processes, can work well in machine-to-human and machine-to-machine environments (Fomin 2018). Due to the novelty and challenges of the application of technology in healthcare, the emerging intelligent machines may lack normative rule sets regarding how they can or should be used, or what their use means.

Even considering that current laws and ethical concepts are largely suited to deal with majority of concerns, particularly related to bias, opacity and failure to model the real world accurately, some prompt clarifications are desirable, especially in the fields of responsibility, fairness and others (Schönberger 2019). Once artificial intelligence-based medical diagnostics are shown to be superior to humans, the risk of over-reliance emerges, requiring a revision of existing medical malpractice law, the standard of care, and liability rules in the clinical settings (Froomkin et al. 2019). However, over-regulation could arbitrarily diminish the value of artificial intelligence systems in healthcare. More control means less freedom and hence over-protectionism might unnecessarily hinder welfare-enhancing innovations. Indeed, as artificial intelligence is still a nascent field in healthcare, hard and premature regulation could stifle many useful innovations that could be beneficial for patients and health system (Schönberger 2019). On the other hand, liberal and innovation-fostering regulation must not lead to a loss of trust between patient and physician or violations of fundamental human rights. Therefore, the main challenge for regulators is to balance freedom and control in a way that would maximize the personal and overall well-being of society, thus promoting the benefits and inhibiting the harmful effects (den Hertog 2010).

Little ethical and legal work related to artificial intelligence in healthcare exists so far (Schönberger 2019), although the body of literature in this field is diverse and rapidly-growing (Ienca et al. 2018). Still, there is a lack of empirical studies to reveal if patients directly benefit from the use of artificial intelligence systems. Although deep learning will not be a panacea, it has huge potential in many clinical areas where datasets are potentially stable over extended periods and high dimensional data is mapped to a simple classification. As such, it will be incumbent on healthcare professionals to become more familiar with this and other artificial intelligence technologies in the coming years to ensure that they are used appropriately (Keane and Topol 2018).

Accordingly, this article contributes to the ongoing scientific discussion on the risks and benefits of artificial intelligence use in healthcare and aims to expose contemporary legal and ethical challenges of informed consent applying artificial intelligence in medical diagnostic consultations. As a first step, the article exposes ethical aspects of the right to explanation, privacy, and informed consent, applying artificial intelligence in medical diagnostic consultations, followed with a legal
analysis of the limits and requirements for it. This analysis is accomplished with conceptual recommendations for action in the concluding part of this article.

2. Ethical Challenges of Applying Artificial Intelligence in Medical Diagnostic Consultations

In general, any type of new technology should not mean new values (European Commission 2018) and any technologies should be regarded as value neutral. Yet, the application of so-called advanced health innovations has been continuously followed by a number of ethical issues while on its way into medical practice. One of the most fundamental questions here is how these radical technological transformations might impact the normative physician and patient relationship in terms of respecting patient’s autonomy, ensuring patient’s informed consent, as well as keeping up with patient’s welfare and primum non nocere principles (Miller 1998)?

In particular, supported by very powerful and smart devices such as big data machines, physicians would obviously be eager to use them in diagnostics, but sooner or later would come across ethical dilemmas - how detailed and comprehensive should the diagnostic information to a patient be and how it should be commented or explained, if any at all. From the perspective of the moral and legislation background of European Union, ideally, “AI systems should function according to values that are aligned to those of humans, so that they are accepted by our societies and by the environment in which they are intended to function” (Rossi 2016). However, if one of the most important values is respect for person including personal autonomy and informed consent, then it imposes the question - what the expected impact of any AI applications is in diagnostic consultations, and how AI assistance may affect physician’s considerations about the best patient interests.

Informed consent is, therefore, an important ethical concern of the “New Health” phenomenon. The concept of informed consent, beside other meanings, is assumed to be an instrument that enables the patients to make their own health related risk and benefit assessments. Patient’s right not to know fearful, stressful, or frustrating information. Hence, despite potentially significant benefits of artificial intelligence application, some moral dilemmas may occur in physician and patient relationships during regular consultations while trying to reveal the causes of any health disorders by using artificial intelligence-based devices. One way to increase patient’s competence and to establish the ethical relationship between a doctor and patient is to reduce opacity behind artificial intelligence. Currently, artificial intelligence systems are portrayed as “black boxes” (Pasquale 2015) in a sense that neither creators nor users of these systems have an explainable understanding of the AI phenomenon (Ferretti et al. 2018; Schönberger, 2019). That is, while inputs and outputs are visible, the internal process of reaching the output from the input remains opaque (Feldman et al. 2019). It also precludes to figure out technical vulnerabilities in order control the diagnostic process. Without sufficient transparency neither patients nor physicians can understand how and why intelligent machines generate the diagnosis. This complicates the interpretation of artificial intelligence-based assessments and reduces the trustworthiness in the overall diagnostic process. Tools behind the AI diagnostics, such as computation of confidence and probability scores, require an understanding of bioinformatics. Interpretation behind these tools may require additional training on the physicians’ part which may not be easy to implement due to big workload on the healthcare industry. Therefore, the “black box” problem of artificial intelligence remains some of the greatest challenges for medical ethicists and regulators (Vayena et al. 2018).

The difficulty of humans to understand and explain how these systems work was defined as ‘opacity’ of artificial intelligence systems (Lipton 2016; Schönberger 2019). Ferretti et al. (2018) observed that there are three different semantic dimensions of opacity particular relevant to medicine: lack of disclosure, epistemic opacity, and explanatory opacity. Lack of disclosure refers to the fact that data subjects are unaware that automated decision-making activities about them are being carried out. This type of opacity does not depend on intrinsic technical characteristics of artificial intelligence systems but prevents data subjects from exercising some specific data-related rights and undermine the fiduciary relationship between patients and physicians. Epistemic opacity occurs when it is not possible to have access to information or there is insufficient understanding. This is either due to procedural darkness or procedural ignorance, of the rules an artificial intelligence system is applying to make predictions, classifications, and decisions. Explanatory opacity relates to the question of why an artificial intelligence system provides a specific outcome. Specifically, the artificial intelligence system may look at properties that human physicians do not use, and whose connection to a given clinical phenotype is not known. Although physicians could say that there is a statistically relevant correlation between the property and the clinical classification, one may not have a scientific explanation of the reason why the property and the classification are linked to one another (Ferreti et al. 2018).

Obviously, opacity of artificial intelligence system depends on the initial design of the model and on responsible, diversified and ethical engineering practices of its developers. Moreover, for some minor diagnostic procedures, overcoming opacity might be neither relevant, nor required, nor linked with the ethical relationship or trustworthy dialog between a doctor and
patient. However, for some decision makers, particularly those in life or death situations where confidence is critical, the benefits of artificial intelligence can be out of reach without interpretability (Frey, 2019). And of course, as explained further, an explanation does not require knowing the flow of bits through an artificial intelligence system, no more than an explanation from humans requires knowing the flow of signals through human brain neurons. However, the problem of algorithmic opacity is not only a technical one, since it leads to another major ethical concern of the “New Health” phenomenon – trust. Creating, preserving, and enhancing trust is understood to be one of the fundamental goals of medical ethics, health care law, as well as public policy (Feldman et al. 2019; Mechanic 1996, 1998). As Feldman (2019) eloquently exposes - how can we trust our health, let alone our very lives, to decisions whose pathways are unknown and impenetrable? Indeed, without established trust, a patient may have little or no incentive to seek the advice of a physician or share sensitive clinical information, which is required by the artificial intelligence algorithms for diagnostic purposes (Dugdale and Epstein 1999; Kullar 2018; Thom et al. 2004).

The lack of data, technological literacy, and consistent standards to guide the use of artificial intelligence and other emerging technologies in healthcare create notable gaps in the calculations of risks and benefits of diagnostic procedures (Nebeker et al. 2019). For instance, as Nebeker et al. (2019) noted, if regulatory bodies or physicians (advised by such algorithms) may not have enough experience or knowledge needed to assess the probability or level of potential harm, this may lead to a lack of awareness regarding privacy concerns, risks assessment, usability, and societal impact. Then, building the trust in technological advancements should start with interdisciplinary dialogue between the developers and physicians in the process of building and clinically validating these systems in prospective studies and randomized clinical trials and, thus, additional technological training of sufficient skills to grasp at least fundamental principles and logic how these sophisticated systems work might be required (Vayena et al. 2018). The task for the creators of artificial intelligence systems is, therefore, to embed the ethical behavioral patterns that require the high level of ethicist involvement in the creation and training processes (Lupton 2018). A carefully set up validation environment should expose the general risks associated with the use of sophisticated technologies, but also set the meaningful explainability framework to be used in physician–patient dialogue. Despite high achievements in medicine, medical diagnostics is still rather the matter of assumption and interpretation of physician than the result of a deductive or observational process, for example in the cases of the medically unexplained disorder (Olde-Hartman 2009). However, building the elements of trustworthy behavior into the systems may be a complicated task that may slow down the development process and may restrict the functionality of the system. These “side-effects” of ethical development may contradict directly the aims of opportunistic and quick or short-term profit-seeking private stakeholders. Hence, some of developers may voluntarily pursue ethical goals following the socially responsible business morale; however, for others the legal pressure through imperative ethical standards is needed.

Moreover, the trustworthiness of artificial intelligence systems is dependent on the amount and quality of the entered data. On one side of this data issue, privacy concerns might emerge.. Without access to sensitive medical records, it is impossible to optimize the algorithms to their precision. But the principles of medical confidentiality and respect for patients’ privacy require the explicit consent by the patient in order to access to his/her health records. Refusal to give consent by some patients may create a situation, where some of the beneficial record-based research is prevented from being used, or it is seriously undermined, and results in a systematic selection bias – a deviation from a representative sample of the general population, thus distorting research findings (Matsuzaki 2018; Porsdam et al. 2016; Schönberger 2019). On the other hand, even if patients are going to share their data, the amount of it might not overcome the risk of discrimination. The medical records of some of the most vulnerable groups, especially from technologically underdeveloped territories, might be poorly collected or digitized, thus resulting in sample size disparity. Therefore, available raw data may reflect and expand existing bias and, in turn, unfairly affect members of protected groups based on sensitive categories like gender, race, age, sexual orientation, ability, or belief (Schönberger2019). To summarize, even if artificial intelligence system is modeled in an explainable way, the interpretability of the outcomes it provides may require assessment of risks related to possible disparities in the raw data that system has been trained on, especially in the high stakes - life or death - situations. .

Fairness appears as another important concern while dealing with disproportionally inaccurate medical decision-making on minorities (Schönberger 2019). Potential bias by artificial intelligence systems, either as a result of unethical engineering, selection bias or some other mistakes in data governance, might affect the access and confront with the results of carefully verified and unbiased diagnostic procedures. As Lencx et al. (2018) suggested, data scientists, security experts, and bioinformaticians should complement the expertise of clinicians, ethicists, and other traditional ethics review committees’ members, to inspect how the data will be securely collected, stored, and shared as well as how it will be classified, sorted, and analyzed. Moreover, it is suggested that research regulators should consider whether complementary governance mechanisms such as data boards, data security committees, or allied bodies are necessary to expand the bandwidth and sensitivity of ethical oversight. Finally, it is legitimate to raise the question of whether ethics review committees should be the only governance body responsible for the evaluation of biomedical big data research. Given their traditional mandate,
which is deeply rooted in the pre-digital era of biomedical research, it might be reasonable to argue that the ethics review committees are ill-suited to exercise exclusive ethical oversight on health-related big data research (Ienca et al. 2018).

Even if collective efforts for artificial intelligence systems reach reasonable level of trustworthiness, desired in conventional physician-patient dialog, the risk to undermine the patient’s autonomy remains. If the patient acts upon a personalized offer, made by artificial intelligence system, more refined personalized offers will follow. Then it might slowly block the patient from making other choices because they are less aware of alternatives. This removal of options from consumer consideration interferes with the principle of free choice (Jetten 2016). Thus, the over-reliance on artificial intelligence systems may reduce the willingness to engage in meaningful physician-patient dialogue, especially if the use of these systems is coupled -with economic pressure, for example by insurance companies, if recommendations made by these systems will become a precondition for financing certain diagnostic procedures (Vayena et al. 2018).

Furthermore, financial incentives to perform procedures may not be medically necessary for the patient and unnecessary medical procedures may expose patients to the superfluous risks of those procedures (Jetten, 2016). To make decisions on a treatment plan, physicians should not merely consider quality of data, collected and used in the process of training artificial intelligence systems (no matter how detailed) but should also employ their experience (to conclude results according to individual situation) and a deeper understanding of illness to conclude everything in the certain contextual picture (to take into account patient’s psychological, emotional, social, or even religious aspects). In this respect, some scholars assume that “although technology can create new risks that need to be disclosed as part of the informed consent process, it also can present opportunities to improve consent discussions with patients” (Michalski et al. 2016). For instance, to reduce the uncertainty of advanced preventive diagnostics (epidemiological statistics) more data and more factors are needed, which are required by societal rather than individual interests. Moreover, this may cause so-called “health violence”, which means that information itself means nothing if there are no measures or opportunities to deal with health disorders due to limited access to health services or scarce resources, as observed in developing countries. Furthermore, advanced preventive diagnostics might also induce moral distress, self-blaming, desperation, etc., and leave a patient alone with a very probable diagnosis for cancer, genetic disease, or an untreatable disease. Therefore, the process of receiving informed consent for diagnostic procedure should also include consideration whether it is ethical to offer such procedures in the first place, despite their efficiency. That is, is it ethical to be informed that one is getting, for example, terminally ill and there is nothing that can be done to help?

Overall, when technology has health implications, regulatory control, or standards to guide new convergence in the health ecosystem becomes critically important (Nebeker et al. 2019). Despite a number of challenges, some ethical issues might be potentially solved if only informed consent is well-balanced and adequate tactics of information sharing are chosen. Members of the medical community and others have already called for changes to ethical guidelines and policies for the training of artificial intelligence devices (Char et al., 2019), and require the mandatory involvement of an ethics board academic research environments (Nebeker et al. 2019). Given the importance of personal interaction, especially in medical diagnostic procedures that determine the need (if any) of following medical intervention, the ethical audit of artificial intelligence tools should become an essential precondition for accreditation as in medical devices regulation (Vayena et al. 2018). This task also calls for educating the new generation of experts with deep interdisciplinary training in medicine, ethics, and technologies.

3. Legal Challenges of Applying Artificial Intelligence in Medical Diagnostic Consultations

As the traditional paternalistic attitude, when physicians unilaterally decided on the well-being of the patients, changed through the 20th century, developing towards the implementation of the principle of patient autonomy, the relationships between patient and healthcare professionals significantly changed. Respecting patient’s dignity and self-determination, communication, and decision-making process became an important part of patient care and the basis for the doctrine of informed consent in healthcare. The legal approach to requirements to get informed consent from patients is based on the legal rules and case law regarding the content of the healthcare professionals’ duties to provide all the necessary information and get competent consent from the patient for the healthcare procedures. These legal requirements are derived from the human right to privacy, which contains also the medical condition and physical integrity of the person and it is generally prohibited to interfere with that right without a justified reason or the consent of the person himself (Glass v. the United Kingdom (Application No 61827/00) [2004] 1 FCR 553).

The landmark legal definition on the concept of informed consent in the court of law was brought by Judge Cardozo in the 1914 of *Schoendorff v. Society of New York Hospitals*. In his holding, Judge Cardozo explained the patient’s right to self-determination by stating that every adult and a person with common-sense has the right to decide what to do with his or her body (Faden and Beauchamp 1986). Later, the necessity of the free consent of the person was clearly established in the context of medical trials during the Nuremberg process, applying principles later enlisted in the Nuremberg Code. Performing any medical procedure without previous consent, unless within the application of justified and legally established exceptions was regarded as a clear violation of legal principles and regulations. Subsequently, in the World
Medical Association’s Declaration of Helsinki, the requirement of informed consent was elaborated taking into consideration the necessary scope of information, voluntary nature of decision-making, protection of vulnerable groups of individuals, as well as having explained the use of sensible personal data. Similarly, the Belmont report was issued in 1979 where the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research acknowledged the importance of informed consent and addressed the controversy surrounding the nature and possibility of such a consent by providing three criteria for assessing the informed consent process: information, comprehension and voluntariness. While the abovementioned documents are directed to biomedical research, the Council of Europe Convention on Human Rights and Biomedicine №164 establishes the legally binding requirement of informed consent for any intervention in the health field. Article 5 of the Convention determines that intervention into health may only be carried out after the person concerned has given free and informed consent to it. As these provisions have laid the foundations for bioethics and promotion of human rights in the field of biomedicine, it is currently undisputed that patient must be provided with the appropriate amount of information before intervention into health as to the purpose, nature, consequences, and risks involved in particular procedures.

Hence, the duty to explain medical procedures before intervention into patients’ health is clearly established in law as one of physicians’ professional duties. In a bilateral legal relation, like one of physician-patient, primary duty of one party correlates with the primary claim right of the other. If the physician has the professional duty of care, the patient has the right to demand maximum efforts providing the services or liability in case of medical malpractice. Accordingly, if the physician has the duty to explain the purpose, nature, results, and risks involved in particular diagnostic or other medical procedures before intervention into patients’ health, the latter has the claim right to such explanation. Thus, the patient has the right to explanation, emerging from the general tort or contractual law principles, that corresponds to the physicians’ duty to explain medical procedures beforehand. Additionally, explainability of medical procedures is an important legal category not only in contract and tort law, but also in data protection law, at least with specific regard to artificial intelligence applications (Schönbeger, 2019) and within the jurisdictional reach of the EU General Data Protection Regulation (GDPR) (Hacker, 2020).

Within this context, there is a yet debated claim that data subject has a general ‘right to explanation’ of all decisions made by artificial intelligence systems. Since the right to explanation is contained only in the (non-binding) recital 71 of GDPR, there is an argument that a right to explanation of individual decisions does not derive from Art. 22(3) GDPR (see for e.g. Wachter et al 2017). Some scholars claim that the right to explanation is an inappropriate remedy given the difficulties of implementing this right on machine learning algorithms (Edwards and Veale, 2017). Moreover, as observed by Malgieri (2019), most Member States of the EU do not include such safeguard in their national data protection law, except for Hungary and France. Other authors claim that in respect to algorithmic decisions that have a legal or significant effect (Art 22(1) GDPR), the right to explanation seems to be a real and tangible right unequivocally provided by the GDPR (see for e.g. Casey et al 2019, Goodman and Flaxman 2017, and others). Yet another group of scholars reaches the right to explanation through systematic interpretation of Article 22(3) and Recital 71, with the right to information and access, set in Arts 13(2)(f), 14(2)(g) and 15(1)(h) of GDPR provisions (see for e.g. Brkan, 2019, Mendoza and Bygrave 2017, Powles 2017, Pagallo 2017 and others).

However, Wachter et al (2017) legal arguments are not an argument against a right to explanation in general, but only their cramped version of it, ignoring the positive value of Recital 71 (Selbst and Powles, 2017). Although recitals have no binding power, the EU Court of Justice applies them now and then, so as to establish the meaning of the valid law (Pagallo 2017). Technical difficulties, on the other hand, should not be used as an argument to refuse protection of core human rights to dignity, physical integrity or life. Indeed, a large numbers of ethicists, practitioners, journalists, and policy-makers, like Microsoft, Google, the World Economic Forum, the draft AI ethics guidelines for the EU commission, etc. agree on a core principle for artificial intelligence that falls under the umbrella of ‘explicability’ (Robbins 2019, Floridi et al 2018). Therefore, we agree with arguments of Selbst and Powles (2017), leading to the conclusion that it makes sense to regard rights to ‘meaningful information about the logic involved’ in automated decisions (Arts 13–15 of GDPR) as a right to explanation, which should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law.

Moreover, as correctly noted by Pagallo (2017), the right to meaningful information about the logic involved, as well as the significance and the envisaged consequences of such data processing for the data subject is associated with an ex ante explanation that regards the system functionality, and it may indeed play the main role in ensuring the fulfillment of the requirement for individual informed consent. It should be noted, that right to ex ante explanation is equivalent to a traditionally well-accepted right, namely, a right to be informed or a right to informed consent (Kim, 2018). Therefore if GDPR indeed contains a novel addition under the new name of a “right to explanation”, it should additionally involve ex post explanations, as a precondition for placing trust intelligently (Kim, 2018). Within this context of ex post explanations, it seems fair to admit that one can contest decisions, only on the basis of the ways, how the decision made; thus, without an explanation of how the algorithm works, it would be hard (if possible at all) to enforce a right to contest automated
decisions and thus the rights to fair trial and effective remedy enshrined in Articles 6 and 13 of the European Convention on Human Rights (Pagallo, 2017).

Therefore, the legal problem of interpretation does not revolve around whether a right to an explanation in the process of informed consent to medical diagnostic procedures exists, especially under the EU law. Under the generally accepted principles of contract and tort law, complemented with data protection law in the context of artificial intelligence applications, physician has the duty to explain and patient has the right to demand such explanation. Rather, the issue concerns the extent of such right, e.g. whether the right to explanation includes the explanation of how algorithms work (Pagallo, 2017).

In some cases, detailed technical explanations about the inner workings of the applied algorithms, may indeed contribute little to facilitate a meaningful choice about a diagnostic or other medical procedure (Schönberger 2019). However, in the high stakes, life or death, cases, such information might be legally required, rather than ethically desired. Physician involved in meaningful and trustworthy physician-patient dialogue on the benefits and risks of diagnostic procedures should be capable to serve both cases. Thus, contemporary legal issues concerning informed consent of the patients focus mostly on the scope of the information that must be sufficiently provided before the patient has to decide which healthcare services or providers to choose. Legal scrutiny shifted on whether consent given by the patient was supported with enough information to make competent decisions before consenting for certain medical treatments. As the scope of such information and the ways it can be provided in different forms and time-consuming approaches are indefinite, the approach for the legal scrutiny is tending to be patient-centered one (Chan et al. 2017). It means that the patient–physician relationship must be built on very effective communication, discussing not only technical aspects of medical treatment possibilities and expected outcomes, but also individual preferences and the values of the patient. Such standard is applied in medical practice, and in case law, allows to determine the scope of information that must be provided and explained to each individual patient and whether it was done appropriately (Chan et al. 2017).

The opacity problem of artificial intelligence seems to be in direct conflict with the right of patients to be provided with meaningful information about the logic involved as well as about the significance and the envisaged consequences of diagnostic or treatment procedures, or other interventions into health (Bachle 2019). Opening of “black box” to understand model algorithms, training data parameters and weights, and even source code might look like a solution to this problem. However, as Hoerenand Niehoff (2018) notes, self-developing models, such as neural networks, are not possible to control in a linear manner neither for its creators, nor users. Moreover, mathematical formulas in bits and bytes, even if disclosed, will not be comprehensible to an average physician or patient without special training in artificial intelligence technologies and, therefore, is useless. Physicians and patients need an explainability model that is understandable by common human beings (Hoerenand Niehoff, 2018). Therefore, the crucial issue here is to determine an optimal level of patient-centered explicability that would not demotivate healthcare stakeholders from implementing advanced technologies in their daily practice (Kiseleva 2019).

Moreover, even if it will be technologically possible, the opening of the “black box” solution might undermine the intellectual or other propriety rights of technology developers and other private stakeholders (Feldman et al. 2019). One’s right to swing arms ends just where the other’s nose begins (Chafee 1919), thus, the physicians’ duty to explain and corresponding patient right to demand an explanation should not adversely affect the rights of others, including the rights in trade secrets or intellectual property. Full disclosure of the artificial intelligence model would render the research and development efforts worthless, which would, in turn, diminish the incentive to invest and innovate (Hoerenand Niehoff, 2018). Accordingly, the balanced approach of required explainability should also consider the level of exposure that does not destroy the motivation of artificial intelligence developers to innovate.

Here, the analogy with the explainability of the physicians’ decisions might be drawn. That is, the assumption might be made that the same ethical and legal principles of explainability apply to the decisions made by physicians alone or with the support of artificial intelligence systems. At least for now, advancements of medical diagnostics tools should not change the level of explanation that is currently expected from the physicians as required by law. Thus, an explanation will have to remain a human-interpretable description of the process by which a decision-maker takes a particular set of inputs and reaches a particular conclusion (Doshi-Velez et al. 2017). An explanation does not require knowing the flow of bits through an artificial intelligence system, no more than an explanation from humans requires knowing the flow of signals through human brain neurons. The European Commission High-Level Expert Group on Artificial Intelligence (2019) also requires the explainability of artificial intelligence systems to those directly and indirectly affected only to the extent possible, as “a suitable explanation of the AI system’s decision-making process.” Auditing mechanisms are proposed as possible solutions that examine the inputs and outputs of algorithms for bias and harms, rather than unpacking how the system functions (Cath 2018). That is, regarding diagnostic outcomes operated by way of algorithms, the right to an explanation requires that the solution is to make artificial intelligence decision-making explainable, not to explain the artificial intelligence model (Spina 2018).
Following this parallel of human decision-making, the patient needs to understand its process and key factors that lead to it (Hoeren and Niehoff 2018). Thus, the explanation process should reveal key inputs, outputs, and causal relations between them. Importantly, these properties can be satisfied without knowing the details of how the system came to its decision (i.e., without revealing the internal contents of the system) (Doshi-Velez et al. 2017). These factors are simple enough and comprehensible for layperson humans like patients, just if like a physician would explain a human decision making (Hoeren and Niehoff 2018). Notably, the process of informed consent in diagnostics requires the physician to provide the patient with meaningful information before or during that process - when no medical treatment decisions are made yet. Hence, in the informed consent process, the physician must only provide the patient with ex ante meaningful information about the general system functionality, not the specific ex post-decision (Hoeren and Niehoff 2018). This forward-looking nature of informed consent, at least within the GDPR context, also does not allow the inference of anything beyond the explanation of system functionality (Schonberger 2019). But even in the ex-ante stage, an explanation as to why the artificial intelligence model generates a particular output and what combination of input factors contributed to that is not always possible due to the opacity problem. In those cases, other explicability measures (e.g., traceability, auditability, and transparent communication, etc.) may be required (European Commission 2019).

Moreover, generating explanations takes time and effort by the physicians, thus, there are concerns of direct cost and lost opportunities to devote that time and effort in providing healthcare services to other patients as well. Therefore, the utility of explanations must be balanced against the cost of generating them (Doshi-Velez et al. 2017). Obviously, novel and highly complex medical diagnostics tools add up these costs. Moreover, economic inefficiencies are also tailored to the failures of human intelligence, which is boundedly rational, opportunistic, and performed under the uncertainty and asset specificity (Williamson 1984). That is, humans are notoriously inaccurate when providing (or receiving) rationales for decisions or predictions (Doshi-Velez et al. 2017; Matsuzaki 2018). Accordingly, the informed consent process is often ill-configured to meet patients’ informational needs and tend to be generic, containing information intended to protect the physician or hospital from litigation. Informed consent documents are often signed minutes before the start of a procedure, a time when patients are most vulnerable and least likely to ask questions—hardly consistent with what a reasonable patient would deem acceptable. That is, at least for now, benefits of the informed consent are weighted to assigning blame to the patient through the carefully drafted consent instruments. So, the law has yet to promote a process that truly supports a reasonable patient–centered standard (Spatz et al. 2016).

Patient decision aids, in various forms providing information about different options for treatment, are considered as a possible solutions and tools for improved and shared decision-making. The purpose of shared decision-making, is to find what is the best for each patient by ensuring that patients are informed and participate in the decision–making process by being involved in a discussion about diagnostic and treatment options within the context of individual patient’s values and preferences. Patient decision aids are being developed in various forms providing information about different options for treatment. Decision aids may vary from set of cards focused on the issues that matter the most to patients and highlight different medications used to treat certain condition (Sepucha, 2016) to videos that cover the main treatment options (Mangla, 2018). Testing and researching the effects of such aids on patient’s perception of self-determination and dignity show positive results, favorable to the quality of healthcare services (Stacey et al. 2017). Although the information provided in such a manner is meant to be more accessible and comprehensible to the patient, still, being only the aid, it does not provide for ready decisions, eliminating the need for final decision-making through the consultation and communication between the patient and physician. Thus, the legal standard of care, applied to the physician’s professional duties in the process of informed consent, requires a full understanding of the medical treatment and its options that must be individually and comprehensively explained to every patient.

Another ultimate solution is to shift the burden by explaining how the artificial intelligence system works from physician to the technology developers. It is technically feasible to create artificial intelligence systems that provide the same kinds of explanations that are currently expected of humans under the law (Doshi-Velez et al. 2017). That is, in a concise, transparent, intelligible, and easily accessible forms, using clear and plain language that the patient can comprehend, verify, and, most of all, effectively contest automated decisions (Hoeren and Niehoff 2018). Such solutions may be cost-effective and solve the failures of human intelligence. However, at least in the European Union, Article 22(1) of GDPR protects the data subject from being a subject to a decision based solely on automated processing without the data subject’s explicit consent or authorization by the state law. Accordingly, explanations built by the developers of artificial intelligence systems, although may cover the technical aspects of their systems, will be incomplete in the context of meaningful physician intervention and oversight. Overall, it is the professional duty of the physician to communicate to patients and discuss diagnosis, possible choices, and expected outcomes; moreover, it is a duty that cannot be delegated to a third party.

Moreover, the technical complexity of artificial intelligence and the possibility of integrating a third party in the process of informed consent raise legal liability issues of complexity of moral authority. In case of conflicting diagnostic conclusions, should the doctor or artificial intelligence system have the moral authority regarding treatment decisions in the face of inconclusive evidence? Who is legally liable for a medical error involving the use of a sophisticated technology? Schiff
standards is legislation. To performance standards of some kind, no such standards currently exist, and the only viable source of legally enforceable interdisciplinary training in medicine, ethics, and technologies. Moreover, while it is conceivable that robots could be held understanding of technological complexities. This need calls for educating the new generation of experts with deep oversight on health-related big data research and development due to the lack of interdisciplinary training and, therefore, ethics review committees, even if supported with legal standardization, may be ill-suited to exercise exclusive ethical artificial intelligence tools should become an essential precondition for accreditation (Vayena et al. 2018). However, the high level of ethicist involvement in the creation and training processes (Lupton 2018). Moreover, the ethical audit of the task for the creators of artificial intelligence systems is, therefore, to embed the ethical behavioral patterns that require the high level of ethicist involvement in the creation and training processes (Lupton 2018). Moreover, the ethical audit of artificial intelligence tools should become an essential precondition for accreditation (Vayena et al. 2018). However, the ethics review committees, even if supported with legal standardization, may be ill-suited to exercise exclusive ethical oversight on health-related big data research and development due to the lack of interdisciplinary training and, therefore, understanding of technological complexities. This need calls for educating the new generation of experts with deep interdisciplinary training in medicine, ethics, and technologies. Moreover, while it is conceivable that robots could be held performance standards of some kind, no such standards currently exist, and the only viable source of legally enforceable standards is legislation.

4. Conclusions

Analysis of artificial intelligence applications in medical diagnostic procedures revealed that these new instruments may be very effective for physicians to make more reasonable and more grounded medical decisions, especially in diagnostics. Currently, the determination of illness in clinical practice is based on objective investigation and proved by evidence of (ontological) factors. Thus, advanced health technology tools are supposed to improve this evidence-based knowledge pool, and so it is expected to increase the probability of true illness as well as successful treatments. Moreover, trustworthy artificial intelligence can also assist on a broader scale examining and identifying general trends in the healthcare and treatment sector, leading to earlier detection of diseases, more efficient development of medicines, more targeted treatments, and ultimately more lives saved (European Commission 2019).

However, the applications of such technologies in everyday practice to this day is ambiguous and debated as it might induce both advantages and disadvantages, raise both ethical and legal concerns. The process of informed consent is one of the major ethical concerns of the New Health phenomenon, especially in the context of technological complexity and opacity of artificial intelligence systems. Without sufficient transparency, neither patients nor physicians can understand how and why intelligent machines generate the diagnosis. Therefore, the “black box” problem of artificial intelligence and related problems of interpretability of inner logic remains one of the greatest challenges for medical ethicists and regulators (Vayena et al. 2018).

Lack of transparency leads to a lack of trust, which is the main element of a successful relationship between a physician and patient. However, it is still not clear if artificial intelligence algorithms are trustworthy, how results of such “diagnostic” procedures should be disclosed to a patient, what sort of information should be given to a patient, and finally how detailed should the explanations be provided. Trustworthy dialogue between physician and patient can be achieved only if a certain level of artificial intelligence transparency and, thus, explainability is satisfied. Legal analysis of the informed consent process revealed that the right to an explanation does not require exposure of the artificial intelligence model and it may be technically feasible to shift the explanation burden from physician to the technology developers. However, an explanation is a professional duty of the physician and, thus, the information provided by the developers of artificial intelligence systems will be incomplete in the context of meaningful physician intervention and oversight in the diagnostic process.

Therefore, despite the possibility to shift at least partially the explanation burden, a physician still needs to understand the complexity of artificial intelligence algorithms used in medical diagnostics. Furthermore, a limited ability to understand the operation of the systems leads to a limited ability to explain the procedures, choices, and predictions to the patient, which also leads to a limited ability of the patient to provide sufficient informed consent for diagnostics and consequent healthcare procedures. Consequently, the relationship between patient and physician, which should be based on trust, could be disturbed (Price 2015).

The task for the creators of artificial intelligence systems is, therefore, to embed the ethical behavioral patterns that require the high level of ethicist involvement in the creation and training processes (Lupton 2018). Moreover, the ethical audit of artificial intelligence tools should become an essential precondition for accreditation (Vayena et al. 2018). However, the ethics review committees, even if supported with legal standardization, may be ill-suited to exercise exclusive ethical oversight on health-related big data research and development due to the lack of interdisciplinary training and, therefore, understanding of technological complexities. This need calls for educating the new generation of experts with deep interdisciplinary training in medicine, ethics, and technologies. Moreover, while it is conceivable that robots could be held to performance standards of some kind, no such standards currently exist, and the only viable source of legally enforceable standards is legislation.
References


Asimov I (1964) Visit to the world’s fair of 2014. New York Times 16


Brkan M (2019), Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond, 27 International Journal of Law and Information Technology 91, pp 110–119


Council of Europe, Convention for the Protection of Human Rights and of the Human Being with regard to the Application of Biology and Medicine (Convention on Human Rights and Biomedicine or the Oviedo Convention) (CETS n. 164)


Edwards L and Veale M (2017) Slave to the Algorithm? Why a ‘right to an explanation’ is Probably not the Remedy you are Looking For, 16 Duke Law & Technology Review 18, p 81


Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a right to explanation. AI Mag 38(3):50–57


Jha S, Topol EJ (2016) Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA Dec 13;316(22):2353-4

Keane PA, Topol EJ (2018) With an eye to AI and autonomous diagnosis. NPJ Digit Med 1:40


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


