Machine Learning in the EU health care context: exploring the ethical, legal and social issues

Abstract

The diagnosis and clinical decision making based on Machine Learning technologies are showing significant advances that may change the functioning of our health care systems. These advances promise more effective and efficient healthcare, at a lower cost. This may allow healthcare professionals to recover 'high-touch' time with their patients. The evidence suggests that all these promises have yet to be demonstrated in clinical practice, but what is undeniable is that these technologies are resignifying the relationships in the health landscape, particularly the physician-patient relationship, which we could already redefine as "physician-computer-patient relationship". Although it is true that today fully automated decision systems are scarce in comparison with integrative decision support systems, we cannot fail to observe the horizon they define. Our most recent regulatory framework, defined by the General Regulation on Data Protection, has tried to avoid this scenario by including the right not to be subject to a decision based solely on automated processing. In this paper, however, we argue that this legal tool is adequate but not sufficient to address the legal, ethical and social challenges that Machine Learning technologies pose to patients' rights and health care givers' capacities.

1. Introduction. AI As a New Promised Land for Health Care Systems: The Pearls and the Perils

In 1955, John McCarthy coined the term ‘Artificial Intelligence’ (AI) as the name for the science and engineering of making intelligent machines (Hamet & Tremblay, 2017). Today, AI focuses on how computers learn from data and mimic human thought processes (Noorbakhsh-Sabet et al., 2019). The field of medicine is one of the most promising application areas for AI (Yu et al., 2018), due to AI’s ability to handle and optimise very complex data sets from very complex systems (Bini, 2018).

Software programs and Machine Learning (ML) are able to convert big data into algorithms, providing advantages such as flexibility and scalability; the ability to analyse diverse data types for disease risk, diagnosis, prognosis, and appropriate treatments; as well tackling unique challenges for model training and refinement; and managing the need for data pre-processing and making crucial ethical considerations (Ngiam & Khor, 2019).

ML algorithms can be classified into two groups. The first involves deep learning platforms, such as IBM’s Watson Oncology -Dr. Watson-, which is fed everything written, in every language, at any time, that is related to cancer diagnostics and treatment (Londhe & Bhasin, 2019). The more information Watson has about a patient, the more accurate it’s assistance will be; but actually, it is not yet perfect (Bini, 2018). In the second group, the algorithms fall under the denomination of pattern medicine, based on data collected through imaging techniques such as x-rays (Kallianos et al., 2019), mammogram images (Le et al., 2019), immunohistochemical stains (Niazi et al., 2019), and retinal images (Schmidt-Erfurth et al. 2018), among others. Some of the pattern medicine algorithms have been approved by U.S. Food and Drug Administration, and most of them have been validated by comparison to the precision exhibited by human beings (Ngiam & Khor, 2019). Nevertheless, the future of AI in diagnosis and treatment should be based on hybrid strategies, since specific medical diagnostic and prognostic success for each concrete matter depends on the nature of the task, type of data, and available information about the related disease (Shahid & Singh, 2019).

All that being said, ML has limited exploratory power: algorithms might be able to identify correlations, but not necessarily prove causation. So, despite their differences, ML and evidence-based medicine can and should complement one another (Scott, 2018). In this scenario, the clinician's role is to be a bridge between machine and decision (Coeira, 2019), and professionals across different fields, speaking different languages, should be trained and integrated with the real benefits and applicability of developed algorithms in health care (Nuñez-Reiz, 2019).

The success of AI, therefore, can bring about a dramatic change in the way medicine is understood, and in the functioning and sustainability of public health systems. However, it also poses considerable challenges. To begin with, it inevitably affects the core of medical practice: the relationship between the care giver and the patient. The emergence of AI means that doctors must consider their own roles. They will not only be responsible for their patients' health; they will be managers of their patients' personal data, with a commensurate obligation to inform them about the use of automated decision-making systems that physicians do not fully understand and the recommendations of which they do not always share. In effect, doctors will be forced to rethink the way they manage the information at their disposal and the very idea of data confidentiality.

This new scenario may cause patients to feel helpless against the use of opaque tools and automated decision-making processes that affect essential aspects of their lives. Faced with this dilemma, the European Union has developed a regulatory framework focused on defending the rights of the data subject, in this case patients. Its General Data

Protection Regulation has proclaimed a patient's right to information and a right not to be subjected to profiling and automated decision-making processes which, it is hoped, will serve as an efficient mechanism to protect patients from the misuse of their data. However, this general framework shows some gaps and deficiencies that need to be clarified.

This paper is intended for this purpose. Its aim is to explain how the implementation of AI can pose problems for patients' and doctors' interests. It analyses the mechanisms created to address these issues, highlighting their weaknesses and incorporating suggestions on how to resolve them. We begin by showing the main technical obstacles that make the guarantee of adequate decision making by patients, doctors, and others responsible for health systems extremely complex. Then, we propose some measure that might contribute to face these challenges successfully.

2. Understanding AI: Intrinsic Issues That Render Transparency Highly Complicated

As discussed, the implementation of AI in health care systems will only respect patients' rights if patients are allowed to make the final decision on whether or not to use automated decision-making systems. However, this is very difficult if patients lack sufficient information, which should be provided by their physicians or health care providers. Achieving the goal of sufficient information transfer is complicated because there are multiple factors that seriously hinder an efficient transmission of information and subsequent decision making. These include: the difficulty of assimilating the paradigm shift introduced by AI; the (current) deficiencies of AI systems applied to the health care sector; the difficulty of reconciling business interests and transparency; and the shortcomings inherent in the construction of algorithms. In this section, we analyse each of these issues.

2.1 The difficulty of assimilating the paradigm shift introduced by Artificial Intelligence

The first issue involving the use of algorithms produced through Deep Learning tools is that the whole philosophy underlying their production differs substantially from the way science has been conducted at least since the main scientific methods were developed. Science generally advances through the formulation of hypotheses (i.e. possible causal associations between two events), which can be subjected to a consistency test through contrast with reality. It is true that in recent years the complexity of some disciplines has forced us to accept alternative models of scientific evidence (Sterky & Lundeberg, 2000), but these subtle exceptions to the general rule have not yet been assimilated by health professionals or their patients. In this particular part of science, the rigid rules of hypothesis-reality contrast continue to apply.

Luckily or regrettably, the algorithms do not fit this form of epistemological functioning. An algorithm does not formulate a hypothesis to contrast with data extracted from the real world, but rather the hypotheses are precisely the result of the analysis of these data. An algorithm only discovers correlations that can predict, not causalities that can explain. In this sense, AI, in its current development, is a complement rather than a substitute for science (Ellis & Silk 2014). If science is assumed to have both the ability to explain and the ability to predict, this part of AI is limited exclusively to the latter (Anderson, 2018). However, as we have anticipated, the value of AI lies in the fact that it is able to achieve acceptably accurate diagnoses with a more efficient use of resources, and much more accurate prognoses.

The question is whether this enormous limitation – the practical impossibility of giving a causal explanation for specific recommendations – will make the use of AI in health care acceptable to health professionals and patients. In the case of professionals, the generalisation of AI will, to a large extent, be an amendment to their entire training, as they will often need to adjust their performance to a mechanism that does not provide them with reasons, but with probabilities. Therefore, the ability to interpret these probabilities clearly and sensitively represents an additional—and essential—educational demand for patients and their families (Wartman & Combs, 2019). In the case of patients, it seems at first glance that the situation may be less complex, but, in a world where conspiracy theories are becoming increasingly predictable, knowing what the reaction will be to the use of an eminently opaque technology is a mystery.

Faced with this situation, it is obvious that the key to transmitting adequate information lies in efficient training of health professionals, which does not exist at the moment. As Char et al. stated, 'Physicians who use machine learning systems can become more educated about their construction, the data sets they are built on, and their lim. HEtations. Remaining ignorant about the construction of machine-learning systems or allowing them to be constructed as black boxes could lead to ethically problematic outcomes' (Char, Shah, & Magnus, 2018). Thus, training is a key concept in terms of efficient information.

2.2. The (current) shortcomings of AI systems applied to the healthcare sector: The 'black box' medicine

The problems described in the previous section would probably be less important were it not for the fact that many of the algorithms developed by machine learning systems are inherently opaque tools. As Ferretti, Schneider and Blasimme (2018) have rightly described:

> "While most people recognize the promise of applying AI systems to medical diagnosis and decision-making, many are worried about the use of partly autonomous computer programs for medical purposes. This fear has to do with a

> characteristic of many ML methods. AI systems that incorporate ML learn with a varying degree of supervision which rules they need to follow in order to perform their task. The programmer sets up the system so that it can learn to do something. However, he or she does not decide, nor is necessarily aware of the rules the AI system has learnt and is following in order to do what it is supposed to do. This characteristic is often referred to as the opacity of ML. For the same reason, AI systems based on ML are often called black boxes, to stress that it is hard or even impossible for human users to open them up, so to say, and see for themselves what the machine is doing (or, which is the same, what rule the machine has learnt and is employing). The possibility that these systems could remain opaque to their own creators as well as to their end-users is a cause of concern."

The issue, in short, is simple: it is very difficult to talk about algorithmic transparency in the case of ML technologies because the operation of these techniques makes it almost impossible to understand how their inferences operate (de Miguel Beriain, 2018); not even their programmer could do it. Indeed, Consequently, the fact must be faced that there is a part of the information that is not available to patients, physicians, or health care providers. In this context, a field of research called explainable AI (xAI) is raising. It is aimed at producing methods that make algorithmic decision-making systems more trustworthy and accountable (Mittelstadt et al., 2019). Nevertheless, further work in this field is mandatory since explanatory systems are focused on programmers or IA experts, not on end users or policy-makers (Gilpin et al., 2018).

2.3. The difficulty of reconciling business interest and transparency

The description of the facts is not complete without an account of a third factor that contributes substantially to the difficulty of understanding the new reality faced by patients and health care givers. The companies that develop the algorithms invest considerable resources in their development. This includes both the need to procure large and well-ordered databases -Smart data- and the development of the AI mechanism itself.

Non-public companies seek a return on that investment. Therefore, opacity is an intentional form of self-protection that attempts to keep trade secrets and the competitive advantages involved (Burrell, 2016). In other spheres of human activity, this is often achieved through mechanisms such as patents or copyright. In the case of data, the system of intellectual property protection offers notable shortcomings. Hence, within the EU, what is known as sui generis database rights, a property right settled by Directive 96/9/EC of 11 March 1996 on the legal protection of databases, according to which

> Member States shall provide for a right for the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database (article 7.1).

Unfortunately, until now we have not been able to develop in parallel a ‘sui generis algorithm right’. Therefore, algorithms continue to be considered as ideas; creations of the mind that do not find accommodation in the intellectual property protection regime, unless we accept the theses proposed by authors such as Minssen and Pierce (2018), who consider that patenting algorithms could be possible in the EU arena. Otherwise, companies have no choice but to hide their algorithms under the trade secret layer in order to maximize their returns. Consequently, the inherent opacity is often exacerbated by this deliberately sought-after form of opacity.

2.4. Inherent flaws in algorithm construction

Finally, distrust of algorithms is by no means unjustified. Evidence shows that machine learning algorithms are often biased and may lead to discrimination based on classes like race and gender (Buolamwini & Gebru, 2018). The content of the dataset determines how the algorithm will make decisions on real-world cases (Wellner & Rothman, 2019). Therefore, significant problems arise from errors and biases latent in data training sets that tend to be reproduced in the outputs of these tools (Zerilli et al., 2018). For example, a database comprised mostly of information about white males will surely produce an algorithm much less accurate for Hispanic women. In other cases, failures stem from the deficiencies generated by a machine learning system that induces unlucky correlations through the incorporation of a human collective thinking system that cannot avoid being biased. As Char et al. have rightly pointed out, 'Subtle discrimination inherent in health care delivery may be harder to anticipate; as a result, it may be more difficult to prevent an algorithm from learning and incorporating this type of bias' (Char et al., 2018).

Patients may therefore legitimately ask whether the algorithm being used to make a diagnosis or assess their response to treatment is adequately adapted to their personal circumstances, or whether it is not. Unfortunately, these questions can only be clarified if care givers ensure that the AI mechanisms have been subjected to validation methods and monitoring systems capable of verifying that there are no biases or errors incompatible with their use in the health care system. And a care giver, of course, will hardly be able to provide the patient with any information other than whether or not these quality control systems have been implemented.

3. Protection of data subjects under EU legislation with respect to AI applications. Right to information and prohibition on fully automated decisions

On the basis of the above limitations (that the operating logic of AI differs substantially from that of science, and that the algorithms are inherently opaque), the legislator has attempted to protect data subjects (patients, in this case) without banishing AI. At the EU level, this attempt has resulted in the development of two normative initiatives: the proclamation of a fundamental right to information and the prohibition of solely automated decision-making. In the following sections, we will attempt to set out the fundamental bases of these policies, and their limitations.

3.1 The right to information

One of the first issues involved in giving patients adequate information about the use of AI in the health care process is that patients need to be made aware that these mechanisms are being used to make decisions that affect them personally. In some cases, it will be easy to guess this, as in order to use AI efficiently it will be necessary to request a huge amount of data from the patient, making it very complex to hide its use. In other cases, however, the controller could use only data that were already available, such as the data already included in the patient's clinical history. It is, therefore, necessary to avoid this possibility by making a rule that obliges the data controller to inform the patient of the intervention of AI mechanisms in decision-making. This is what the Regulation provides for in Articles 13 and 14. Article 13 (Information to be provided where personal data are collected from the data subject) states:

> "1. Where personal data relating to a data subject are collected from the data subject, the controller shall, at the time when personal data are obtained, provide the data subject with all of the following information: (…) c) the purposes of the processing for which the personal data are intended as well as the legal basis for the processing."

Furthermore, its number 2 states that

> "2. In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: (...) g) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic

> involved, as well as the significance and the envisaged consequences of such processing for the data subject”

Thus, the Regulation grants the patient the right to be fully aware of the use of personal data collected by the controller if these data are to be used for automated decision-making purposes. It is necessary to point out that the Regulation does not use the term ‘solely automated decision-making’, but only ‘automated decision-making’. This seems reasonable, since otherwise the obligation to communicate the fact that an AI tool would be involved would be reduced if the process included some form of human supervision. In this way, the GDPR confronts the secret use of automated decision systems, which has been claimed to be harmful (O’Neil, 2016): every patient has the right to know if her personal data has been subjected to this kind of automated processing. It is important, on the other hand, to underline the fact that this obligation applies not only to automated decision-making but also to profiling, that is:

> “any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements”[art. 4].

The provisions of Article 13, which relate to personal data provided by the subject to the controller, are complemented by those of Article 14, which applies to information to be provided where personal data have not been obtained from the data subject. This clause states that, for such data, controllers must also inform the data subject about the purposes of the intended processing of the personal data as well as the legal basis for that processing [Art. 14.1.c], and about the existence of automated decision-making, including profiling, and *‘meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject’*.

To summarize, the Regulation has shaped a scenario in which the right to an explanation about the use of AI tools for profiling or automated decision-making plays a dominant role. This is of crucial importance in terms of the reliability of the health system, as it avoids reasonable suspicions about the ultimate purpose of introducing AI into the process. Moreover, the proclamation of this right contributes to a reinforcement of the trust between the health personnel (who can exercise the role of the data controller, as has been said) and the patient. With regard to this caregiver–patient relationship, it is mandatory to evaluate the impact that the introduction of AI in clinical decision-making will have. The right to information avoids giving patients the impression that *"they are being marginalized in decisional processes regarding their health, thus affecting their decisional autonomy and their sense of self-determination. In light of these considerations, restricting disclosure to solely-automated activities may turn out to be insufficient"* (Ferreti et al., 2018).

However, this apparently strong legal structure hides some important holes, mainly related to the content of this general right, proclaimed in the legislation, to receive an explanation. What does this right mean in practical terms? Does it mean that patients are given a right to know about the technicalities of the decision-making tool? Does it only mean that they should be informed that an AI tool will be used? In the legal arena, this issue has raised a profound discussion, which is still far from being resolved (Brkan, 2019; Goodman & Flaxman, 2017; Selbst & Powles, 2017; Wachter, Mittelstadt & Floridi, 2018). To enter into the subtleties of this discussion would clearly go beyond the boundaries of this text. However, we believe it is possible to set out some of the issues that seem most pertinent now.

3.2. What the right to explanation is not: a right to disclosure

First, we must highlight that the right to an explanation by no means implies that the data subject is empowered to have access to the algorithm as such. This would clearly render industrial secrecy impossible and would deprive the developer of the algorithm of any way to exploit the result of his investment commercially. This result is unacceptable for both legal and practical reasons. From the legal point of view, it would contradict the spirit of the Regulation, whose Recital 63 states *"that right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software"*.

As Ferretti et al. (2018) wrote,

> "It follows that, while data controllers must disclose that they are conducting profiling or automated data processing, they are not obliged to reveal all details about their AI systems. In practical terms, this entails that data controllers may still be required to provide information regarding the general characteristics of their system, but they may not be compelled to explain what rules the AI system follows, how it has reached a conclusion, or how it has taken a given decision about a particular data subject."

Moreover, from a practical point of view, disclosing the algorithm would just provide patients with information that they could not really understand, a situation that is far removed from their needs and from the spirit of the Regulation. Indeed, one must consider that information about the logic must be meaningful to the data subject, who is, notably, a human being who can be presumed to have no particular technical expertise (Selbst & Powles, 2017).

Therefore, we must conclude that the right to an explanation does not include disclosure and, furthermore, that the right could not be satisfied by disclosing the algorithm (that is, the controller would not comply just by providing the patient with the algorithm used in the automated decision-making). Obviously, this does not mean that controllers can rely on the protection of their trade secrets as an excuse to deny access, and nor can they refuse to provide information to the data subject (A29WP, 2018).

3.3. What the right to explanation must include: a right to know the type of information that is being used and the general principles involved in the design of the algorithm

In our opinion, patients may, in short, assume that they will probably never know exactly how the algorithmic mechanism that will intervene in a crucial decision in their life works. This is not necessarily new; the sorts of explanations we cannot obtain from AI are the same as those we cannot obtain from humans either (Zerilli et al., 2018). However, this does not mean that patients cannot be provided with any form of relevant information. Indeed, there are some fruitful ways to guarantee that the explanation is sufficient to facilitate the exercise by patients of the rights granted to them by the GDPR and human rights law. For instance, Article 12 emphasises intelligibility and contains the requirement that '[t]he controller shall facilitate the exercise of data subject rights' (Selbst & Powles, 2017).

To begin with, it is perfectly possible to provide a layperson with general information about how an algorithm has been constructed or what type of data categories it uses. This has been understood, for example, by the Article 29 Data Protection Working Party, an advisory body made up of a representative from the data protection authority of each EU Member State, which played a prominent role in terms of the interpretation of the Regulation until it was replaced by the European Data Protection Board (EDPB) under the GDPR. The Working Party has stated:

> "Article 15 gives the data subject the right to obtain details of any personal data used for profiling, including the categories of data used to construct a profile. In addition to general information about the processing, pursuant to Article 15(3), the controller has a duty to make available the data used as input to create the profile as well as

> access to information on the profile and details of which segments the data subject has been placed into." (A29WP, 2018)

Similarly, patients must be made aware of the importance of the contribution made by the AI system in the final decision, including receiving all available information on the main factors in the decision, whether changing a certain factor would or would not have changed the decision, and why different decisions are reached in similar-looking cases, or the same decision in different-looking cases (Doshi-Velez et al., 2017). On the one hand, this also means that from the very first moment patients should know about the use that might be made of their data and the foreseeable consequences of the data processing for this purpose, as, indeed, is required by the Regulation; however, this requirement could be very limited in an actual scenario of big data analytics, where new data are created from inferred and derived data. Looking at how the automated processing of data and profiling works, it is undeniably true that the GDPR focuses primarily on mechanisms to manage the input side of the processing, and that the legal mechanisms that address the outputs of the processing, including inferred and derived data, profiles, and decisions, are far weaker (Wachter & Mittelstadt, 2019). On the other hand, it also means that physicians and/or health care providers must explain to patients the weight that automated decision-making and profiling represented in their final decision, and provide understandable explanations for why the automated decision-maker's suggestions were or were not followed. It might happen, indeed, that physicians have to confess that the only reason they followed the machine's advice is that they could simply find no justification to contradict its opaque conclusion. But, if this is the case, this information and no other should be shared with the patients. For this purpose, a flexible, functional approach will be most appropriate for understanding the term 'meaningful information' that is included in the right to an explanation (Selbst & Powles, 2017).

3.4. But… the issues that remain

The construction of an apparently sound legislative framework, such as the one we have described, will not, however, serve to address all the problems that the introduction of AI will bring to the management of health information. To begin with, it is difficult to know how it will be possible to reconcile a patient's right to restrict the use of his or her health data with increasingly automated health systems. If in the future most decisions are made on the basis of AI recommendations, patients who refuse to provide their data for that purpose will have to rely on physicians who will probably have lost some of the skills of traditional medicine. Thus, the configuration of the medicine of the future may end up dividing patients into two groups, those who are reconciled to the use of their data in AI systems and those who refuse to take this step. It is not clear what the consequences of this division will be, or whether we should start warning of these dangers right now.

From the doctor's point of view, the introduction of AI creates a growing challenge in terms of the concept of confidentiality and the fiduciary relationship between a patient and a physician. As Char et al. have written,

> In the era of electronic medical records, the traditional understanding of confidentiality requires that a physician withhold information from the medical record in order to truly keep it confidential. Once machinelearning–based decision support is integrated into clinical care, withholding information from electronic records will become increasingly difficult, since patients whose data aren't recorded can't benefit from machine-learning analyses. The implementation of machine-learning systems will therefore require a reimagining of confidentiality and other core tenets of professional ethics. What's more, a learning health care system will have agency, which will also need to be factored into ethical considerations surrounding patient care." (Char et al., 2018)

Third, even if it is convenient that physicians' skill sets include collaborating with and managing AI devices that aggregate big data (Wartman & Combs, 2019), one cannot ignore the fact that it will be hard for physicians to acquire all the technical capacities needed to provide accurate information about those devices directly. Therefore, taking care of these issues may take us into a highly undesirable scenario in which patients do not receive accurate information and physicians are stressed by the need to perform tasks they are not trained to perform. In our view, this could be prevented if we let physicians primarily communicate to the patient how the use of AI has influenced their diagnosis or choice of treatment, including the reasons that would have supported that conclusion. It would be better if health care providers could designate other professionals who are more familiar with AI to convey technical information about how AI works in each particular case. For this reason, the creation of new roles, such as that of health information counsellors (HICs) (Fiske, Buyx, & Prainsack, 2018), is of particular interest. These counsellors would be professionals with a broad knowledge of various kinds of health data and data quality evaluation techniques, as well as analytical skills in statistics and data interpretation, who could offer patients information about AI much more efficiently than health care givers. As Fiske et al. (2018) propose, *'trained also in interpersonal communication, health management, insurance systems, and medico-legal aspects of data privacy, HICs would know enough about clinical medicine to advise on the relevance of any kind of data for prevention, diagnosis, and treatment'*. Therefore, both patients and physicians would profit from the intervention of this new role.

Last, but not least, we must keep in mind that the right to information concerns not only what we have historically referred to as health data, but also what the GDPR calls data concerning health: *'personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status'* (Article 4.15). According to this definition, the concept extends to an increasing variety of data generated and collected outside the clinical setting, such as lifestyle data, data about dietary habits, socio-economic data, and data included in patients' health records, but also data collected through smartphones, direct-to-consumer testing, online platforms, apps, and wearables (Frisse, 2016). This means that the obligation to provide explanations may extend to data controllers who are not health care providers as such. If this is the case, we should design new policies regarding informed consent that apply to the use of these devices and deal with the obligations to which these providers are subject. Quite a number of tasks to perform there.

4. The general prohibition on fully automated individual decision-making

The General Data Protection Regulation has directly addressed its concern for decisions based solely on automated data processing, especially when it affects special categories of data, a concept which includes 'data concerning health' (article 9.1). In this sense, its Recital 71 states that

> The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her (…) Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her (…) In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and 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 (…) Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

The binding part of the Regulation reflects the intentions made in the Recital in its article 22, which is quite complex, but might be summarized by stating that *“the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her”* (article 22.1). This clause does not rule whether the decision is: (a) necessary for entering into, or performance of, a contract (b) authorised by Union or Member State law or (c) based on the data subject’s explicit consent. Additionally, where the automated processing is based on special categories of personal data, such as data concerning health, data subjects have to explicitly consent to the use of such data or processing needs to be justified by a substantial public interest, and the data controller must adopt suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests (article 22.4). As we have seen above, according to Recital no. 71 those measures include providing specific information to the data subject and 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.

A first crucial issue that seemed unclear is the nature of this right (Brkan, 2019). It could be understood both as a right to object, where automated decision-making is restricted only to cases in which the data subject actively objects, or as a prohibition, where data controllers will not be allowed to make automated decisions about a data subject until one of the legal requirements is met (Wachter et al., 2017). This understanding is crucial since it is in no way realistic to believe that there is effective control of personal information through consent – or objection in this specific case – and the rights that complement it (Cotino, 2017). Fortunately, this point has been addressed the Article 29 Data Protection Working Party, the guidelines on automated individual decision-making and profiling declared *that a general prohibition on this type of processing exists to reflect the potential risks to individuals' rights and freedoms*.

Therefore, unless we met a legal exception to the general prohibition, there is no room for solely automated decision-making in the EU zone if it *produces legal effects concerning him or her or similarly significantly affects him or her*. This seems to be the case of most of the clinical decisions, even if it remains arguable what kind of decision-making significantly affects such an individual (Brkan, 2019). The risk categorization framework proposed by the FDA for the use of AI systems (U.S. Food & Drug Administration, 2019), based on the state of healthcare situation or condition of the patient and the significance of the information provided by the system to the healthcare decision, might be useful in this scenario.

However, the Regulation also fails to make clear what counts as a decision based solely on automated decision-making. Indeed, one cannot deduce from the literacy of the clauses what kind of human intervention is needed to make the difference between automated and solely automated decision-making. The only concretion made by the Regulation is that *"the controller must put in place suitable measures to safeguard the data subject's rights and freedoms and legitimate interests"*. This, definitively, does not provide too much concretion (Zarsky, 2017), what is certainly worrying, particularly if the condition 'solely' in 'decisions made solely by automation' is interpreted narrowly, because the safeguards and associated requirements of meaningful information will have limited applicability (Andrew et al., 2017).

Once again, the Article 29 Data Protection Working Party has provided some clarifications on what should be understood by solely automated decision-making,

> "The controller cannot avoid the Article 22 provisions by fabricating human involvement. For example, if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing. To qualify as human intervention, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the available input and output data." (A29WP, 2018)

The current state of the art (Yu et al., 2018) suggests that, for the moment, fully automated clinical decision systems, which could be understood as solely automated decision systems, are scarce in comparison with integrative decision support systems, where clinicians still need to make the final decision and, therefore, they are invested with authority and competence to change the algorithmic decision. Of course, neither the Regulation, nor the Statement by the Working Party focus specifically on the health care arena (indeed, this last document only mentions health care marginally). At first sight, it is perfectly clear that patients are entitled to claim for the intervention of a human being in the process of decision-making, but although the intervention of a human with authority and capability to change the decision may be legally appropriate and societally desirable, it might present enormous difficulties in practice (Brkan, 2019). This approach does not serve us well to specify what kind of obligations physicians and caregivers who play the role of data controllers will have to assume. Limitations are both relevant for fully automated clinical decision-making systems, where human intervention is claimed, and for decisions based on integrative decision support systems, where human intervention remains in control. Moreover, we have to focus on the kind of practical consequences this system will bring and which dynamics might arise within the health caregiving community.

In our opinion, this is quite difficult to determine, since circumstances can change substantially depending on the time of care – diagnosis or therapy – or even depending on the circumstances of each specific case. To begin with, it is obvious that a physician must, in any case, supervise that the diagnosis or treatment recommendation provided by the AI does not blatantly contradict what medical science has been able to determine in a well-known situation. Thus, for example, if the AI recommends a treatment or a dose of medicine that would surely cause unnecessary harm or even death to a patient, physicians must be able to detect it and impose their judgement on the machine (and report the failure to improve the system, obviously).

Much more complex is determining what to do in cases where a machine recommendation challenges what we might call its intuitions. In these circumstances, we face a complex dilemma. On the one hand, if we concede that it has to be the medical criterion, we would be largely denying one of the bases that justify the use of AI: its ability to make a diagnosis or recommend treatment more efficiently than a human in unclear circumstances. On the other hand, it seems complex to force physicians to act against their own inclinations. Another relevant point in the evaluations of those dynamics is how automation may reinforce the (mal)practice of 'defensive medicine' (Perin, 2019). Surely, the solutions to these dilemmas can only be traced by leaving the final decision in the hands of a patient who has been adequately informed of the circumstances at hand. This will include, of course, the possible consequences of an error in the suggestion of the machine or in human intuition. Once again, it seems necessary to resort to some kind of advice that goes beyond that which can be provided by the doctor who is directly involved in the dilemma. And once again it seems recommendable to introduce the figure of the Health Information Counsellor in the equation.

References

Afzal Hussain Shahid, A. H. & Singh, M. P. (2019). Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments. Biocybernetics and Biomedical Engineering (in press). https://doi.org/10.1016/j.bbe.2019.05.010

A29WP WP251. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. European Commission, 22 August, 2018.

Beer, D. (2017). The social power of algorithms. Information Communication and Society, 20(1), 1-13.

Bini, S.A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? The Journal of Arthroplasty, 33, 2358-2361.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification∗ Joy. Proceedings of Machine Learning Research, 81, 1–15.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care: Addressing Ethical Challenges. New England Journal of Medicine, 378, 981–983.

Coeira, E., (2019). On algorithms, machines, and medicine. The Lancet Oncology, 20(2), 166–167.

Cotino, L. (2017). Big Data and Artificial Intelligence. An Approach from a Legal Point of View about Fundamental Rights. Dilemata, 24, 131–150.

de Miguel Beriain, I. (2018). Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling. Law, Probability and Risk, 17, 45–53.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, … Wood, A. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134, 2017.

Ellis, G. & J. Silk (2014). Defend the integrity of physics, Nature, 516, 321-323.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M.A., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80-89.

Ferretti, A., Schneider, M., & Blasimme, A. (2018). Machine Learning in Medicine: European Data Protection. Law Review, 4, 320 – 332.

Fiske, A., Buyx, A., & Prainsack, B. (2018). Health Information Counselors: A New Profession for the Age of Big Data. Academic Medicine : Journal of the Association of American Medical Colleges, 94, 37–41.

Frisse, M.E. (2016). The business of trust. Academic Medicine, 91: 462–464.

Goodman, B. & Flaxman, S. (2017). European Union Regulations on Algorithmic DecisionMaking and a "Right to Explanation". AI Magazine, 38, 50–57.

Hamet, P. & Tremblay, J.. (2017) Artificial intelligence in medicine. Metabolism, 69, S36–S40.

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), 900–915.

Kallianos, K., Mongan J., Antani, S., Henry, T., Taylor, A., J.Abuya, J.& Kohli, M. (2019). How far have we come? Artificial intelligence for chest radiograph interpretation. Clinical Radiology, 74, 338-345.

Kemper, J., & Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. Information, Communication & Society, Published online: 18 Jun 2018, 1–16. https://doi.org/10.1080/1369118X.2018.1477967

Le, E. P. V., Wang, Y., Huang, Y., Hickman S. & Gilbert F. J. (2019). Artificial intelligence in breast imaging. Clinical Radiology, 74, 357-366.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

Londhe, V. Y. & Bhasin, B. (2019) Artificial intelligence and its potential in oncology. Drug Discovery Today, 24, 1228-232.

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? EMBO Reports, 16(10), 1250–1255.

Minssen, T., & Pierce, J. (2018). Big Data and Intellectual Property Rights in the Health and Life Sciences. In I. Cohen, H. Lynch, E. Vayena, & U. Gasser (Eds.), Big

Data, Health Law, and Bioethics (pp. 311-323). Cambridge: Cambridge University Press.

Ngiam, K. Y. & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. Lancet Oncology, 20: e262–73.

Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. (2019) Digital pathology and artificial intelligence. Lancet Oncol, 20: e253–61.

Noorbakhsh-Sabet, N., Zand, R., Zhang, Y. & Abedi, V. (2019). Artificial Intelligence Transforms the Future of Health Care. The American Journal of Medicine (in press). https://doi.org/10.1016/j.amjmed.2019.01.017

Nuñez-Reiz, A. (2019). Big data and machine learning in critical care: Opportunities for collaborative research. Medicina Intensiva, 43, 52-57.

O'neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. (Random House LCC US, Ed.).

Perin, A. (2019). Standardizzazione, automazione e responsabilità medica. Dalle recenti riforme alla definizione di un modello d'imputazione solidaristico e liberale. Biolaw Journal - Rivista Di BioDiritto, (1), 207–235.

Schmidt-Erfurth, U., Sadeghipour, A., Bianca S. Gerendas, B. S., Waldstein, S. M. & Bogunović H. (2018) Artificial intelligence in retina. Progress in Retinal and Eye Research, 67, 1-29.

Scott, I. A. (2018) Machine Learning and Evidence-Based Medicine. Annals of Internal Medicine, 169, 44-46.

Selbst, A. D. & Powles, J. (2017). Meaningful Information and the Right to Explanation. International Data Privacy Law, 7, 233–242.

Sterky, F. & Lundeberg, J. (2000). Sequence analysis of genes and genomes. Journal of Biotechnology 76, 1-31.

U.S Food & Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. Published online: April 2, 2019. Retrieved from: https://www.fda.gov/media/122535/download

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7, 76–99.

Wartman, S. A., & Combs, C. D. (2019). Reimagining Medical Education in the Age of AI. AMA Journal of Ethics, 21(2), 146–152.

Wellner, G., & Rothman, T. (2019). Feminist AI: Can We Expect Our AI Systems to Become Feminist? Philosophy & Technology. https://doi.org/10.1007/s13347-019-00352-z

Yu, K.-H., Beam, A. L. & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2, 719–731.

Zarsky, T. (2017). Incompatible: The GDPR in the Age of Big Data (2017) Seton Hall Law Review, 47, 995-1019.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philosophy & Technology. https://doi.org/10.1007/s13347-018-0330-6

Acknowledgements:

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 788039