Consequences of unexplainable machine learning for the notions of a trusted doctor and patient autonomy

Michal KLINCEWICZ [a,1], and Lily FRANK [b]
[a] *Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands*
[b] *Ethics and Philosophy, Technical University of Eindhoven, The Netherlands*

Abstract. This paper provides an analysis of the way in which two foundational principles of medical ethics–the trusted doctor and patient autonomy–can be undermined by the use of machine learning (ML) algorithms and addresses its legal significance. This paper can be a guide to both health care providers and other stakeholders about how anticipate and in some cases mitigate ethical conflicts caused by the use of ML in healthcare. It can also be read as a road map as to what needs to be done to achieve an acceptable level of explainability in an ML algorithm when it is used in a healthcare context.

Keywords. machine learning, explainability, health care, ethics

Introduction

Machine Learning (ML) is used here to refer to a class of statistical models primarily used to yield repeatable and accurate predictions [1]. ML models result from 'learning', broadly understood, on large amounts of data. This learning process determines what the model will be able to predict. Given this, a ML model becomes the basis for predictions about features that were in the data from which it 'learned'. For example, a ML model that is trained with cardiograms can 'learn' to predict which cardiogram is associated with heart disease, but may not 'learn' what a healthy heart rate is.

ML can be used in any domain of inquiry where large amounts of data can be found, including, but not limited to, all aspects of healthcare. This development can be attributed to two independent factors: an increase in the availability of large health-related datasets, on the one hand, and a decrease in the expense of the computationally intensive learning process, on the other. As central processing units in computers become cheaper and faster, it takes less time and energy to generate a useful and accurate ML model.

Non-ML statistical modelling techniques used in healthcare can and often are used to enable interpretations of data and to provide a basis for causal inference. The main reason for this is that researchers can validate their interpretations of data by, for exam-

[1]Corresponding Author: m.w.klincewicz@uvt.nl

ple, checking for statistical significance and then comparing their results with accepted practice in their field. For example, a correlational model of cardiograms of people with and without heart disease may tell researchers which features are relevant and significant in contributing to its predictions. This is typically not something we can or even want to do with ML models. The notion of statistical significance has little place in making sense of why a particular ML model makes predictions the way it does.

In sum, ML is limited to those domains that can provide an adequately large and rich dataset. The use of ML, however, comes with a trade-off. Typically an increased accuracy of prediction coincides with a simultaneous increase in the opaqueness of the factors that come to play a role in that accuracy. This means that robust ML models are unlikely to tell us much about the factors, variables, patterns, and relationships that are responsible for the predictions that the model makes. There is ongoing research [2], [3] into making ML models interpretable and explainable, which, if successful, could be a straightforward way to resolve the trade-off. Until that time comes, however, this particular feature of ML raises significant ethical and legal concerns within the healthcare domain, especially in contexts where transparency and explainability play a foundational role.

In this paper, we focus on two moral foundations that are especially important in the healthcare domain: the position of *the trusted doctor* and *respect for patient autonomy*. We focus on these two issues because they connect most obviously to established legal frameworks within which healthcare professionals typically operate. Since obligations of medicine are importantly distinct from the obligations of everyday morality, we ignore the more general issues connected to the explainability of ML. In section 1 we provide a brief review of the legal and ethical foundations of the notions of a trusted doctor and patient autonomy. In section 2 we take up the moral and legal difficulties that ML generates for the trust in doctors and patient autonomy. In section 3 we sketch some of the methods that can be used to mitigate the problems we discuss in section 2.

1. The Reasonable Physician and Patient Standards in the Law and Medical Ethics

Idealizations play an important role in common law. For example, when the issue under consideration is someone's intent, one way in which the law sees it being determined is by examining the understanding of an idealized reasonable person. During this process, consideration is given to all relevant circumstances of the case to determine what a reasonable person would intend or do in these circumstances. Similar idealizations are used in the medical context to determine whether there was informed consent and to ascertain liabilty. It is an issue of ongoing debate in the ethics and law of medicine precisely which pieces of information must be disclosed to patients in order to fulfill this commitment to truth-telling and subsequently to avoid allegations of medical malpractice [4]–[6]. Some matters are uncontroversial, for example, the most common and most serious risks of undergoing or forgoing a treatment must be disclosed and explained. But it is practically impossible and probably ethically undesirable to disclose all relevant information to patients before they make a medical decision.

In the U.S. two different legal and ethical standards dictate which information must be disclosed to a patient in order for them to give informed consent to medical interventions or participate in clinical research: the reasonable physician (medical practice

or professional) standard and the reasonable patient (or person) standard [7], [8]. These two standards provide different answers to the question: 'what information must be provided to a patient before they are capable of giving truly informed consent?' The reasonable physician standard answers it by referring to the broadly accepted professional standards and practices relevant to the specific context. The reasonable patient standard answers it by referencing what an average patient would want to know, find relevant to their decision making, or be expected to be informed about.

In the context of informed consent to interventions or treatments, the reasonable physician and reasonable patient standards create distinct challenges. The reasonable patient standard generates problems connected to it being vague, since it assumes that patients across demographic groups are sufficiently similar. This notion does not rely on empirical evidence regarding patient preferences or expectations, so it is sometimes difficult to imagine what the idealized person would want or expect to know about their diagnosis. The problems with the reasonable physician standard are slightly different. The standards of medical practice and thus what can be expected from a reasonable physician is not constant over time or place, so what is expected of a physician is highly dependant on context.

For example, in the United States for at least four decades there was professional consensus that, with rare exceptions, competent patients must be informed of their diagnosis, if they wish to be. But a study in 1961 revealed that a vast majority of physicians routinely witheld cancer diagnosis from patients [9]. Practice does not always follow the standards set out by idealizations. A second shortcoming of this standard is that even a high percentage of professional physicians can be mistaken, biased, or unaware, when it comes to information that is relevant to patient decision making. This is unfortunate, since the physician has to determine which information to provide to patients before asking them to decide on treatment or care. It is in these sorts of situations that the notion of a reasonable physician is used when U.S. courts need to determine whether the physician satisfied standards of informed consent. Informed consent, which we turn to later, is crucial to determine negligence and malpractice. A physician that did not abide by the standard of informed consent may be liable.

The reasonable physician standard is used differently in other common law traditions, such as U.K. and Canada, and significantly different in civil law contexts. For example, in Canada (except Quebec), doctors are legally required to answer all questions posed by patients, including about benefits, risks, and treatment alternatives. In Australia, the professional does not incur a liability in negligence, if it is established that "the professional acted in a manner that at the time the service was provided was widely accepted by peer professional opinion as competent professional practice" (Civil Liberties Act of 2002, Section 5O). In the U.K., these issues are typically resolved be referring the tort standard embodied in the Bolam Standard. On this standard, "a doctor is not guilty of negligence if he has acted in accordance with a practice accepted as proper by a responsible body of medical men skilled in that particular art" (Bolam v Friern Hospital Management Committee [1957] 1 WLR 583). All of these common law standards are, to some extent, idealizations of what a reasonable physician is expected to do, even though their practical and legal bases are different. In the civil law tradition, we have examples such as the *la responsabilite civile* (Code of Ethics of Physicians) in France, which is unequivocal in demanding that "a doctor must in all circumstances be trustworthy and act with integrity and devotion to duty, essential for the practice of medicine. Confidentiality

is a patient's right. It is mandatory for all doctors as required by law" (Articles R.4127-3 and R.4127-4).

Zooming out from nation-specific legal instruments to the international level, the medical profession has distinct mention in the 1948-2017 Declaration of Geneva, in the 1964 Declaration of Helsinki, as well in the 1997 Council of Europe Convention on Human Rights and Biomedicine, among others. Arguably, these international instruments, like their nation-specific counterparts, embody at least in spirit two closely related foundational principles of medical ethics: *the trusted doctor* [10], [11] and respect for *patient autonomy*, which is closely related to truth-telling and informed consent [12], [13]. Violating these two ethical principles will be, in many cases, also a violation of related nation-specific and international legal standards that protect medical professionals and patients. Therefore, focusing on the way in which ML will affect these two ethical principles is a way of addressing possible legal consequences, without focusing on nation-specific and international similarities and differences across legal contexts.

The trusted doctor principle is grounded in the claim that medicine has a distinct set of moral responsibilities and physicians have a fiduciary duty to their patients [10], [14]. For example, a doctor may be free to allocate attention and empathy to only those people in their circle of intimates for whom they care, but in a healthcare setting they are required to put personal preferences aside and allocate attention and care on the basis of considerations like medical need and urgency [10], [14]. This centers the physician's obligation to earn trust and be trustworthy, as they are given a special set of rights and privileges in society. In general, the medical notion of a patient's trust in their doctor can be understood as:

> ... an attitude of willingness to rely on another person or entity to perform actions that benefit or protect oneself or one's interests in a given sphere of activity, together with a normative expectation: the person or entity should perform in a particular way[15, p. 355].

Physicians are obligated to provide care in accordance with the principle of beneficence, that is, to act for the benefit of the patient. Simultaneously, they are expected and obligated to act in accordance with a wide range of other moral and professional commitments, such as the commitment to staying up to date with respect to scientific developments in their field, transparency about the limitations of their expertise, respect for patient confidentiality, truth telling, and respect for patient autonomy and informed consent, broadly construed.

The attitude of trust that many patients have in their physicians cannot be taken for granted and is mediated by several factors. Research shows that patient characteristics like race and socio-economic status impact their trust in physicians (c.f. Kennedy, Mathis and Woods 2007 on urban African Americans trust in the health care system) as can characteristics of the physician or the institutions they are embedded in. Patients who believe that their physicians are being compensated based on the number of medical tests they request or prescriptions they write in a managed care system are (unsurprisingly) seen as less trustworthy [16]. The introduction of new technologies in medical practice is not by any means a novel phenomenon and how new technologies impact patient trust is a perennial issue [17]. For example, Promberger and Baron [18] found that patients have greater trust in and are more likely to follow medical recommendations provided to them by a physician rather than by a computer.

The other foundational moral commitment of medicine that is embodied in legal instruments and that we discuss here is *patient autonomy*, which includes the interrelated ethical principles of truth-telling and informed consent [12], [13]. Informed consent is central to many of legal instruments discussed already, but it also has a special place within medical ethics, especially when coupled with respect for patient autonomy on which it arguably depends. One way in which requirement of respect for autonomy or, more broadly, respect for persons, is operationalized in the medical context is through the requirement that physicians obtain consent or refusal for any medical intervention they consider from the patient, assuming that the patient has decisional capacity. The closely related legal concept of patient competence is determined on the basis of four criterion [19], [20]. When making a medical decision patients must be able 1) "communicate a choice;" 2) "Understand the relevant information;" 3) "Appreciate the situation and its consequences;" and 4) "Reason about treatment options" [19, p. 1836]. In order for patients to be able to demonstrate these capacities physicians must provide them with the relevant information in a form that the patient is able to understand and then follow up with an assessment of their understanding through the use of specific questions, such as: "Why do you think your doctor has [or I have] recommended this treatment?" (Ibid p. 1836).

Respect for autonomy requires that patients have the opportunity to make their own medical decisions which are consistent with their own values, preferences, and understanding of a good life, even when these decisions may conflict with what others, including the medical team, see as in their best interests. Given this, most major medical interventions require that the patient be given adequate and truthful information about the risks, benefits, and alternative treatments and be given opportunity to discuss and ask questions about their treatment. This means that the requirements of truth-telling and informed consent are to some extent derivative from the requirement of respect for individual autonomy–hence the aforementioned inter-relatedness of these principles. Truth-telling in medicine makes it possible for patients to be able to make their own decisions about matters of their health care. Deception in medicine is a form of expressing a lack of respect for the rationality and autonomy of the patient, which interferes with a patient's ability to exercise his or her decisional capacity.

2. The Effect of ML on the Trusted Doctor and Patient Autonomy Principles

A crucial question for the ethical use of ML in medicine is whether or not patients' attitude of trust will be undermined as it becomes difficult or impossible to explain to the patient or their family member what lies behind a diagnosis or recommendation of course of treatment. And once this question is answered, we also need to answer a follow-up question: Will the special set of rights and privileges that medical professionals are endowed with on the basis of that trust appear unwarranted from the patient's perspective as a result? Two considerations suggest that such a situation is likely and that the use of ML in medicine may undermine patient trust. The first has to do with responsibility/explainability and the second to do with perceived objectivity/bias.

To maintain trust physicians must be able to unpack their diagnoses and recommendations in lay person's terms and create a shared understanding of the medical facts. This allows patients to make informed and autonomous decisions about the course of their

care. Simultaneously, it is during this process that patient and physician take shared responsibility for the course of care [21]. In order for trust to be maintained physicians must be able to explain the role that ML-models or algorithms played in the diagnosis or recommendation. They must also communicate and be justified in communicating that the doctor is ultimately responsible for a patient's evaluation or care, despite the role of a black-box algorithm. If this is not possible, in time the trust that governs the doctor-patient relationship will be undermined and the expectations that patients have of their doctors will be changed. There is now significant evidence that trust is indeed undermined by computer systems in the medical context, if such explanations are not provided [22]. ML-models will similarly be unable aid patients in a way that keeps them informed in the ethically significant sense.

Second, physicians are expected to treat their patients with nonjudgmental regard and in a manner free from personal bias [23]. This is not easily accomplished by physicians and there is significant evidence that they fail to live up to this obligation when treating, for example, patients with eating disorders [24] or patients who are obese [25]. Although physicians are imperfect in setting aside biases and personal preferences when delivering care, they are bound by a moral duty to strive to do so. If physicians fail to prevent their biases from impacting their perceptions and treatment of their patients, the expanded use of diagnostic and treatment recommending technology seems like an appealing way to ameliorate this problem. Machines are, after all, objective, their results free from interpretation and associated human frailties, one might think.

This sort of techno-optimism is potentially problematic because, somewhat famously now, ML algorithms can themselves become biased in a variety of ways [26]–[28]. We already now know cases of ML algorithms in healthcare that turned out to be biased [29]–[31]. From the perspective of clinical justice this is a problem in and of itself that is likely to compound preexisting physician biases, rather than counteract them. As the existence of bias in ML-aided healthcare becomes widely known there is a further risk that patient trust in its recommendations will also be undermined. This presents a troubling dilemma for the project of maintaining the trusted doctor standards. To disclose to patients the extent to which ML-aided healthcare is subject to bias undermines trust in the system as a whole, but to fail to disclose these limitations may violate the obligations of truth telling and robust informed consent.

The legal consequences of undermining *the trusted doctor* standard must be carefully considered and this is outside the scope of the present article. Regarless, we can here at least focus on the legal basis for the final word in diagnosis and treatment recommendations, which lies with medical professionals precisely because of the privileged epistemic and moral position that they have within that domain. When the epistemic and moral bases for that position are undermined, we can expect the legal basis to be similarly undermined in time. Someone that does not or cannot fulfill an obligation to do something, eventually is relieved of that obligation, all things being equal. This in turn would pose a fundamental challenge to medical moral responsibility and to the way that the legal system deals with malpractice, patient death, and liability in cases of disagreement among medical professionals. In all of these cases, the medical professional's authority and protection under law will be diminished or disappear altogether, as a result of the diminishing of the expectations on their performing their duty to explain things to patients. The remaining question would be then to assign responsibility to someone when things go awry.

It is worthwhile noting here that there are related discussions of the ways in which the introduction of ML into new spheres of human activity (e.g. self driving cars, surgical robots, or automated loan eligibility assessment) impacts responsibility attribution [32]–[35]. Troubles with assigning responsibility in a world full of automation and ML is not unique to healthcare. In those other domains the question of who to blame when things go awry is far from being solved. We can also expect medical professionals to be put in an uncomfortable position to have to justify their diagnoses or treatment decisions in cases where they themselves cannot state reasons or explain the performance of an ML algorithm. At best, medical professionals will have to offer *post hoc* rationalizations that the performance of the ML model is in line with what they would have decided independently as an appropriate course of action themselves.

Similarly to the requirement of *the trusted doctor*, it is difficult to fulfill the *patient autonomy* requirement, if the answers to questions about treatment or diagnosis are in principle difficult or altogether impossible to give. To see this consider a medical professional that answers questions about the risks and benefits of treatment with only predictions of success or failure, rather than with reasons why these predictions are as low or high as they are. A patient that asks for such reasons and does not get answers will not be in a position to have informed consent to the treatment. Similarly, a medical professional that fails to answer questions about alternative treatments would be failing to respect patient autonomy. That choice is effectively not given to the patient. Finally, a medical professional that cannot or will not discuss possible manners of treatment, but merely provides a recommendation, will be violating the requirement of truth-telling. An opaque ML-aided algorithm that recommends or diagnoses in healthcare will be just like that medical professional, unable to provide answers to questions about risks and benefits, reasons for predictions, or alternative treatments. This situation threatens respect for patient autonomy by effectively removing a patient's decisional capacity from the calculus that determines courses of treatment and care.

Removing patient decisions from that calculus can be legally significant. In Canada, for example, doctors are legally required to answer all questions posed by patients, including about benefits, risks, and treatment alternatives. Similar laws can be found in the European Union and the United States. It is simply not clear how these legal requirements of informed consent doctrines can be met in good faith when ML-aided healthcare interventions or diagnoses are involved. Medical professionals cannot be expected to understand the operation of an ML-model when they are in principle opaque, even to the computer scientists that may be 'teaching' them to recognize patterns. Again, at best, medical professionals can offer *post hoc* rationalizations of the operation of the ML-model, assuming that it is doing what they would do, but without ever knowing that it actually does so.

The legal consequences of undermining patient autonomy, as with the trusted doctor standard, are likely to be profound. The professional and legal obligations that doctors have with respect to patients will likely come under pressure. In a legal context, a doctor's answer to a patient's question that cites the opaqueness of an ML model may be in violation of the requirements of informed consent, effectively putting in question a diagnosis that is responsible for a possible later mistake or simply by ignoring what a patient may find particularly important in their medical situation. Again, an answer that simply assumes that the ML model is doing what a doctor would do is a misrepresentation that may itself be in violation of the legal standard for evidence. Evidence that, say, a doctor

provided informed consent in a way 'I-know-not-how' is setting the bar for admissibility extremely low.

To sum up, the two foundational principles of medical ethics, *the trusted doctor* and *patient autonomy*, are undermined by the opaqueness of ML-aided medicine. If these principles are undermined, then we can expect significant downstream conflicts with international and nation-specific legal standards that govern the the medical profession. In particular, the legal basis for informed consent, liability, and standard of malpractice will be out of sync with the reality of day-to-day doctor-patient interactions. Physicians will not be in a position to live up to the standards of the trusted doctor standard that is required of them to secure authority in diagnosis and consent. Furthermore, patients will have their autonomy challenged by physicians that are either unable to tell them the actual basis for diagnoses or treatment recommendations or will outright misinform them about that basis. The remaining question is how to deal with these consequences, especially since ML-aided medicine is already here.

3. Three Ways Forward for Machine Learning in Healthcare

Using ML in a medical setting is bound to have a positive effect in a number of areas, including effectiveness of diagnostics, decreases in cost, and better resource management. These positives are in danger of being outweighed by the negative consequences that arise from the nature of ML technology and the downstream effects of its widespread use. Among these is the already mentioned lack of transparency inherent in ML models, but also the use of massive amounts of private data to 'teach' ML models, and finally the possible introduction of biases into the predictions that ML models make. Here we discuss three ways of resolving problems caused by the lack of transparency, without addressing these other potential problems: (1) saliency methods, (2) limiting the role of ML to specific domains of healthcare where its lack of explainability does not undermine ethical foundations or legal norms, or (3) changing the reasonable patient and reasonable physician standards. This list is not meant to be exhaustive.

1) Saliency methods analyze a ML model that has already 'learned' to recognize patterns in an image by piecemeal subtracting parts of an image until the part most relevant to the classification and/or prediction is found. This process can be repeated to obtain a ranking of parts of an image that can then comprise something very close to an explanation as to why the image was classified in the way that it was. At that point, medical professionals can also provide something like reasons to the diagnosis that are based on the ranking, if asked to do so by the patient. Saliency methods for generating explainable ML models may work whenever images are involved. This is by no means exhaustive of the possible ways in which ML diagnostics can be made explainable, but is the one that most obviously connects to healthcare diagnostics that use images.

There are two problems with the saliency approach. First, it is not generalizable to all areas of healthcare and specifically to those that do not rely on images. Cardiograms, X-rays, MRIs, or just photographs may all be essential for diagnosis, but they are not always relevant to the recommendations that a medical professional ultimately makes with respect to treatment or care. There are also a variety of diagnostics that do not rely on imagining. Second, saliency has recently come under pressure as a method of explaining the performance of ML [36]. It turns out that that two distinct ML models may

perform identically under the same conditions, which would likely generate disparate explanations via saliency for the same performance. What this means, is that explanations via saliency can be a dime a dozen, depending on the model that happens to be used– not something that a patient or doctor are likely to accept as an accaptable standard of explainability.

(2) The most drastic option and perhaps also the easiest way to deal with the problems of ML in context of healthcare is to advocate for strict legal regulations both at a national level and internationally. In cases where the use of ML may directly undermine the doctor-patient relationship or undermine legal standards for informed consent, it should not be used, full stop. One good example of a context where such a limitation may be particularly important is in diagnostic algorithms that take disparate data about a large number of people and construct a model that can predict the presence of early stages of a chronic degenerative disease, such as Alzheimer's Disease or multiple sclerosis. While undoubtedly such a tool could be useful in a non-clinical setting to nudge people to seek physician-assisted diagnosis, it should under no circumstances be used instead of such a diagnosis or in conjunction with it. Doing so opens up a slew of moral and legal difficulties, not limited to those outlined in the two sections above. The main problem of the regulative approach is that enforcement of laws that regulate ML-aided medicine is likely to be extremely difficult internationally, especially in an era of medical tourism.

3) The third option is to reassess and change the reasonable patient and reasonable doctor standards in such a way that the idealization that are based on them are sensitive to the advent of ML-aided medicine. We cannot offer speculations about what these changes could entail–this is work for legal scholars working within specific legal contexts. However, this article sketches what we hope are useful guidelines and framing conditions that such speculations could follow. First, the reasonable doctor cannot be expected to explain the actual basis of an ML-aided advice and a reasonable patient cannot expect to receive such information from their physician. Second, attempts by physicians to justify ML-based advice or diagnoses on the basis of what they assume they would have done themselves are dangerous and should be avoided. Thirdly, patients and physicians should be vigilant about the potential biases that are at the heart of the diagnoses and advice provided by ML-aided medicine. Any reasonable idealization that takes into account the trusted doctor standard and patient autonomy should have room for flagging potential hidden biases that drive ML-aided medicine, in lieu of evidence to the contrary.

4. Acknowledgement

This paper was partially financed by the Polish National Science Centre (NCN) SONATA 9 Grant, PSP: K/PBD/000139 under decision UMO-2015/17/D/HS1/01705.

References

[1] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nature Methods*, 2018, ISSN: 1548-7091. DOI: 10.1038/nmeth.4642.

[2] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 2018, ISBN: 9781538653777. DOI: 10.1109/ICHI.2018.00095.

[3] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning," *arXiv:1806.00069*, 2018, ISSN: 00224936. DOI: arXiv:1806.00069v2.

[4] D. J. Mazur, "What should patients be told prior to a medical procedure? Ethical and legal perspectives on medical informed consent," *The American Journal of Medicine*, 1986, ISSN: 00029343. DOI: 10.1016/0002-9343(86)90405-5.

[5] B. Murray, "Informed consent: what must a physician disclose to a patient?" *AMA Journal of Ethics*, vol. 14, no. 7, pp. 563–566, 2012.

[6] A.-E. Ciortea, "What Medical Risks Should Physicians Disclose to their Patients? Towards a Better Standard in American and French Medical Malpractice Law," *Journal of Civil Law Studies*, vol. 10, no. 1, p. 9, 2018.

[7] R. R. Faden, C. Lewis, C. Becker, A. I. Faden, and J. Freeman, "Disclosure standards and informed consent," *Journal of health politics, policy and law*, vol. 6, no. 2, pp. 255–284, 1981.

[8] J. Greenblum and R. Hubbard, "The common rule's 'reasonable person' standard for informed consent," *Bioethics*, 2019, ISSN: 14678519. DOI: 10.1111/bioe.12544.

[9] D. Oken, "What to tell cancer patients: a study of medical attitudes," *Jama*, vol. 175, no. 13, pp. 1120–1128, 1961.

[10] R. Rhodes, *Understanding the trusted doctor and constructing a theory of bioethics*, 2001. DOI: 10.1023/A:1014430208720.

[11] P. Illingworth, "Trust: The Scarcest of Medical Resources," *The Journal of Medicine and Philosophy*, 2002, ISSN: 0360-5310. DOI: 10.1076/jmep.27.1.31.2969.

[12] R. R. Faden and T. L. Beauchamp, *A history and theory of informed consent*. Oxford University Press, 1986.

[13] T. L. Beauchamp and J. F. Childless, "Principles of Biomedical Ethics," *International Clinical Psychopharmacology*, 1991, ISSN: 0268-1315. DOI: 10.1097/00004850-199100620-00013.

[14] R. Rhodes, "Clinical justice guiding medical allocations," *The American Journal of Bioethics*, vol. 4, no. 3, pp. 116–119, 2004.

[15] P. J. Nickel, "Ethics in e-trust and e-trustworthiness: the case of direct computer-patient interfaces," *Ethics and information technology*, vol. 13, no. 4, pp. 355–363, 2011.

[16] A. C. Kao, D. C. Green, N. A. Davis, J. P. Koplan, and P. D. Cleary, "Patients' trust in their physicians: Effects of choice, continuity, and payment method," *Journal of General Internal Medicine*, 1998, ISSN: 08848734. DOI: 10.1046/j.1525-1497.1998.00204.x.

[17] P. Nickel and L. Frank, "Trust in Medicine," in *The Routledge Handbook of Trust and Philosophy*, Routledge, 2020.

[18] M. Promberger and J. Baron, "Do patients trust computers?" *Journal of Behavioral Decision Making*, 2006, ISSN: 10990771. DOI: 10.1002/bdm.542.

[19] P. S. Appelbaum, *Assessment of patients' competence to consent to treatment*, 2007. DOI: 10.1056/NEJMcp074045.

[20] T. Grisso, A. Grisso, and P. S. Appelbaum, *Assessing competence to consent to treatment: A guide for physicians and other health professionals*. Oxford University Press, USA, 1998.

[21] A. Edwards and G. Elwyn, *Shared decision-making in health care: Achieving evidence-based patient choice*. Oxford University Press, 2009.

[22] T. Wangmo, M. Lipps, R. W. Kressig, and M. Ienca, "Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders," *BMC Medical Ethics*, vol. 20, no. 1, p. 98, 2019, ISSN: 1472-6939. DOI: 10.1186/s12910-019-0437-z. [Online]. Available: https://doi.org/10.1186/s12910-019-0437-z.

[23] R. Rhodes *et al.*, "The professional responsibilities of medicine," *The Blackwell guide to medical ethics. Oxford: Blackwell*, pp. 71–87, 2007.

[24] J. Fleming and G. I. Szmukler, "Attitudes of medical professionals towards patients with eating disorders," *Australasian Psychiatry*, 1992, ISSN: 10398562. DOI: 10.3109/00048679209072067.

[25] H. J. Wiese, J. F. Wilson, R. A. Jones, and M. Neises, "Obesity stigma reduction in medical students," *International Journal of Obesity*, 1992, ISSN: 03070565.

[26] M. Garcia, "Racist in the machine: The disturbing implications of algorithmic bias," *World Policy Journal*, 2016, ISSN: 19360924. DOI: 10.1215/07402775-3813015.

[27] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[28] O. Osoba and W. Welser, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. 2017. DOI: 10.7249/rr1744.

[29] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, *Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data*, 2018. DOI: 10.1001/jamainternmed.2018.3763.

[30] D. S. Char, N. H. Shah, and D. Magnus, *Implementing machine learning in health care ' addressing ethical challenges*, 2018. DOI: 10.1056/NEJMp1714229.

[31] A. Yapo and J. Weiss, "Ethical Implications of Bias in Machine Learning," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018. DOI: 10.24251/hicss.2018.668.

[32] D. J. Gunkel, "Mind the gap: responsible robotics and the problem of responsibility," *Ethics and Information Technology*, pp. 1–14, 2017.

[33] R. de Jong, "The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm," *Science and Engineering Ethics*, 2019, ISSN: 14715546. DOI: 10.1007/s11948-019-00120-4.

[34] S. Köhler, N. Roughley, and H. Sauer, "Technologically blurred accountability?: Technology, responsibility gaps and the robustness of our everyday conceptual scheme," in *Moral Agency and the Politics of Responsibility*, Routledge, 2017, pp. 51–68.

[35] S. Nyholm, "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci," *Science and Engineering Ethics*, 2018, ISSN: 14715546. DOI: 10.1007/s11948-017-9943-x.

[36] P. J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (Un)reliability of Saliency Methods," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. DOI: 10.1007/978-3-030-28954-6{_}14.