The Jean Monnet Center for International and Regional Economic Law & Justice

THE NYU INSTITUTES ON THE PARK

THE JEAN MONNET PROGRAM

J.H.H. Weiler, Director
Gráinne de Burca, Director

Jean Monnet Working Paper 4/19

Yiannos Tolias

Explainable AI in Medicine, Confidence Intervals and Warnings

NYU School of Law • New York, NY 10011
The Jean Monnet Working Paper Series can be found at
www.JeanMonnetProgram.org

ISSN 2161-0320 (online)
Copy Editor: Danielle Leeds Kim

New York University School of Law
New York, NY 10011
USA

Publications in the Series should be cited as:
AUTHOR, TITLE, JEAN MONNET WORKING PAPER NO./YEAR [URL]

Explainable AI in Medicine, Confidence Intervals and Warnings

Yiannos S. Tolias*

Abstract:

The advancements in artificial intelligence (AI) have created a promising potential to revolutionize a number of domains including healthcare. These rapid developments in AI led policy makers and legal scholars to also look at AI legal implications. In addressing any legal implications, it is first necessary to understand the technical character of AI and specifically the branch of AI that deals with machine learning (ML). This understanding enables us to identify any new risks and/or safety concerns that differentiate ML systems from other conventional products and services. If such new risks and concerns exist then a differentiated legal treatment for ML systems could be justified. This paper focuses on ML liability in medicine and particularly on warnings. In this regard, the crux of the matter concerning ML systems performing medical tasks, such as diagnosis, is the type of information (warning) that a manufacturer/ML medical system should be providing to the physician. The physician would be acting as a learned intermediary and the manufacturer would be held to the standard of an expert in the field. The manufacturer/ML medical system should be providing information concerning the ML medical prediction to the physician in a manner suitable for her expertise that would enable the physician to provide appropriate explanations to the patient in order to obtain the patient's informed consent. This paper sets the foundations for a correlation between explainable ML, ML confidence intervals and warnings.

* Senior Emile Noël Global Fellow - The Jean Monnet Center, NYU School of Law and lawyer at the European Commission. I am grateful to Mark Geistfeld for the long and inspiring discussions we had on these issues, for all his comments on my project and support as well as for his brilliant classes on Products Liability and Tort. I am also grateful to Cem M. Deniz, Rajesh Ranganath and Narges Razavian, for the exciting discussions we had on machine learning and for their stimulating classes on Deep Learning in Medicine, Machine Learning for Healthcare and Deep Learning in Medicine respectively. I would also like to thank Gráinne De Búrca and Joseph Weiler for sharing their great ideas with me on how to develop this project and for the amazing research environment at The Jean Monnet Center. Additionally, many thanks to Catherine Sharkey for her excellent comments on legal issues related to this project. Moreover, I would like to warmly thank Claudia Golden for her superb support at The Jean Monnet Center. The views expressed are personal, do not necessarily represent the official position of the European Commission and any errors are attributable to me alone. All comments are welcomed at Yiannos.Tolias@ec.europa.eu.

Tables of Contents

1. Introduction

The advancements in artificial intelligence (AI) and particularly in machine learning (ML) have created a promising potential to revolutionize a number of domains including healthcare. There is no generally accepted definition of AI.[1] Scholars in legal and social sciences literature often criticize algorithms but few have considered in much depth their mathematical design.[2] Additionally, we see that AI poses specificities in some domains. Within this labyrinthic environment, many legislatures around the world have undertaken the challenging task to examine the legal, ethical, social and economic implications of AI.[3] In the course of these examinations, there were a number of novel questions posed, for example, whether there is a need for creating a "specific legal status for robots."[4]

This paper focuses on ML liability in medicine and particularly on *warnings*.[5] One of the underlying legal issues in this respect is whether existing legal frameworks on liability are fit for the development and deployment of AI. In addressing this issue, it is first necessary to understand the technical character of AI, specifically the branch of AI that deals with machine learning (ML). It is important to understand how a ML system carries out intelligent tasks, such as medical diagnosis, that have traditionally required human

---

[1] House of Lords Select Committee on Artificial Intelligence, *AI in the UK: ready, willing and able?*(2018), *available at* https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.; JANET FINLAY & ALAN DIX, AN INTRODUCTION TO ARTIFICIAL INTELLIGENCE 1 (UCL Press. 1996).

[2] Jenna Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, 3 BIG DATA & SOCIETY, 2 (2016).

[3] For example, see European Parliament, *European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics*(2017), *available at* http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//EN#BKMD-12.; House of Lords Select Committee on Artificial Intelligence. 2018.; Congressional Artificial Intelligence Caucus, *available at* https://artificialintelligencecaucus-olson.house.gov.

[4] European Parliament, *European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics*(2017).

[5] In a public consultation carried out by the European Parliament on artificial intelligence, 74% of the respondents considered "liability rules" as the second most important concern regarding regulation purposes (see European Parliament, *European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics*(2017)). As Geistfeld notes in the autonomous vehicles context, the liability question is the most important source of legal uncertainty manufacturers currently face and that eliminating this uncertainty and costs it would facilitate the safe deployment of autonomous vehicles (Mark A Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CALIF. L. REV., 1684 (2017).).

intelligence. This understanding enables us to identify whether there are any new risks and/or safety concerns that differentiate ML systems from other conventional products and services. If such differences exist, it could then justify a differentiated legal treatment for ML systems. This paper explores specificities of ML in the medical domain. Considering these specificities, the crux of matter for ML systems performing medical tasks, such as diagnosis, concerns the type of information (warning) that the ML manufacturer should be providing to the physician.

In carrying out this analysis, first, the paper identifies what distinguishes a conventional medical device from a ML system in medicine. It concludes that a fundamental distinction is that a conventional medical device, in effect, carries out mechanically based tasks whereas a ML medical system is carrying out a task that has been traditionally perceived as requiring human intelligence and decision making. But what is intelligence and how do machines learn to become intelligent? This paper looks deeper into these issues in order to identify any distinct risks in AI decision making in medicine. It reaches the conclusion that ML medical systems have some characteristics similar to human learning and intelligence and this encompasses novel legal challenges. It is this intelligent behavior of ML systems that renders the quest for designing adequate legal frameworks and warnings more challenging.

As explained in this paper there are technical and legal specificities in developing and deploying ML in medicine compared to other domains. Considering the specificities of the medical domain, the crux of matter for ML systems performing medical tasks, such as diagnosis, concerns the type of information (warning) that a manufacturer should be providing to the physician. In this regard, the physician would be acting as a learned intermediary and that the manufacturer would be held to the standard of an expert in the field.[6] The manufacturer should be providing information (warning) concerning the ML medical prediction to the physician in a manner appropriate for her expertise that would allow the physician to obtain the patient's informed consent.[7] After the physician is given

[6] Point raised by Mark Geistfeld (NYU Law School) in a discussion we had on this subject.
[7] Id.

an adequate warning appropriate for her expertise, she would then be able to combine it with her expert knowledge and then "translate" this information for the patient in her verbal explanation in order to obtain the patient's informed consent.[8]

In the quest of identifying the type of information that a manufacturer of a ML system should be providing to the physician, the paper takes step backwards and examines the doctrinal reasons behind the existence of warnings for conventional products. This examination aims to find out why do we need warnings accompanying products? What are the basic elements that need to be considered and balances that need to be struck in formulating adequate warnings? What should constitute a waring defect that would ensure that both innovation and consumer protection are safeguarded? What type of information should a warning encompass that would allow the consumer to make an informed choice on whether or not to use a particular product? How do warnings expose and explain risks associated with the functioning of conventional (not based on AI) products to consumers? Understanding the doctrinal mechanism behind warnings for conventional products allows us to understand the type of warning mechanism that would be more suitable for ML in medicine. Warning defects play a fundamental role in products liability litigation concerning medical devices and pharmaceuticals and this provides inspiration to identify the type of warning that could be fit for ML in medicine.

Finally, in the course of this assessment, this paper raises a widely discussed subject within the ML community, especially in relation to AI in healthcare, namely explainable AI. Explainable AI refers to AI that is able to provide some explanations for its predictions. This is an emerging topic that is currently the subject of substantial research by the ML community. The paper introduces the concept of explainable AI and indicates how explainable AI systems could be providing necessary and useful information to physicians and patients. It sets the foundations for a correlation between explainable ML, ML confidence intervals and warnings. This correlation would be further developed in another paper. This study does not seek to put forward a specific proposal on how to

[8] Id.

resolve the challenge concerning ML medical warnings it rather aims to discuss trade-offs between various feasible solutions.

2. What is intelligence and how do machines perform medical tasks?

It may be wondered at the outset what fundamentally distinguishes a conventional medical device, such as a magnetic resonance imaging machine (MRI), from a ML[9] medical system.[10] A fundamental difference is that a magnetic resonance imaging machine provides a physical measurement, such as images of an organ, by measuring signals generated from protons in tissues containing water molecules. In contrast, a ML medical device tries to interpret this physical measurement and draw inferences and predictions about the latent cause of the measurement such as the presence of a life-threatening tumor. In other words, the ML system is carrying out a task that has been traditionally perceived as requiring human intelligence and decision making.[11] Consequently, as explained in this paper, a new relationship appears to be formed between the AI system, the physician and the patient. But what is intelligence and how do machines become intelligent?

As early as 1955, a proposal was made for the Dartmouth summer research project on artificial intelligence to carry out a 2-month study on artificial intelligence based on the assumption that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[12] More than sixty years later, artificial intelligence is a thriving field with lots of practical applications.[13] However, it still remains a challenge for AI to solve tasks that people easily

[9] As explained below ML is one of the branches of AI.
[10] Such as, the first one approved by the U.S. Food and Drug Administration (FDA) to detect the eye disease diabetic retinopathy; see FDA press release - FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems (April 11, 2018).
[11] In this context, as Friedler et al also note machine learning systems carry out many of the decision-making activities that used to be done by humans, such as, credit ratings for loans, find patterns, assist in decisions that have significant impact on our lives and take decisions in the course of criminal proceedings; see Sorelle A. Friedler, et al., *On the (im)possibility of Fairness*, ARXIV.ORG/PDF/1609.07236.PDF, 1 (2016).
[12] John McCarthy, et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955*, 27 AI MAGAZINE (2006).
[13] IAN GOODFELLOW, et al., DEEP LEARNING 1 (MIT Press. 2016).

can perform but hard for them to precisely describe – problems that people solve intuitively such as recognizing faces in images and spoken words.[14]

Sejnowski argues that intelligence, like consciousness, are hard to define.[15] As he points out, scholars have written more on intelligence than any other subject in psychology after consciousness.[16] Psychologists since the 1930s, draw a distinction between two types of intelligence namely fluid intelligence and crystallized intelligence.[17] The former, uses reasoning and pattern recognition when faced with new situations in order to solve new problems without depending on previous knowledge, whereas the later, depends on previous knowledge.[18] Fluid intelligence relates to a developmental trajectory, reaching a peak in young adulthood and decreases with age, whereas crystalized intelligence, increases with age.[19] Fluid intelligence encompasses inductive reasoning and deductive reasoning.[20] There are different categories of reasoning including induction, abduction and deduction.[21] Reasoning is the ability to rely on existing knowledge to draw conclusions or infer something new about a particular field.[22] Without the ability to reason one simply recalls accumulated information.[23] Reasoning is particularly useful when knowledge is unreliable or incomplete.[24]

Inductive reasoning is particularly useful in machine learning.[25] Finlay and Dix summarize inductive reasoning as "generalization from cases seen to infer information about new cases unseen".[26] Inductive reasoning or inductive inference encompasses the

[14] Id.; In this regard, see also Sejnowski who argues that computers can now recognize objects in images almost as well as most adults can and there are computerized vehicles that drive themselves more safely than an average sixteen-year-old could (TERRENCE J. SEJNOWSKI, THE DEEP LEARNING REVOLUTION 3 (The MIT Press. 2018).).
[15] SEJNOWSKI, 20. 2018.
[16] Id.
[17] Id.
[18] Id.
[19] Id. Sejnowski points out that for example AlphaGO demonstrates both crystalized and fluid intelligence in a limited domain.
[20] https://en.wikipedia.org/wiki/Fluid_and_crystallized_intelligence
[21] FINLAY & DIX, 31. 1996.
[22] Id.
[23] Id.
[24] Id.
[25] Id. at, 33.
[26] Id. at, 32.

ability to use specific examples or observations to draw broader generalizations that allow one to successfully predict the label (category) of unseen information.[27] For example, we infer that all crows are black because every crow we see is black and thus inductive reasoning could be unreliable.[28] However, inductive reasoning is very useful and is the foundation of much of our learning.[29] A successful learner should be able to develop specific examples to broader generalization and hence why inductive reasoning or inductive inference is crucial to our learning.[30] In machine learning, the term generalization denotes how well a machine learning model can apply what it learned during its learning period to new cases that has never seen before. It is a core challenge in machine learning that algorithms must perform well on new previously unseen inputs and not just on those that the model was trained.[31] This type of learning should be contrasted to "learning by memorization" that lacks a fundamental element of learning systems namely the ability to label (categorize) unseen information.[32] The terms overfitting and underfitting that are explained below are used in machine learning to indicate how well the model learns and generalizes to new data.[33]

Shalev-Shwartz and Ben-David also explain how inductive reasoning could sometimes lead to false conclusions. They use the so called 'superstition' in the pigeon experiment carried out by psychologist B.F.Skinner and published in 1948 to demonstrate the problem with inductive reasoning.[34] In this experiment, food was delivered to a group of hungry pigeons at regular intervals. When the food was first delivered the pigeons were engaged in some random activity (pecking, turning head etc.)[35] The pigeons developed a superstitious behavior believing that the random activity that they were engaged in when food was first delivered was associated to the delivery of food. Consequently, they

[27] Shai Shalev-Shwartz & Shai Ben-David, *Understanding Machine Learning : From Theory to Algorithms*, 2 (2014).
[28] FINLAY & DIX, 32. 1996.
[29] Id.
[30] Shalev-Shwartz & Ben-David, 2 (2014).
[31] GOODFELLOW, et al., 108. 2016.
[32] Shalev-Shwartz & Ben-David, 2 (2014).
[33] Jason Brownlee, *Overfitting and Underfitting With Machine Learning Algorithms*(2016), *available at* https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/. These concepts are further explained below.
[34] B. F. Skinner, *'Superstition' in the pigeon*, 38 JOURNAL OF EXPERIMENTAL PSYCHOLOGY (1948).
[35] Shalev-Shwartz & Ben-David, 2 (2014).

continued to perform this useless activity believing that food would arrive. They ask what distinguishes the learning mechanisms that result in superstition from useful learning. The answer to this question is also fundamental to the development of automated learners. Shalev-Shwartz and Ben-David explain that human learners can rely on *common sense* to filter out random meaningless learning conclusions. Consequently, when the task of leaning is carried out by a machine one should provide well defined principles that will shield the program from reaching useless or senseless conclusions.[36]

They contrast the useful learning mechanism of rats that helps them avoid poisonous baits with the meaningless reasoning of pigeons. When rats come across food with novel smell or look they will initially eat a very small amount. If the food produces an ill effect namely nausea, the rat negatively labels it, and if it encounters this food in the future it will predict that this food will also have an ill effect and hence not eat it.[37] They refer to experiments carried out by Garcia and Koelling to identify why the rats' learning mechanism is more effective than that of the pigeons. In this experiment electric shocks were artificially inflicted on the rats following the consumption of particular food. Interestingly, the rats did not draw a correlation between electric shocks and that particular food. Thus, the rats were not avoiding that food when later encountering it. Similarly, the rats did not draw any correlations between sounds leading to nausea. As they highlight the rats appear to have some "built in" prior knowledge instructing them that, while temporal correlation between food and nausea can be causal, it is improbable that there would be a causal relationship between food consumption and electric shocks or between sounds and nausea.[38] They argue that the distinguishing feature between the rats' learning mechanism and the pigeons' learning mechanism is the incorporation of prior knowledge that biases the learning mechanism. This is what is referred to as inductive bias. Similarly, the incorporation of prior knowledge, biasing the learning process is inevitable to successfully design a learning algorithm and this is what is known as the "No-Free-Lunch theorem" that is examined below. [39]

[36] Id.
[37] Id. at, 1.
[38] Id. at, 3.
[39] Id.

What the above analysis shows, is that some form of bias, which could take the form of common sense for humans, is necessary for a successful learning process. It is also evident that the human learning mechanism as well as certain animal's successful learning mechanisms have inherent biases. These human and animal biases could be even possibly genetic ones that developed through evolution. In this respect, Shalev-Shwartz and Ben-David explain that the development of tools that express domain expertise, translating it into learning bias and quantifying the impact of this bias on the success of learning is a core aspect of the theory of machine learning.[40] They underline that the stronger the prior knowledge (or prior assumptions) that one starts the learning process with, the more successful it is to learn from further examples. However, they emphasize that the stronger these prior assumptions are the more rigid the learning is. Thus, there appears to be a fine line between strong prior assumptions that lead to more successful learning and rigidity in learning. Once this fine line is set, the initial question that appears to emerge is whether the machine learning algorithm's inductive bias has any distinct characteristics compared to human bias such as human common sense. The answer to this question would also determine whether there are any practical implications in the algorithm's medical prediction, for example, how it prioritizes patients in an intensive care unit. If there are distinct implications, then the next question is to identify a type of legal framework that is more appropriate for these algorithms. These issues will be further examined below.

As we have seen above, reasoning is a core element of learning and intelligence.[41] In addition to the types of reasoning noted above, reasoning can also progress in two directions – forward to the goal or backwards from the goal.[42] Both are used in AI depending on the circumstances.[43] In cases, like medicine, where the goal or hypothesis could sometimes be easily generated and where the physician wishes to have information (problem data) in order to proof or disproof a hypothesis, backward reasoning would be most likely applicable.[44] On the other hand, in cases where the problem data is given but

[40] Id.
[41] FINLAY & DIX, 31. 1996.
[42] Id. at, 33.
[43] Id.
[44] Id. at, 34.

the goal is unknown or there are many different possible goals, forward reasoning would be useful.[45]

Sometimes, knowledge is incomplete or changing and this requires reasoning methods that can deal with uncertainty.[46] Almost all activities necessitate some ability to reason in the presence of uncertainty.[47] In fact, save mathematical statements that are true by definition, it is hard to contemplate of any proposition that is absolutely true or any event that is absolutely guarantee to occur.[48] Goodfellow et al point out, that while it should be clear that we need a way of representing and reasoning about uncertainty, it is not evident that probability theory can deliver all the requisite tools for machine learning applications.[49] As they indicate, probability theory was developed to analyze the frequencies of events. For example, probability theory can be applied to study events like drawing a certain hand of cards in a poker game. These types of events are usually repeatable. They explain that, when we say that an outcome has a probability *p* of occurring, it means that if we were, for example, to be infinitely drawing a hand of cards, then a proportion *p* of the repetitions would result in that outcome. However, as they highlight, this type of reasoning does not appear immediately applicable to propositions that are not repeatable. Usefully for the purposes of this paper, they use the example of a physician who tells a patient that she has a 40 percent chance of having a flu. This, they point out, means something very different as we cannot make infinite replicas of the patient, nor there is a reason to assume that different replicas of the patient would present with the same symptoms yet have varying underlying conditions. In this respect, they distinguish between two types of probability namely frequentist probability and Bayesian probability. Frequentist probability relates to the rates at which events occur such as drawing a certain hand of cards in a poker game. Bayesian probability relates to the qualitative levels of certainty such as in the case where the physician diagnoses a patient. In this case, probability represents a degree of belief with 1 denoting absolute certainty

[45] Id.
[46] Such as non-monotonic reasoning, probabilistic reasoning, reasoning with certainty factors and fuzzy reasoning; there two more methods of reasoning namely reasoning by analogy and case-based reasoning; See further explanations, id. at, 34-43.
[47] GOODFELLOW, et al., 52. 2016.
[48] Id.
[49] Id. at, 53.

that the patient has flu and 0 denoting absolute certainty that patient does not have flu. They explain that if one lists several properties that we expect common sense reasoning to have about uncertainty then the only way to satisfy those properties is to treat Bayesian probabilities as behaving in the exact same way as frequentists probability.[50] Concluding, they state that probability can be perceived as the extension of logic to deal with uncertainty. As they note, probability theory provides formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions.[51]

The ability to successfully deal with uncertainty, is sometimes also useful in the medical domain, such as in intensive care units (ICU) as discussed below. Machine learning is useful to the kind of problems for which encoding an explicit logic of decision-making performs very poorly.[52] Machine learning is dealing with uncertain quantities and at times stochastic (nondeterministic) quantities.[53] In contrast to a number of branches of computer science that mainly deal with entities that are completely deterministic and certain.[54]

3. Introducing machine learning

As noted at the outset, there is no generally accepted definition of AI.[55] Be that as it may, machine learning is one of the branches of AI and in fact the most widely used technique in developing AI.

[50] See for more explanations on this conclusion id. at, 53-54.
[51] Id. at, 54.
[52] Burrell, BIG DATA & SOCIETY, 6 (2016).
[53] GOODFELLOW, et al., 52. 2016.
[54] Id.
[55] There are a number of definitions provided. For example, the House of Lords (House of Lords Select Committee on Artificial Intelligence, 14. 2018.) adopted the one used by the UK Government in its Industrial Strategy White Paper that defines AI as "Technologies with the ability to perform tasks that would otherwise require human intelligence, such as visual perception, speech recognition, and language translation" (Energy and Industrial Strategy, Industrial Strategy: Building a Britain fit for the future 37 (Department for Business ed., 2017). The European Commission in its Communication on Artificial Intelligence for Europe adopted the following definition "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analyzing their environment and talking actions – with some degree of autonomy – to achieve specific goals" (Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe. (2018).).

Goodfellow et al define machine learning as "essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions."[56] They describe a machine learning algorithm as an algorithm that is able to learn from data.[57] As to the meaning of leaning in the context of machine learning, they reiterate the neat definition provided by Tom Mitchell in 1997:

"A Computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

They provide helpful explanations to the above definition that are useful to summarize in order to set the foundations in machine learning before embarking into analyzing the limitations and legal implications of machine learning systems in medicine.[58]

Regarding, the task T, they explain that machine learning helps us to deal with tasks that are too difficult to solve with fixed programs written and designed by humans. They point out that the process of learning itself is not the task. Learning is the means of acquiring ability to perform the task. Machine learning tasks are often described in terms of how a machine learning system should process a collection of features that have been quantitatively measured from some event or object that we want the machine learning system to process.[59]

---

[56] GOODFELLOW, et al., 96. 2016.
[57] Id. at, 97. See further as to what constitutes machine learning: Sejnowski, who notes that learning algorithms instead of being told how to see or drive they learn from experience. As he points outs, data are the new oil, and learning algorithms are refineries that extract information from raw data (SEJNOWSKI, 3. 2018.). Shalev-Shwartz and Ben-David explain that machine learning refers to the automated detection of meaningful patterns in data. They also note that machine learning tools are related to programs with the ability to learn and adapt. They note that machine learning is extensively used in scientific applications such as bioinformatics, medicine and astronomy (Shalev-Shwartz & Ben-David, xv (2014).)
[58] GOODFELLOW, et al., 97-105. 2016.
[59] Id. at, 97. They explain that a number of tasks can be solved with machine learning including classification, classification with missing inputs (e.g. in medical diagnosis where many medical tests are expensive or invasive), annotation of the locations of roads in aerial photos, and machine translation.

As to the performance measure *P*, in order evaluate the abilities of a machine learning algorithm one should design a quantitative measure of the performance of this algorithm.[60] As they point out, normally this performance measure *P* specifically relates to the task *T* that is carried out by the system. They add that measuring the accuracy of the model refers to the proportion of examples for which the model produces the correct output. Alternatively, they note, that one can obtain equivalent information by measuring the error rate. They also explain that what normally interests one is how well the machine learning algorithm performs on data that it has not seen before as this will indicate how well it will perform when deployed in the real world. This evaluation takes place on the basis of a test set of data that is different from the data used for training the machine learning system.[61] They make an interesting observation that it might also have legal implications. The choice of performance measure might seem easy and objective but is frequently difficult to choose a performance measure that corresponds well to the desired behavior of the system.[62] The reason for this challenge is sometimes the difficulty of deciding what should be measured.[63] It thus appears that a machine learning algorithm might be presented as highly accurate on the basis of what was chosen to measure. Especially, in the medical domain where sometimes the aimed targets might not be clear, physicians might have different targets and/or patients' different expectations, these accuracy rates might be to a certain extent misleading.

Regarding the experience *E*, machine learning algorithms can generally fall under the categories of supervised or unsupervised depending on the type of experience that are allowed to have during the learning process.[64] Similarly, many of the human skills are acquired or refined though learning from our experience rather than following precise instructions given to us.[65]

[60] Id. at, 101-102.
[61] Id. at, 102.
[62] Id.
[63] Id.
[64] Id.
[65] Shalev-Shwartz & Ben-David, xv (2014).

Thus, machine learning at its current stage of evolution has at least some similarities to human learning and some dissimilarities to conventional products which rely on mechanical processes to perform tasks and/or fixed programs written by humans. Machine learning algorithms can be used to perform tasks that are traditionally performed by humans such as driving, speech recognition and image understanding.[66] Even more interesting, machine learning can perform tasks that exceed human capabilities. For example, the ability to learn to draw meaningful patterns from big and complex data sets seems to be very promising especially in the medical domain. [67] As Price argues, in the context of genetic testing and other "omics" technologies,[68] beyond the relatively simple links that can be explicitly labelled and comprehended, the observance or use of a number of complex relationships necessitate different type of algorithmic tools, what he calls "black-box algorithms".[69] Another distinct characteristic of a machine learning algorithm is that it can adapt to changes in the environment that they interact with.[70]

A machine learning algorithm generally includes two parallel operations or two distinct algorithms: a "classifier" and a "learner."[71] Classifiers receive the input (known as a set of "features") and produce an output (a "category").[72] For example, a machine learning algorithm that performs diagnosis may receive the input such as, symptoms or clinical

[66] Id. at, 3.
[67] See id. at, 4.
[68] Such as, testing of large panels of metabolites, gene expression levels, and protein levels.
[69] In the medical domain, he explains that in order "to discover new complex relationships black-box medicine relies on computer systems that improve their performance over time by trying a certain solution, evaluating the outcome, and then modifying that solution accordingly to improve future outcomes"; as to what constitutes "black-box medicine," he notes that first, the information used to develop the relationships and predictions is based on a large set of information; secondly, the predictions are based on complex connections between patient characteristics and anticipated treatments without understanding or identifying the underlying connections; finally, the relationships generally cannot be confirmed through clinical trials. See further, W. Nicholson II Price, *Black-Box Medicine* 28 HARV. J. L. & TECH. 419, 429-432 (2015).
[70] Shalev-Shwartz & Ben-David, 4 (2014).
[71] Burrell, BIG DATA & SOCIETY, 5 (2016).
[72] Id.

presentations, and produce a disease diagnosis as output.[73] However, the machine learning algorithms known as the learners must first train on test data.[74]

As we have seen, humans learn from experience rather than following specific instructions given to them. This type of learning also very much characterizes machine learning. As we have also seen, machine learning methods are largely divided into supervised and unsupervised learning algorithms.[75] Unsupervised learning algorithms process data without labels (i.e. without human categorization of the particular data) and are trained to discover patterns.[76]

4. How is machine learning used in medicine?

Healthcare is a domain where machine learning techniques could have a great impact hence the great momentum for research in deploying machine learning algorithms in this domain. Machine learning in medicine could be enhancing personalized medicine to make predictions and treatment recommendations by identifying complex, implicit biological relationships from large heath datasets.[77] Ghassemi et al. identify three main categories where machine learning could provide great healthcare opportunities: automating clinical tasks,[78] providing clinical support[79] and expanding clinical capacities.[80] In pursuing such objectives, machine learning could for example have a

[73] Id. at, 5.
[74] It is important to point out that this refers to the machine learning approach called "supervised" learning; id.
[75] G. Litjens, et al., *A survey on deep learning in medical image analysis*, 42 MED IMAGE ANAL, 62 (2017).
[76] Id.
[77] W. Nicholson II Price, *Describing Black-Box Medicine* 21 B.U. J. SCI. & TECH L. 347, 349 (2015).
[78] Automating clinical tasks during diagnosis and treatment (automating medical image evaluation and automating routine processes).
[79] Optimizing clinical decision and practice support (standardizing clinical processes and integrating fragmented records).
[80] New potentials in screening, diagnosis and treatment (expanding the coverage of evidence, moving towards continuous behavioral monitoring and precision medicine for early individualized treatment). See for further details on these categories, Marzyeh Ghassemi, et al., *Opportunities in Machine Learning for Healthcare*, ARXIV:1806.00388V2, 4-6 (2018). See also work by N. Coudray, et al., *Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning*, 24 NAT MED (2018). This work shows how deep-learning convolutional neural networks could aid in assisting pathologists in their classification of lung tissue images; this information can be very important in applying appropriate and tailored targeted therapies to patients with lung cancer. See also work on the application of deep convolutional neural networks to breast cancer screening by Krzysztof J. Geras, et al., *High-*

promising potential in bioinformatics.[81] The application of machine learning in bioinformatics could be useful in the prediction of biological processes, prevention of diseases and personalized treatment.[82] Machine learning technology can help medical researchers to discover hidden patterns from the huge amount of electronic health records (EHR) data and develop predictive models that could help the physicians with clinical decision making.[83] Concretely, for example, machine learning algorithm could be used to predict diagnoses given features from the EHR.[84]

Similarly, in recent years, deep neural networks and particularly convolutional neural networks (CNNs) have been adapted rapidly by the medical imaging research community.[85] Deep learning applied to medical imaging delivers automatic discovery of object features and automatic exploration of feature hierarchy and interaction.[86] These deep learning advancements in pattern recognition could have concrete successful uses in for example detection of metastatic breast cancer in images of lymph nodes biopsies.[87] In this respect, it was found that the accuracy rates of detection of pattern recognitions was reaching an almost 0.995 when physicians worked together with deep learning systems; machine learning algorithms and physicians did better together than either alone.[88] However, deep learning requires the availability of large labelled datasets in

---

Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks, ARXIV:1703.07047V3 (2018).

[81] Bioinformatics aim to examine and understand biological processes at a molecular level (Daniele Ravì, et al., *Deep Learning for Health Informatics*, 21 IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 10 (2017).)

[82] Id.

[83] Xinyuan Zhang, et al., *Multi-Label Learning from Medical Plain Text with Convolutional Residual Models*, ARXIV:1801.05062V2, 1 (2018).). See also Price, B.U. J. SCI. & TECH L. , 348 (2015). Price argues that this approach promises a faster, less costly path to empower many novel biological relationships, enhancing possibilities for treatment decisions and promoting new therapeutics.

[84] Zhang, et al., ARXIV:1801.05062V2, 1 (2018). Mukherjee investigated how doctors learn to diagnose in contrast to AI. It was indicated in this investigation that a successful reading of a radiological image was partly based on the subconsciousness of the radiologist (see Siddhartha Mukherjee, *A.I. VERSUS M.D. What happens when diagnosis is automated?* , THE NEW YORKER 2017.); the extent to which such observations could differentiate an AI radiological tool from a human radiologist would become clearer in this paper.

[85] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 11 (2017).

[86] Id. at, 13. See also deep learning uses in medical imaging and specific uses in anatomical application areas (brain, eye, chest, digital pathology and microscopy, breast, cardiac, abdomen, musculoskeletal and other) by G. Litjens, et al., *A Survey on Deep Learning in Medical Image Analysis*, ARXIV:1702.05747V2, 7-23 (2017).

[87] SEJNOWSKI. 2018.

[88] Id.

order to be successful in disease detection and classification.[89] Furthermore, as further explained below, beyond the challenge concerning the availability of a large labelled dataset, it is sometimes also a challenge to have the "correctly" labelled data in medicine as sometimes domain experts disagree on the labeling themselves (e.g. whether or not there is a nodule on a computed tomography (CT) image).

Other uses of machine learning in medicine include helping the physicians to choose between a selection of known interventions, recommend an off-label use of an approved intervention, suggest what drug is most likely to treat the specific tumor of a particular patient more effectively based on the genetic sequence of the patient's tumor, continuously evaluate a patient's vital signs and sound an alarm at an earlier stage, allocate scarce resources by suggesting which patient might benefit most from, for example, a transplant and generate hypotheses for traditional biomedical research that might lead to the uncovering of the underlying mechanisms.[90]

A different type of use of machine learning in healthcare concerns pervasive sensors, such as wearable, implantable and ambient sensors that are used to continuously monitor certain features related to health and wellbeing.[91] Deep learning has substantially contributed to the utility of such pervasive sensing in a wide variety of health applications by improving the accuracy of sensors that measure food calorie intake, energy expenditure, activity recognition, sign language interpretation and detection of irregular events in vital signs.[92] However, a major challenge still remains namely the selection of features that can generalize across the wide variety of food and daily activities.[93]

It has been also suggested that predictive systems analyzing large quantities of data could be used in six concrete cases: high-cost patients, readmissions, triage, decompensation when a patient's condition worsens, adverse events and treatment optimization for

[89] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 13 (2017).
[90] W. Nicholson II Price, *Medical Malpractice and Black-Box Medicine in* BIG DATA, HEALTH, LAW AND BIOETHICS 297, (Glenn Cohen ed. 2018).
[91] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 13 (2017).
[92] Id. at, 14.
[93] Id. at, 13.

diseases having an impact on multiple organs.[94] For example, the ability to accurately predict the trajectory of a patient's disease could allow the clinician to target more accurately expensive and complicated therapies to patients who stand to benefit the most from them.[95] However, such predictions, as useful as they may be, at the same time raise a number of sensitives. Such predictions could be determining whether a patient should be administered certain life-saving treatments, and if so, which patient should be prioritized. Irrespective of the accuracy of machine learning algorithms in making such predictions it might still necessitate some form of reasoning as to how this prediction has been reached. Both patients and physicians in most cases would justifiably require some sort of explanation on these predictions that could have a paramount impact on the life of a person. Similarly, different research work in medicine aims to develop accurate predictions on future illnesses, readmissions and mortality of individuals in order to improve clinical decision making and optimize clinical approaches.[96] Such predictions could have valuable impacts on the care of a patient but at the same time could in the future have an impact on health and life insurances of individuals.[97] Consequently, once again, some form of explanation as to how the particular prediction has been reached might be necessary in this respect as well. Furthermore, machine learning could also aid in drug discovery and development.[98] Moreover, machine learning could be applicable in public health that aims to prevent diseases, prolong life and enhance healthcare by examining the spread of disease and social behaviors in connection to environmental factors.[99]

---

[94] D. W. Bates, et al., *Big data in health care: using analytics to identify and manage high-risk and high-cost patients*, 33 HEALTH AFF (MILLWOOD), 1124-1127 (2014).
[95] Id. at, 1127.
[96] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 15 (2017).
[97] In this respect it is interesting to point out that the cost of health care in the US is almost double than that in other developed countries; this unsustainable development has led to calls for improving the value of health care (see for more details Bates, et al., HEALTH AFF (MILLWOOD), 1123 (2014).). These cost related challenges might also have an impact on how machine learning algorithms are trained and how they are used the medical domain that could be raising concerns.
[98] Price, HARV. J. L. & TECH. , 435-437 (2015). For example, it could broaden the already-widespread notion of off-label use (i.e. use that has not been approved by the FDA) beyond the use founded on practitioner experience or limited clinical trials; it could also aid the discovery of new drugs and aid and reduce the costs of clinical trials.
[99] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 15 (2017).

Deep learning systems would not only be applicable in the clinic but could soon be directly used by patients outside the clinic as well without physician's presence or examination. For example, it could be possible for anyone with a smartphone to photograph a suspicious skin lesion and having it diagnosed immediately.[100] This direct connection between the manufacturer of machine learning medical systems and patients raises its own challenges and legal implications including warnings.

As it is apparent from the above, machine learning has already started changing the practice of medicine. These changes encompass new challenges and could have new legal implications as examined throughout this paper.[101] A fundamental question in this respect would be the extent to which doctors and patients would trust machine learning algorithms. In the case of conflicting opinions which opinion should prevail? As already mentioned above, in contrast to conventional medical devices, machine learning algorithms assess themselves the relevant images or features concerning a patient and reach their own predictions. Whereas, it would up to the doctor to make this assessment and reach a conclusion when relying on conventional medical devices. Who should be liable in case where the wrong opinion was followed? Should the physician follow the machine learning algorithm's "opinion" if the physician does not know the biological relationships underlying the algorithm's recommendation and cannot explain it to the patient?[102] Will the patient accept the advice of the algorithm if they cannot understand even the basic links made to reach that advice?[103] The question is how should law and policy choices facilitate the deployment of such algorithms in medicine and ensure the acceptance by patients and health providers?[104] As Price points out, the argument is sometimes made, admittedly somewhat cynical, that there are lots of medical treatments provided and medicines subscribed where the exact mechanism is unknown.[105] As we

[100] SEJNOWSKI. 2018.
[101] As to obstacles that there may be erected from changing the practice of medicine through statistical learning approaches and how these might be overcome see Rahul C. Deo, *Machine Learning in Medicine*, 132 CIRCULATION (2015).
[102] Price, HARV. J. L. & TECH. , 465 (2015).
[103] Id. at, 466. This challenge concerning the patient might also have an impact on informed consent.
[104] Id.
[105] Price points out that it has been argued, admittedly somewhat cynical, that doctors and patients already accept treatments even though they don't have much understanding for those treatments; that physicians could only 55% of the time correctly identify whether a drug was FDA approved for a particular indication;

noted above, machine learning algorithms in medicine introduce a distinct addition to medical diagnosis and treatment compared to conventional medical devices and medicines. Machine learning algorithms, fully autonomously assess images and/or features concerning a patient and reach their independent and precise recommendation on what should be done. Whereas with conventional medical devices or medicines, the physician still plays a major role in assessing the treatment.

Machine learning algorithms also play a very important role to the rising and promising personalized medicine.[106] Physicians are already using, for example, diagnostic genetic tests to tailor treatments to the individual patient.[107] However, there are a number of distinctions between such conventional approaches and machine learning approaches to personalized medicine. Some of these distinctions could be also encompassing new legal implications that would need to addressed. First, conventional approaches to personalized medicine use only a limited and simple set of relationships but machine learning algorithms would be able to make connections between massive amounts of data and identify more complex biological relationships.[108] Secondly, conventional approaches, what Price calls explicit personalized medicine, is based on scientifically identified and understood relationships.[109] This approach is based on scientific and clinical research aiming to identify and explain simple biological relationships between certain measurable features of a patient and possible medical outcomes for that particular patient.[110] This approach uses relationships between different biomarkers and medical responses to identify diagnoses and medical treatments.[111] On the other hand, the machine learning approach, what Price calls "black-box" medicine, is grounded on

finally, that the mechanism of actions is unknown for many drugs; he points out even given this possibility, it still appears that some knowledge is better than none; moreover, he puts forward some proposals on how this problem could be legally addressed (id.).

[106] Personalized medicine refers to the tailoring of treatment based on the different characteristics of individuals. The idea behind personalized medicine is that all patients are different and treatment should be tailored to the particular patient. In addition to medical treatment, personalized medicine could aid drug discovery (see id. at, 424-425.).

[107] Id. at, 424.

[108] See id. at, 424 and 429.

[109] Id. at, 425.

[110] Id. at, 427. He notes that "explicit" refers to the fact that these relationships are explicitly identified and validated; in other words, knowing why a treatment is tailored to a particular patient.

[111] Often these biomarkers are genomic variations and genetic diagnostic tests and it in these areas that most research on personalized medicine takes place (see id.).

relationships that are not understood and often, based instead on non-transparent computational algorithms. Black-box medicine makes predictions and improve treatments on the basis of large datasets and algorithms but it does not provide any explanations or identifying complex relationships.[112] Machine learning in medicine could prove extremely successful in identifying connections that cannot be easily labelled or understood.

Having referred to the opacity of machine learning algorithms, it should be also pointed out that radiologists or other physicians might, like ML algorithms, be unable to articulate the internal algorithm they use to take a decision.[113] They often use their experience to recognize connections between image and implicit biology but those relationships cannot be explicitly indicated.[114]An interesting question that arises in this respect is whether the type of opacity inherent in a human medical decision is similar or not to the one inherent in a machine learning algorithm. Similar or not, still a human physician is subjected to different legal rules than products. A physician might be sued for medical malpractice and would be required to provide explanations concerning her opaque at the time decision. Regarding products, as shown throughout this paper, there are unique distinctions between conventional products and machine learning technologies.

5. Supervised and unsupervised machine learning and challenges in medicine

As we have introduced above, machine learning algorithms could generally fall under the categories of supervised or unsupervised depending on the type of experience that are allowed to have during the learning process. [115]

[112] Id. at, 429.
[113] Id. at, 433.
[114] Id.
[115] GOODFELLOW, et al., 102. 2016.

Supervised learning, deep or not,[116] is the most common form of machine learning.[117] Supervised machine learning algorithms experience a dataset which is a collection of features that are associated with a label or target.[118] The idea behind supervised learning is that the target or label (y) is provided by the teacher who shows the machine learning system what to do.[119] The machine is basically taught during its learning process to predict (y) from different input values (x). It does that by trying to find a function that best connects input (x) and output (y).[120] In supervised machine learning we have both a target (or label) and a collection of features. The labeling of data (e.g. "tumor" or "no tumor") means that we know the output or target (y). Supervising learning techniques have been quite popular in medical image analysis.[121] In designing such systems it is required to select discriminant features from images that is done by human researchers hence why these systems are known as systems with handcrafted features.[122] However, as explained below, sometimes the selection of features is a challenging task in medicine and that could translate to detrimental diagnosis and/or treatment. Moreover, defining target (y) might also prove challenging in medicine especially in cases where targets are not binary or there is no agreement on the labeling of the data.[123] For instance, in designing a machine learning system to determine the action to be taken in an intensive care unit, there could be different targets that need to be simultaneously balanced. For example, a balance would need to be struck between prolonging the life of a patient and providing a good quality of life to a patient. On what basis should a doctor disregard a machine learning prediction? What information is needed to enable the doctor to reach such a decision?

[116] Deep learning is explained below.
[117] Yann LeCun, et al., *Deep Learning*, 521 NATURE, 436 (2015).
[118] GOODFELLOW, et al., 103. 2016.
[119] Id.
[120] The way to achieve good predictions during training is to compute an objective function that measures the error between the machine's prediction and the desired prediction. The machine then modifies its internal adjustable parameters in order to reduce this error. These adjustable parameters are referred to as weights. In machine learning systems there could be millions of these adjustable weights and millions of labelled examples used to train these machines (LeCun, et al., NATURE, 436 (2015).).
[121] See how supervised learning is used in medicine at Deo, CIRCULATION, 1920 (2015).
[122] Litjens, et al., MED IMAGE ANAL, 60 (2017).
[123] The challenges concerning the labeling of data are further examined below.

In unsupervised learning there is no teacher so the algorithm should learn by itself to make sense of the data.[124] In other words, there is no need for human labor to annotate examples during learning as the algorithm extracts by itself information from the dataset.[125] Thus, in unsupervised learning there is no need to have labelled data.[126] However, the dividing line between the two types of learning is often blurred and many machine learning systems can be used to perform both tasks.[127] Moreover, the distinction between supervised and unsupervised is not clear as it is not clear whether a value is a feature (an input) or a label (a target).[128] Thus, in principle, features can be selected in a manner that could "bias" the unsupervised learning algorithm towards a certain target despite that no labels were provided during its training. This is another challenge with practical impacts in medicine and legal implications that would need to be addressed. If there is a way to reveal such limitations might provide better medical care and avoid any potential legal liability concerns.

Unsupervised machine learning might have a great potential in radiology because machines could learn to detect patterns like humans recognize objects and structures without any need for labels. Hence why, there are unsupervised machine learning techniques that are developed which are expected to have an impact in medical imaging.[129] Moreover, unsupervised learning might prove useful in the so called precision medicine.[130] Most common diseases are inherently heterogeneous hence the quest to redefine disease according to pathophysiologic mechanisms, which could in turn, indicate new paths to therapy.[131] The challenge is to identify such mechanisms for complex

[124] GOODFELLOW, et al., 103. 2016.
[125] Id. at, 142.
[126] Human and animal learning is mainly unsupervised learning (LeCun, et al., NATURE, 442 (2015).).
[127] GOODFELLOW, et al., 103. 2016. It is also interesting to point out that some machine learning algorithms do not solely experience a fixed dataset. Reinforcement learning algorithms interact with an environment consequently there is a feedback loop between the learning system and its experiences. However, most machine learning algorithms simply experience a dataset; a dataset is a collection of features (id. at, 104.). Systems that combine both reinforcement learning and deep learning are at their infancy but they have already outperformed other systems and have impressive results in video games (LeCun, et al., NATURE, 442 (2015).).
[128] GOODFELLOW, et al., 142. 2016.
[129] See further Litjens, et al., ARXIV:1702.05747V2, 27 (2017).
[130] See more details Deo, CIRCULATION, 1921 (2015).
[131] Id.

multifactorial diseases.[132] Unsupervised learning might identify patterns that could prove useful in this quest.[133] Unsupervised learning, in contrast to supervised learning, aims to identify patterns in the data and there is no predicted outcome. In fact, using supervised learning techniques in this regard it could miss different subgroups completely and consequently not allowing for the identification of novel disease mechanisms.[134]

6. Deep learning potentials and inherent challenges

Deep learning has in recent years set an exciting momentum in machine learning.[135] Deep learning is a particular type of machine learning.[136] Neural networks are a type of learning algorithm which constitutes the basis of most deep learning methods.[137] Deep learning technique that has its foundations in artificial neural networks, is emerging as a powerful tool for machine learning, promising to reshape the future of artificial intelligence.[138] Deep neural networks can be trained by both supervised and unsupervised learning techniques.[139] Regarding the relationship between deep learning and conventional machine learning, LeCun et al note that conventional machine-learning algorithms had limited ability to process natural data in their raw form.[140]As they point out, for many years constructing a conventional machine learning system necessitated careful engineering and substantial domain expertise to design a feature extractor that transformed the raw data into suitable internal representation or feature vector from which the learning subsystem (usually the classifier) could detect or classify patterns in the input. Whereas, representation learning is a group of methods that allows machines to be fed with raw data and to automatically discover the representations needed for

[132] Id.
[133] A similar approach albeit on genomics led to finding an eosinophilic subtype of asthma; (see further id.)
[134] Id.
[135] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 4(2017).
[136] GOODFELLOW, et al., 96. 2016.
[137] Litjens, et al., MED IMAGE ANAL, 62 (2017). In this respect, Finlay and Dix note that the development of artificial neural networks, modelled on the human brain, has been welcomed by some as the foundations for "genuine machine intelligence and learning" (see FINLAY & DIX, 6. 1996.).
[138] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 4(2017). In this context they also note that in contrast to more traditional use of neural networks deep learning accounts for use of many hidden neurons and layers, typically more than two, and this is an architectural advantage combined with new training paradigms (at 4).
[139] Id. at, 5.
[140] LeCun, et al., NATURE, 436 (2015).

detection or classification.[141] As they explain, "[d]eep-learning methods are representation-learning methods with multiple levels of representation obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned."[142] Adding more hidden layers to the network enables a deep architecture to be build that can express more complex hypotheses due to the ability of the hidden layers to capture nonlinear relationships.[143] The crucial aspect of "deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure."[144] This is the key advantage of deep learning that good features can be learned automatically using a general-purpose learning procedure rather than resorting to hand designing good feature extractors.[145] Deep learning was designed to overcome certain obstacles faced by conventional algorithms.[146] Conventional machine learning algorithms were failing to generalize well on certain tasks such as recognizing speech and objects.[147] This challenge becomes exponentially more difficult when working with high-dimensional data and the mechanism used to achieve generalization in conventional machine learning are insufficient to learn complex

[141] Id.
[142] Id. They use the example of an image to provide further clarifications on this process. As they state, an image comes in the form of a collection of pixels values; the learned features in the first layer of representation normally represent the presence or absence of edges at specific orientations and locations in the image; the second layer normally detects motifs by identifying specific arrangements of edges, irrespective of small variations in the edge positions; the third layer could assemble motifs into larger combinations that match parts of familiar objects and subsequently layers would identify objects as combinations of these parts.
[143] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 5(2017).
[144] LeCun, et al., NATURE, 436 (2015). They note that deep learning dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection as well as in other domains such as drug discovery and genomics; additionally, deep learning techniques are used to match news items, posts or products with user's interests, select relevant results of search. Google's automatic translator has also resulted from the deep learning technique (as noted by Larry Hardesty, *Explained: Neural networks* (2017), *available at* http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.).
[145] LeCun, et al., NATURE, 438 (2015).
[146] GOODFELLOW, et al., 152. 2016.
[147] Id.

functions in high-dimensional spaces.[148] Deep learning turned out to be successful in discovering intricate structures in high-dimensional data.[149]

It is clear that deep learning could prove particularly useful in medicine where multiple factors could be corelated in a complex network. However, we also see that deep learning architectures at their current stage of evolution could be highly non-interpretable (i.e. ML system not able to provide an explanation for its prediction) and this could be a particularly problematic in medicine. As we have seen above, the ability to generalize well is key element to successful learning. As we have also seen, to generalize well, machine learning algorithms should be guided by prior beliefs on the type of function they should learn.[150] In this respect it is interesting to point out that the so called no free lunch theorem for machine learning (Wolpert, 1996) provides that, "averaged over all possible data-generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points."[151] This means, as Goodfellow et al explain, that in some sense, no machine learning algorithm is universally any better than any other. However, they emphasize that if one makes assumptions about the types of probability distributions one encounters in real-world applications, then one can design learning algorithms that perform well on these distributions. Thus, the goal of machine learning research they explain is not trying to find the best learning algorithm. Instead the goal is to understand the types of distributions that are relevant to the "real world" that an AI

[148] Id. As they also state the majority of deep learning algorithms are based on an optimization algorithm called stochastic gradient descent (see at 96 and 149-150); they also describe how one can combine various algorithm components, such as an optimization algorithm, a cost function, a model, and a dataset to build a machine learning algorithm (see at 97 and 151-152)

[149] LeCun, et al., NATURE, 436 (2015). For example, predicting what would be the effects of mutations in non-coding DNA on gene expression and disease.

[150] GOODFELLOW, et al., 154. 2016. As they explain these prior beliefs can take different forms. For example, prior beliefs can be implicitly expressed by choosing algorithms that are biased toward choosing some class of function over another. They explain how deep learning introduces additional (explicit and implicit) priors in order to reduce the generalization error on complex tasks. They refer to K-nearest neighbors' algorithm and decisions trees (see at 154-157). Furthermore, they point out that other approaches to machine learning make stronger, task-specific assumptions. However, they note that normally one does not include strong, task specific assumptions in neural networks in order to generalize to a much wider variety of structures. They point out that the central idea in deep learning is that we assume that the data was generated by composition factors or features and that many other similarly generic (mild) assumptions can further improve deep learning algorithms (see at 157).

[151] Id. at, 115-116.

agent experiences, and the type of machine learning algorithms that perform well on data drawn from the kinds of data-generating distributions we care about.[152]

Deep learning is a promising machine learning technique with great potential. Machine learning will have many more successes in the near future as it necessitates very little engineering by hand, thus can easily take advantage of increases in the amount of available computational data.[153] Provided that a deep learning network is optimally weighted it leads to an effective high-level abstraction of the raw images or data.[154] This high level of abstraction result in an automatic feature set, which otherwise would have necessitated hand-crafted or bespoke features.[155] Consequently, deep learning, for example, in the domain of medical imaging can generate features that are more sophisticated and harder to elaborate in descriptive ways.[156]

Research work[157] in the fields of dermatology and ophthalmology has shown that it is even possible to outperform medical experts in certain tasks using deep learning for image classification.[158] Deep learning techniques are the main techniques now used in medicine but as explained throughout this paper there are numerous challenges and limitations in

[152] Id. at, 116. In this context, Deo notes that simple algorithms can perform as well as more complex ones when the number of training examples is low and hence more complex models are likely to overfit and generalize poorly (Deo, CIRCULATION, 1924 (2015).).

[153] LeCun, et al., NATURE, 436 (2015). As they point out the deep learning discovers complex structure in large data sets by using backpropagation algorithm to show how a machine should change its internal parameters which are used to compute the representation in each layer from the representation in the previous layer (they make references to deep convolutional nets and recurrent nets). In this context, Litjens et al note, that the concept that lies at the basis of many deep learning algorithms is that computers themselves learn the features that optimally represent the data for the problem in question: models (networks) composed of many layers that transform input data (e.g. images) to outputs (e.g. indicating whether disease is present or absent) while learning increasingly higher-level features (Litjens, et al., MED IMAGE ANAL, 60 (2017).).

[154] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 4(2017).

[155] Id.

[156] Id. As they note implicit features could determine fibroids and polyps and characterize anomalies in tissue morphology such as tumors (at 4).

[157] Andre Esteva, et al., *Dermatologist-level classification of skin cancer with deep neural networks*, 542 NATURE (2017);Varun Gulshan, et al., *evelopment and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus PhotographsAccuracy of a Deep Learning Algorithm for Detection of Diabetic RetinopathyAccuracy of a Deep Learning Algorithm for Detection of Diabetic Retinopathy*, 316 JAMA (2016).

[158] Litjens, et al., ARXIV:1702.05747V2, 26 (2017). However, as Litjens et al. explain such findings need to put into context relative to medical image analysis in general, as the majority of tasks can be no means be treated as "solved."

applying machine learning to different medical tasks.[159] As we will see below, there are tradeoffs with deep learning including the important issue of lack of interpretability despite that deep learning could prove more accurate for some cases. At the current stage of evolution, neither the engineer who developed the deep learning system nor the system itself would be able to provide an explanation why that prediction was reached. This is the point where deep learning algorithms especially when applied to medicine could pose certain challenges and consequently legal challenges as well. Additionally, it has been shown that the exact deep learning architecture to be used in medical image analysis is not the most important determinant in obtaining good solutions.[160] An important aspect is that expert knowledge about the task aiming to solve can provide advantages that go beyond the addition of more layers to a deep learning network.[161]

Although deep learning architectures are delivering substantial improvements compared to conventional machine learning algorithms, many scientists and researchers remain skeptical of their application to the medical domain.[162] As already noted, deep learning models are often not interpretable and this lack of interpretability has distinct impacts in healthcare where the patient and physician could justifiably require some explanation on the suggested predictions concerning the course of treatment. Additionally, deep learning models are used as a black-box without the researchers be able to explain why is in certain cases successful or without the ability to modify them in cases where there are misclassification problems.[163] Furthermore, in order to train effectively deep learning models, they require large sets of training data and thus rare diseases or events might not be well suited to deep learning.[164] Also, for many applications, raw data cannot be directly used as training input for deep neural networks.[165] Thus, in addition to the burden of transforming this raw data into appropriate training inputs, as explained below, labeling

[159] See id. at, 2. Out of 308 papers reviewed in the survey by Litjens et al. it was clear that deep learning has spread all over medical image analysis (see at 23).
[160] Id. at, 23.
[161] Id. Convolutional neural networks (CNNs) that are the top deep learning performers in medical image analysis.
[162] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 16 (2017).
[163] Id. at, 17.
[164] Id.
[165] Id.

of these data carries its own separate challenges. Moreover, many deep neural networks could easily reach wrong predictions when some noise (artifact) is applied to a medical image.[166] Finally, as we have seen, the layers of features in developing deep learning architectures are not designed by human engineers but they are learned from data using a general-purpose learning procedure.[167] Consequently, even the ML engineer who develops such architectures in many cases does not know precisely how the prediction is reached or how the deep learning devices might precisely perform in a particular occasion. As it appears, there are tradeoffs and a number of challenges in developing deep learning architectures. These challenges could have practical impacts especially in the medical domain and legal implications especially regarding warning obligations.

7. The challenge of learning from medical data

As we have seen, machine learning algorithms, especially deep learning ones, have the ability to learn from data and thus require less engineering by hand.[168] Despite the many advances and great potentials of the use of machine learning in medicine, the direct application of machine learning in medicine remains filled with many pitfalls.[169] Many of these challenges arise from the objective to make personalized predictions using large volumes of noisy and biased data.[170] Failing to address these challenges could hamper the validity and utility of machine learning methods.[171] Healthcare has become a natural domain for the application of machine learning particularly due to the increasingly large amounts of data resulting from electronic health records (EHRs).[172] Machine learning

[166] Id.
[167] LeCun, et al., NATURE, 436 (2015). They note that deep learning dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection as well as in other domains such as drug discovery and genomics; additionally, deep learning techniques are used to match news items, posts or products with user's interests, select relevant results of search. Google's automatic translator has also resulted from the deep learning technique (as noted by Hardesty. 2017.).
[168] See LeCun, et al., NATURE, 436 (2015).
[169] Ghassemi, et al., ARXIV:1806.00388V2, 1 (2018).
[170] Id.
[171] Id.
[172] Id. Sometimes, the term big data used. Big data refers to the analysis of large amounts of data and collecting new insights from that analysis (noted by Bates, et al., HEALTH AFF (MILLWOOD), 1123 (2014).). See also Price's explanations on big data; he also notes that with big data far more relationships could be used than the current version of personalized medicine; (Price, HARV. J. L. & TECH. , 424 and 430-432 (2015).). See also Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 4(2017).; Litjens, et al., MED IMAGE ANAL, 60 (2017).; W. Nicholson II Price, *Medical Malpractice and Black-Box Medicine in*

technology can aid in discovering hidden patterns from massive EHR data and develop predictive models that could be used for medical decision making.[173] However, clinical data and practice present distinct challenges that create complications to the use of common methodologies.[174] There are a number of challenges associated with the collection and use of data.[175]

When medical care is provided, care staff collect clinical data about a patient and consider knowledge from the general population to decide how to treat the patient.[176] As explained by Ghassemi et al, different data types come with different challenges.[177] For example, high-frequency monitors are used to record real time data at a patient's bedside such as oxygen saturation. These signals frequently have artifact corruption (e.g. sensors falling off), hence why such data must be aggregated, filtered or discarded to remove these artifacts before any learning or feature extraction.[178] Another challenge concerns vital signs, laboratory tests and other numerical measurements that are noted by medical staff.[179] Such tests and numerical measurements are often irregularly ordered; non-invasive values can be in conflict with high-frequency invasive data;[180] staff may have a feeling about the patient's state and preferentially record results that concur with that understanding;[181] clinical staff order laboratory tests related to the amount of variability

---

BIG DATA, HEALTH LAW, AND BIOETHICS 296, (I. Glenn Cohen ed. 2018).; Harini Suresh, et al., *Clinical Intervention Prediction and Understanding using Deep Networks*, ARXIV:1705.08498V1, 1 (2017). See other types of data that can be used in machine learning including billing code labels and patient physiological signals to predict mortality at id. at, 2-3.

[173] Zhang, et al., ARXIV:1801.05062V2, 1 (2018).

[174] Ghassemi, et al., ARXIV:1806.00388V2, 1 (2018).

[175] These include data complexity due to varying length, irregular sampling, lack of structured reporting and missing data and long-term dependencies between clinical events and disease diagnosis and treatment that creates complications to learning (see further on these Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 15 (2017).).

[176] Ghassemi, et al., ARXIV:1806.00388V2, 1 (2018). In this context it should be pointed out that clinical data is available in many different forms which can be relevant to understand patient health. Electronic health record is one of these types; other types may include for example scans (G. M. Weber, et al., *Fiinding the Missing Link for Big Biomedical Data* 311 JAMA (2014). Other data in acute care for example can include laboratory tests, written notes, vital signs, high frequency data sampled hundreds of times per second and static demographic data (as noted by Ghassemi, et al., ARXIV:1806.00388V2, 2 (2018).).

[177] Analysis that follows and references were obtained from Ghassemi, et al., ARXIV:1806.00388V2, 2 (2018).

[178] Id.

[179] Id.

[180] H. L. Li-wei, et al., *Methods of Blood Pressure Measurement in the ICU*, 41 CRIT CARE MED. (2013).

[181] CW Hug, et al., *Clinician blood pressure documentation of stable intensive care patients: an intelligent archiving agent has a higher association with future hypotension*, 39 see id. at (2011).

they expect in the test;[182] for instance, the *absolute time* that a laboratory test occurs can be more predictive of patient health than the *value* of the test.[183] Note records and the interaction between a patient and the healthcare team is another challenge.[184] Normally, clinical notes are aimed to provide trained professionals a quick glance into important issues concerning the patient's condition.[185] However, even clinical natural language processing (NLP) packages designed to process clinical text can be deceived.[186]

Another challenge with the use of data in machine learning concerns the problem of missing data. The problem of missing data arises very often in practice.[187] As Ghassemi et al note in designing a learning algorithm in medicine, the sources of missingness must be carefully understood.[188] Moreover, as they interestingly emphasize the fundamental feature of missing data is that there may be information conveyed by the absence of an observation, and ignoring the correlation may lead to models that make wrong and even harmful predictions.[189] Multivariate time series data that are prevalent in a variety of practical applications in medicine very frequently carry missing observations due to different reasons including medical events, saving costs, inconvenience and anomalies.[190] The problem with such missing values and patterns is that they provide rich information about target labels in supervised learning tasks.[191] Additionally, recurrent neural networks (RNN) models that are using EHR data and often used for modeling diseases

[182] G. Hripcsak, et al., *Characterizing treatment pathways at scale using the OHDSI network*, 113 PROC NATL ACAD SCI U S A (2016).
[183] GM Weber & IS. Kohane, *Extracting physician group intelligence from electronic health records to support evidence based medicine.*, 8 PLOS ONE. (2013).
[184] Ghassemi, et al., ARXIV:1806.00388V2, 2 (2018).
[185] Id.
[186] G. K. Savova, et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*, 17 J AM MED INFORM ASSOC (2010). as noted by Ghassemi, et al., ARXIV:1806.00388V2, 2 (2018). It is interesting in this respect to point out what Ghassemi et la note namely that clinical NLP tool trained on big amounts of medical text may incorrectly identify many autistic patient records with cancer. As to how medical narrative information could be converted into relational database tables of patient information see N Sager, et al., *Natural language processing and the representation of clinical data*, 1 J AM MED INFORM ASSOC. (1994).
[187] Donald B. Rubin, *Inference and Missing Data*, 63 BIOMETRIKA, 581 (1976).
[188] Ghassemi, et al., ARXIV:1806.00388V2, 3 (2018).
[189] Id.
[190] Zhengping Che, et al., *Recurrent Neural Networks for Multivariate Time Series with Missing Values*, 8 SCIENTIFIC REPORTS, 1 (2018).
[191] Id.

and patient diagnosis in medicine do not systematically handle missing values in data.[192] Recent research has addressed this problem of missing data and proposed models to predict the missing values[193] and handle better the inherent characteristics of EHR data.[194] The manner of how this is done and the effectiveness of the process could prove challenging with practical medical and legal consequences as further explained below.

In addition to the challenges of learning from data, the costs of collecting these data is a hurdle in developing machine learning algorithms in medicine. Depending on the machine learning technique used, data needs to be gathered, "cleaned" of unreliable observations and checked for quality and then put into compatible formats for a unified database.[195] More data and diverse data often leads to the development of more accurate medical algorithms. Diverse and abundant data would have a greater capacity to identify complex implicit relationships in medicine.[196] These challenges as well as other legal restrictions in collection of data could render the development of machine learning algorithms more difficult. Thus, in assessing regulatory frameworks and liability regimes these unique hurdles causing sometimes limitations in ML medical predictions should be also considered as further explained in this paper.

8. The difficulty of choosing the "correct" features and indicating the "correct" labels

We have seen above, that different types of data as well as missing data pose challenges in machine learning. In addition, it remains a challenge in medicine what features are selected to best capture the complexity of a disease process.[197] It is primarily this challenge of collecting training examples with a set of (sufficiently) informative features that has limited the contribution of machine learning to complex tasks of prediction and

[192] Zhengping Che, et al., *Recurrent Neural Networks for Multivariate Time Series with Missing Values*, ARXIV:1606.01865, 6 (2016).
[193] Xenia Miscouridou, et al., *Deep Survival Analysis: Nonparametrics and Missingness*, 85 PROCEEDINGS OF MACHINE LEARNING RESEARCH (2018).
[194] Rajesh Ranganath, et al., *Deep Survival Analysis*, ARXIV:1608.02158 (2016).
[195] Price, HARV. J. L. & TECH. , 438 (2015).
[196] Id.
[197] Deo, CIRCULATION, 1921 (2015).

categorization in medicine.[198] The use of the growing amount of medical data enables one to ask evidence-based questions concerning the need and benefits of, for example, particular clinical interventions in critical-care settings across large populations.[199] It appears, that the challenges of collecting the "correct" data and extracting the "correct" features from these data could have substantial consequences in decision making in such critical-care interventions.

Therefore, the first task in training a machine learning algorithm is to identify some features or predictors.[200] Focusing on supervised learning (we will see unsupervised learning later), the fundamental question is how do we identify the "correct" features and we ensure that we do not miss out "important" ones? One simple way to choose features could be by identifying correlations between features and a disease or a medical event, for example a heart attack, and maintaining those that are significant.[201] However, such an approach could miss out a substantial number of features that might be useful to a group of patients who have had for example heart attack.[202] In some instances our limited understanding of a disease pathogenesis would render it unlikely that we are collecting all the correct features that are needed for accurate predictions.[203] Even worse, it could be that there features that are useful in combination with other features but not on their own.[204] It might be tempting in an attempt to resolve this problem to throw in all possible features but such an approach might make things even worse.[205] Hence why variable and feature selection has drawn the attention of much of research in cases where there are hundreds of thousands of variables in datasets that are used to train machine learning algorithms.[206]

---

[198] Id. at, 1923.
[199] M Wu, et al., *Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database*, 24 J AM MED INFORM ASSOC. , 488 (2017).
[200] Deo, CIRCULATION, 1921 (2015).
[201] Id.
[202] Id.
[203] Id. at, 1922.
[204] Id. at, 1921.
[205] Id.
[206] See Isabelle Guyon & Andre´ Elisseeff, *An Introduction to Variable and Feature Selection*, 3 J MACH LEARN RES 1157(2003).

Medical definitions and the behavior of clinicians in practice pose another distinct challenge in the development of machine learning algorithms in medicine. Medical definitions are working models based on the present scientific understanding of a disease.[207] However, as these scientific understandings evolve, so do these definitions.[208] Thus, in the context of machine learning, good predictive performance labels that are based on such definitions is only as good as the underlying criteria.[209] Similarly, it might be tempting to use the actual behavior of clinicians as the correct labels, but they could very well not be.[210]

Another challenge in this respect, is the link made between features and targets in machine learning. While this relationship between features and targets is an object of learning, information leakage can render the prediction useless.[211] For example, a machine learning algorithm could be designed to predict mortality of hospital patients using all available data until their death. If a machine learning algorithm is trained naively on predicting death when the ventilator is turned off in the preceding hour, such an algorithm would have high predictive performance, yet would have absolutely no clinical utility.[212] It is often the case that patients and/or their families decide to withdraw care at a terminal stage of illness.

It is often, and in most cases correctly, argued that the lack of training data is one of the major challenges in training machine learning algorithms. However, for example, in the area of radiology, this argument is partially true as most western hospitals have acquired millions of images as picture archiving and communication systems (PACS) have become a routine in the course of at least the last decade.[213] Thus, in such cases, the challenge is not the availability of image data but the acquisition of relevant annotations/labeling for these images.[214] For example, it is necessary to have a large amount of labeled data in

[207] Ghassemi, et al., ARXIV:1806.00388V2, 4 (2018).
[208] Id.
[209] Id.
[210] Id.
[211] Id.
[212] Id.
[213] Litjens, et al., ARXIV:1702.05747V2, 25 (2017).
[214] Id.

order to train a deep architecture.[215] There at least five major challenges in this regard. First, PACS normally store free-text reports by radiologists who explain their findings. Turning these reports into accurate annotations or structured labels in an automated manner necessitates complex text-mining methods, which is a field of study where deep learning is also used.[216] Thus, deep learning techniques also substantially contribute in this regard but the problem that is particularly associated with deep learning namely interpretability still remains. Secondly, even in cases where the data is annotated by domain experts, label noise (i.e. distorted/corrupted data) can be a substantial limiting factor in the development of algorithms.[217] For example, four different radiologists annotated pulmonary nodules in a widely used dataset[218] for evaluating image analysis algorithms to detect nodules in lung CT (computed tomography).[219] Interestingly, the number of nodules for which they did not unanimously agree on to be a nodule was three times greater than the number they fully agreed on.[220] This example shows another distinct limitation of applying machine learning in medicine compared to for example autonomous vehicles where in the latter case there would at least be fewer disagreements on the annotation/labeling of data. Thus, training deep learning algorithms in medicine on such noisy data could intensify the call for more interpretability as further explained below. Thirdly, often classification in medical imaging is presented as a binary task – normal versus abnormal.[221] However, such classification can be over simplified in many instances as both classes can be highly heterogeneous.[222] Fourthly, another data-related challenge is class imbalance.[223] In medical imaging, sometimes might be hard to find images for the abnormal class. For example, there are abundant of mammogram images but the majority of these images are normal and in cases where the mammogram does contain a suspicious lesions this is most of the times not cancerous and even most

[215] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 5(2017).
[216] Litjens, et al., ARXIV:1702.05747V2, 25 (2017).
[217] Id.
[218] LIDC-IDRI dataset; see S. G. Armato, 3rd, et al., *The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans*, 38 MED PHYS (2011).as noted by Litjens, et al., ARXIV:1702.05747V2, 25 (2017).
[219] Litjens, et al., ARXIV:1702.05747V2, 25 (2017).
[220] Id.
[221] Id.
[222] Id.
[223] Id. at, 26.

cancerous lesions will not lead to the death of the patient.[224] Finally, in many instances useful information is not just contained in the images themselves but physicians rely on a wealth of other data including patient's history, age, demographics and other factors to reach a better decision.[225] Thus, the weight that should be given to a machine learning prediction in reaching certain complex medical decisions could also pose challenges.

Labeling of data could pose another distinct challenge in the medical domain. A patient can be associated with multiple diagnoses simultaneously[226] as one patient may suffer from several connected illnesses.[227] Zhang et al. acknowledging this complexity, they propose a convolutional residual model for multi-label classification from doctor notes in EHR data.[228] They indicate that a comparison between their proposed model with several well-known baselines, to predict diagnosis based on doctor notes, showed the superiority of their proposed model.[229] The extent to which machine learning interpretability could also substantially contribute to the resolution of such complexities is examined below.

9. Machine learning bias and inherent tradeoffs

The issue of bias in artificial intelligence has been widely debated by different groups in different disciplines and fora.[230] The importance of examining algorithmic bias in medicine is that biased algorithms could be creating limitations in diagnosis and medical treatment. Having said that, as we will see below certain forms of bias could be beneficial to machine learning development.

---

[224] Id. In such cases where a class is unrepresented a typical technique used is the application of specific data augmentation algorithms to the unrepresented class.
[225] Id. Some researchers considered combining all these factors into deep learning networks but as they acknowledged the improvements were not as large as expected (see T. Kooi, et al., *Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network*, 44 MED PHYS (2017). as noted by Litjens, et al., ARXIV:1702.05747V2, 26 (2017).)
[226] E.g. cough, fever, and viral infection.
[227] Zhang, et al., ARXIV:1801.05062V2, 2 (2018).
[228] Id. at, 1.
[229] Id.
[230] See for example, issues raised by Margaret Mitchell, et al., *Model Cards for Model Reporting*, ARXIV:1810.03993V2 (2019).; Joy Buolamwini, *How I'm fighting bias in algorithms*(2017), *available at* https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?language=en.

There are three points that need to be set at the outset before embarking into further analysis. First, in many instances, it is even challenging to agree on what constitutes bias, what are the causes of bias and how bias could be rectified.[231] For example, a study on sex bias in graduate admissions to the University of California, Berkeley showed that measuring bias is more difficult than is usually assumed and the evidence is sometimes even contrary to expectation.[232] This challenge renders the task of finding appropriate solutions to the problem of machine learning bias even more challenging.[233] Sometimes,

[231] In this respect, DeepMind poses two open questions in developing artificial intelligence: What approaches are needed to fully understand biases in AI systems and data? What strategies should those designing AI systems use to counteract or minimize these effects? (DeepMind, *Privacy, transparency and fairness*, *available at* https://deepmind.com/applied/deepmind-ethics-society/research/privacy-transparency-and-fairness/.) Hacker for example, identifies two causes of bias in the context of machine learning: (1) biased training data and (2) unequal ground truth. He subdivides biased training data into two subcases: (1) incorrect handling of training data (e.g. in supervised machine learning having incorrectly labelled data as a result of implicit bias or sampling bias where some part of the population is unrepresented) and (2) historical bias (e.g. historically successful candidates to a UK medical school were predominantly white males). Regarding unequal ground truth, he refers to cases where, for example, risks or other target variables are unevenly distributed between protected groups (Philipp Hacker, *Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law*, 55 CMLR 1143, 1146-1150 (2018).; He also distinguishes between direct and indirect discrimination under EU law in the context of machine learning algorithms (see id. at, 1151-1154.). For another approach, see Google's crash course on fairness, explaining different types of human bias that can be inadvertently reproduced by machine learning algorithms and how to identify them and evaluate their effect at https://developers.google.com/machine-learning/crash-course/fairness/video-lecture.

[232] P. J. Bickel, et al., *Sex Bias in Graduate Admissions: Data from Berkeley*, 187 SCIENCE (1975). Another example includes the COMPAS algorithm that predicts the recidivism risk of criminal defendants; questions were raised on whether this algorithm discriminates against black people (see State v. Loomis 881 N.W.2d 749, (Wis. 2016).; Note, *State v. Loomis*, 130 HARV. L. REV. 1530(2017).; Jeff Larson, et al., *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublica(2016), *available at* https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.; William Dieterich, et al., *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*, Northpoint Inc (2016), *available at* http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.). In this context it is generally argued that recidivism is predicted at higher rates among certain minority groups in the US (see Julia Angwin, et al., *Machine Bias*, ProPublica(2016), *available at* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.); Nabi et al. ask to what extent are these predictions discriminatory? (Razieh Nabi & Ilya Shpitser, *Fair Inference On Outcomes*, ARXIV:1705.10378V4 1(2018).

[233] For example, as to the ways to rectify discrimination, Hacker refers to a number of fairness criteria that have been advanced in algorithmic fairness research and distinguishes between two main concepts: (1) individual fairness and (2) group fairness. Some of these fairness criteria are mutually incompatible e.g. full individual and full group fairness; he argues that in the case where an algorithmic procedure is found to be discriminatory then bias minimization strategies would be applied; He divides these strategies into three different approaches: pre-processing approaches that modify the input data; in-processing approaches that aim to control the mapping from input to output data; and post-processing approaches that aim to transform the algorithmic output into a fair representation; he argues that in applying these techniques a careful proportionality assessment must be undertaken (Hacker, CMLR, 1175-1177 and 1182-1183 (2018).); Regarding biased training data, he argues that mitigation of this bias will improve the predictive accuracy; however, regarding cases where the algorithm correctly picks up features that are substantially correlated with a particular protected group, reduction of discrimination will imply a decrease in predictive accuracy (id. at, 1183-1184.)

the term fairness is used to denote, in the context of machine learning, the quest for designing algorithms that don't learn to be bias and are "fairness-aware."[234] However, the selection of the appropriate fairness measures remains a deeply normative and challenging question.[235] Affirmative actions or the positive actions, which is the term used in EU anti-discrimination law, remain in many cases controversial and a challenging question for courts.[236] Secondly, as much as it is desirable to eliminate bias, it should be acknowledged that it would be impossible to ever completely eliminate it. Different forms of bias would be emerging in one form or another. As we will see below there are trade-offs between machine bias and machine accuracy. Some forms of bias, such as inductive bias, could be useful for successful learners. Thus, in certain cases, bias might be even desirable and in others a necessary evil in designing machine learning algorithms. Still, we should be reminded that machine learning bias in medicine could be acting as a limitation that could translate to detrimental diagnosis and treatment. Hence why, it becomes of paramount importance to address such limitations but at the same time understand the complex trade-offs in this respect. It is in this context that the question arises whether machine learning interpretability could provide a constructive solution to such problems. The advantage of ML interpretability (i.e. ML system providing an explanation for its prediction) is that it provides certain safeguards. Consequently, it allows for the deployment of machine learning algorithms in medicine rather than restraining their deployment until the "perfect" algorithm is designed. This issue will be further analyzed below when discussing machine learning interpretability and warning obligations.

We have seen above that machine learning algorithms have the ability to learn from data. The argument that is usually presented in this regard is that in the case where the datasets themselves are biased, the machine learning algorithms also become biased.[237] In

[234] Friedler, et al., ARXIV.ORG/PDF/1609.07236.PDF, (2016).
[235] Hacker, CMLR, 1183-1185 (2018).
[236] Regarding for example, the EU law challenges see explanations by Hacker id. at, 1179-1183.
[237] For example, in the medical domain, it has been argued that the medical datasets used by AI researchers are notoriously biased and that health care data is extremely male and extremely white (Dave Gershgorn, *If AI is going to be the world's doctor, it needs better textbooks*(2018), *available at* https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.). In this context see also Obama administration report on Big Data that maps pathways for fairness and opportunity

healthcare, this means that biased predictions would provide better medical treatment to one group of people than another.[238] For example, a machine learning study on predicting pneumonia revealed that the machine learning system was giving a lower risk score to patient who also have asthma.[239] However, in reality the data were biased as the patients in the lower risk group were given extra care hence the inaccurate results.[240] Ahmad et al. make two interesting observations in this respect.[241] First, problems like these get lost in black box machine learning systems. Secondly, in medical image diagnosis where deep learning techniques are often used with excellent predictive power could be fooled into making mistakes which human experts would never make. These observations provide additional support for the need of some form of machine learning interpretability especially in the medical domain.

Another interesting aspect in discussing bias is that many medical datasets are not as such biased but imbalanced. They are imbalanced as they simply reflect the population that suffers from a given condition.[242] For example, as we already seen, in the field of mammography, there are abundant mammogram images but the majority of these images are normal and in cases where the mammogram does contain a suspicious lesions this is most of the times not cancerous and even the most cancerous lesions will not lead to the death of the patient.[243] As we have also seen, in such cases where a class is unrepresented a typical technique used in machine learning is the application of specific data augmentation algorithms (i.e. a technique that artificially expands the size of a training

but also cautions against re-encoding bias and discrimination into algorithmic systems (Megan Smith, et al., *Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights*(2016), *available at* https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights.).

[238] Gershgorn, 7. 2018. In support of this argument Gershgorn notes that in 2017 a Stanford study claimed that an AI system was more accurate than dermatologists in diagnosis of malignant skin lesions from images. However, he states that the datasets of malignant and benign melanoma used overwhelmingly featured lighter skin (at 10).

[239] Muhammad Aurangzeb Ahmad, et al., *Interpretable Machine Learning in Healthcare*, *in* PROCEEDINGS OF THE 2018 ACM INTERNATIONAL CONFERENCE ON BIOINFORMATICS, COMPUTATIONAL BIOLOGY, AND HEALTH INFORMATICS 560, (2018).

[240] Id.

[241] Id.

[242] Gershgorn, 11. 2018.

[243] Litjens, et al., ARXIV:1702.05747V2, 26 (2017).

dataset by creating modified versions of images in the dataset)[244] to the unrepresented class.[245] However, such techniques might also have the their own drawbacks and incorporate their own biases.

Machine learning research focused on different approaches in tackling different forms of bias. Kusner et al. for example, presented a model of fairness that they refer to as counterfactual fairness.[246] This model enabled them to develop algorithms that are able to consider the various social biases that may arise towards individuals based on ethically sensitive attributes and compensate for these biases effectively.[247] They distinguish this type of algorithm from other algorithms that tackle bias by simply ignoring protected attributes such as gender, race or religion. Nabi and Shpitser proposed in their work to model discrimination based on "sensitive features," such as race or gender in relation to an outcome.[248] A growing community is now examining matters of fairness and transparency in data analysis by addressing harmful effects of algorithmic bias from a variety of perspectives and frameworks.[249] Thus, researchers are already addressing some of the factors that could be causing bias but it is also evident that it is not easy to pinpoint to one perfect technique as each technique has its own drawbacks. Furthermore, each technique could be incorporating new and different forms of bias. As Hardt et al. point out, a naïve approach might require to design an algorithm that ignores all protected

[244] Jason Brownlee, *How to Configure Image Data Augmentation When Training Deep Learning Neural Networks*(2019), *available at* https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/.
[245] Litjens, et al., ARXIV:1702.05747V2, (2017).
[246] Matt J. Kusner, et al., *Counterfactual Fairness*, ARXIV:1703.06856V3 (2017). Their definition of counterfactual fairness incorporates the intuition that "a decision is fair towards an individual if it at the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group" (at 1).
[247] Id. at, 9. See also other research work that tackles discrimination by developing algorithms incorporating other techniques including Niki Kilbertus, et al., *Avoiding Discrimination through Causal Reasoning*, ARXIV:1706.02744V2 (2018).); see also Jon Kleinberg, et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV:1609.05807V2 (2016).
[248] Nabi & Shpitser, ARXIV:1705.10378V4 1(2018).
[249] Id. See Sam Corbett-Davies, et al., *Algorithmic decision making and the cost of fairness*, ARXIV:1701.08230V4 (2017).; Michael Feldman, et al., *Certifying and removing disparate impact*, ARXIV:1412.3756V3 (2015).; Moritz Hardt, et al., *Equality of Opportunity in Supervised Learning*, ARXIV:1610.02413V1 (2016).; F. Kamiran, et al., *Quantifying explainable discrimination and removing illegal discrimination in automated decision making*, 35 KNOWLEDGE AND INFORMATION SYSTEMS 613(2013).

attributes such as a race, color, religion, gender, disability or family status.[250] Such an approach, they point out, that is based on the philosophy of "fairness through unawareness," is not helpful due to the existence of "redundant encodings", ways of predicting protected attributes from other features.[251] Furthermore, bias in machine learning can arise as much from human choices on how they design or train an algorithm as they can from human errors in judgment when interpreting the predictions.[252] Hence why, different forms of biases in machine learning would always be present in one way or another and thus the question is whether there are mechanisms that could provide novel solutions to such problems.

The contrast between human and machine learning biases is another interesting aspect in discussing machine bias. Nabi and Shpitser in examining the extent to which a specific approach is "fair" they note that the gold standard is human intuition.[253] They argued that it is unfortunate that data analysis is based on statistical models that do not by default encode human intuitions about fairness and bias.[254] However, the question arises whether human intuition about fairness and bias is or should be the baseline to assess machine learning bias. Hardt et al.[255] argue that reliance on data can aid in quantifying and eliminating existing biases but they still point out that some scholars warned that algorithms can also introduce new biases or continue existing ones.[256]

---

[250] Hardt, et al., ARXIV:1610.02413V1, 1 (2016).
[251] Dino Pedreshi, et al., *Discrimination-aware data mining*, *in* PROCEEDINGS OF THE 14TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (2008). As noted by Hardt, et al., ARXIV:1610.02413V1, 1 (2016).
[252] Regarding the interpretation of the predictions, it refers to, for example, how administrative agents and other decision-makers use these prediction outputs to implement different policies and everyday practices. See further, AI Now Institute, *Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies*(2018), *available at* https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde.
[253] Nabi & Shpitser, ARXIV:1705.10378V4 3(2018).
[254] Id. at, 1.
[255] Hardt, et al., ARXIV:1610.02413V1, 1 (2016).
[256] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIFORNIA LAW REVIEW 671(2016). See also Burrell who also argues that the claim that algorithms will classify more "objectively" cannot be taken at face value considering the degree of human that is still incorporated in designing the algorithms (Burrell, BIG DATA & SOCIETY, 3 (2016).). However, in this respect, at least as far as deep learning is concerned, as LeCun et al. noted the crucial aspect of "deep learning is that [t]he layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure;" as LeCun et al. note deep learning will have many more successes in the near future as it necessitates very little engineering by hand (LeCun, et al., NATURE, 436 (2015).).

Furthermore, supervised machine learning algorithms present another different type of bias challenge, the so-called, bias-variance tradeoff (problem). Even if we manage to pass the challenge of collecting all the correct features, we still need some function to combine them to achieve the desired task.[257] As we mentioned above, even the desired task itself could sometimes be controversial in medicine – for example in training an algorithm, is the desired task the prolongation of the life of the patient, ensuring a good quality of life for the patient or a balance of the two? Be that as it may, assuming a clear target can be set, the goal of supervised machine learning algorithms is to find the best possible mapping function (or target function) (f) for the output variable (y) given the input data (x).[258] Any machine learning algorithm will have a prediction error that can be divided into three categories: bias error, variance error, irreducible error.[259] The irreducible error cannot be reduced irrespective of the algorithm used. Bias are the simplifying assumptions made by the model in order to make the target function less difficult to learn.[260] Thus, a model with high bias relies very little on the training data and oversimplifies the model.[261] Variance is the amount that the estimate of the target function will change in the case where a different training dataset was used.[262] Thus, a model with high variance heavily relies on the training data and does not generalize well on data which has not seen before.[263] In supervised learning, models that usually have high bias and low variance will underfit. Underfitting occurs, when a model is not able to capture the underlying pattern of the data.[264] Basically, the model is not able to obtain a satisfactorily low error value on the training set.[265] In such instances, we will need to

[257] Deo, CIRCULATION, 1922 (2015).
[258] Jason Brownlee, *Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning*(2016), *available at* https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/.
[259] Id.
[260] Id.
[261] Seema Singh, *Understanding the Bias-Variance Tradeoff*(2018), *available at* https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229.
[262] Brownlee. 2016.
[263] Singh. 2018.
[264] Id.
[265] GOODFELLOW, et al., 109. 2016.

estimate the training error and use a loss function that is adapted to reflect what type of errors are more tolerable than others.[266]

However, even if we manage to minimize the training error we might still have an algorithm with bad generalization ability. In other words, the algorithm might be still bad in giving the right medical diagnosis or treatment for new cases that it has never seen before. Models that usually have low bias and high variance will overfit. Overfitting occurs when the model captures the noise (artifacts) along with the underlying pattern in data.[267] In other words, overfitting takes place when the model learns the details and noise in the training data to the extent that it detrimentally affects the performance of the model on new data.[268] Consequently, the gap between the training error and the test error[269] would be too large.[270] The goal of supervised machine learning algorithms is to achieve a low bias and a low variance. However, the problem is that increasing bias will decrease variance and vice-versa. In practice, we cannot calculate the real bias and variance error terms as we do not know the actual underlying target function.[271] However, both bias and variance provide the tools to comprehend the behavior of machine learning algorithms in the quest of predictive performance.[272] It is evident from the above, that there are multiple considerations that need to be taken into account in striking the best possible balance and that the perfect algorithm cannot exist.

In this context, it is also interesting to refer to the "no free lunch theorem" we introduced above, that suggests that we must design a machine learning algorithm to perform well on a particular task.[273] The way to do this, is by building a set of preference into the learning algorithm.[274] When these preferences are aligned with the learning problems we need to solve, then this learning algorithm performs better.[275] A learning algorithm could

[266] Deo, CIRCULATION, (2015).
[267] Singh. 2018.
[268] Brownlee. 2016.
[269] Test error is the accuracy of predictions on a dataset that the algorithm has never seen to before.
[270] GOODFELLOW, et al., 110. 2016.
[271] Brownlee. 2016.
[272] Id.
[273] GOODFELLOW, et al., 116. 2016.
[274] Id.
[275] Id.

be given a preference for one solution over another in its hypothesis space.[276] This means that both functions are eligible but we show preference to one of them.[277] Additionally, we may restrict the learner algorithm to choose from only a particular set of predictors.[278] Such restrictions are what we refer to as inductive bias as we bias the learner towards a particular set of predictors.[279] The choice of such predictors is determined before the learner algorithm sees the training data consequently this choice should ideally be founded on some prior knowledge about the problem to be learned.[280] On the one hand, choosing a more limited hypothesis group would probably avoid overfitting, on the other, it might cause a stronger inductive bias. [281] Hence why a balance needs to be struck. Some forms of bias are not necessarily bad, on the contrary, certain forms of bias could be even beneficial for a successful learner. As we have seen when discussing inductive bias, the incorporation of prior knowledge, biasing the learning process is inevitable to successfully design a learning algorithm.[282] However, whereas human bias in some instances could be exposed or become explainable when humans provide justifications for their decisions, machine bias would be hidden. The medical domain is a particularly sensitive domain as bias could be indicating for example which patient should be prioritized in an intensive care unit or which one should get a transplant. Products liability law, provides a framework to identify product warning defects. However, the question remains whether and to extent this framework is fit for machine learning algorithms in medicine. Could for instance the current legal framework on product warnings address the question of machine learning bias and other inherent limitations? If not, how could warnings be adapted to address these challenges? As Goodfellow explains the "ideal [ML] model is an oracle that simply knows the true probability distribution that generates the data. Even

[276] Id. at, 116.
[277] Id. at, 117. There are many ways of expressing preferences for different solutions and these different approaches are known as regularization; Goodfellow et al. define regularization as "any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error" (see at 117). Moreover, as Goodfellow et al. argue foundational concepts in the field of statistics such as parameter estimation, bias and variance are useful to characterize notions of generalization, underfitting and overfitting (at 120).
[278] Shalev-Shwartz & Ben-David, 16 (2014).
[279] Id.
[280] Id.
[281] Id. at, 17.
[282] Id. at, 116.

such a model will still incur some error on many problems."[283] It is in this challenging environment in the medical domain that this paper examines the type of warnings appropriate for ML medical devices.

10. Diverse medical opinions - the example of intensive care units

Real time prediction of clinical interventions is one of main challenges faced by clinicians within intensive care units (ICUs).[284] In the ICU setting clinical decision-making takes place in an environment of limited knowledge and high uncertainty; for instance, only 10 of the 72 ICU interventions that were evaluated in randomized control trials have shown to relate to improve outcomes.[285] This task gets even more complicated due to data sources that are sparse, heterogenous, noise and outcomes that are imbalanced.[286] In these complex environments, different machine learning techniques could be of use in order to predict the onset and weaning of multiple invasive interventions.[287] For example, prolonged dependence on mechanical ventilation or premature extubation are related to increase complications of the health of a patient and higher costs.[288] It is interesting for the purposes of this paper that clinical opinion on the most appropriate protocol to be followed for weaning patients off of a ventilator differs.[289] Thus, an additional opinion by a machine learning algorithm to the already diverse physicians' opinions could place some physicians in the situation where they would need to choose between the machine learning algorithm's prediction of what needs to be done or their own initial decision.[290] In this respect, Prasad et al. have worked on developing a machine learning decision

[283] They note that the error incurred by an oracle making predictions from the true distribution p(x,y) is called the Bayes error; GOODFELLOW, et al., 114. 2016.
[284] Suresh, et al., ARXIV:1705.08498V1, 1 (2017).
[285] Gustavo A Ospina-Tascón, et al., *Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail?*, 36 CRITICAL CARE MEDICINE (2008). As noted by Suresh, et al., ARXIV:1705.08498V1, 2 (2017).
[286] Suresh, et al., ARXIV:1705.08498V1, 1 (2017).
[287] In particular, machine learning could be used to predict intervention tasks concerning invasive ventilation, non-invasive ventilation, vasopressors, colloid boluses and crystalloid boluses; See work done by Suresh id.
[288] Niranjani Prasad, et al., *A reinforcement learning approach to weaning of mechanical ventilation in intensive care units*, ARXIV PREPRINT ARXIV:1704.06300, 1 (2017).
[289] Id.
[290] A key question would be how machine learning algorithms would impact medical malpractice liability of clinicians; how should tort liability should apply to providers who are not aware of basis of the treatment they recommend? (Price, Medical Malpractice and Black-Box Medicine 295. 2018.

support tool to predict the time-to-extubate readiness of a patient and to recommend a personalized regime of sedations dosage and ventilator support.[291] The aim of such a tool is to recommend a personalized treatment protocol.[292] Specifically, they used off-policy reinforcement learning algorithms to determine the most appropriate action to be taken at a given patient state from sub-optimal historical ICU data.[293] They argue that their paper adopts a novel approach as they incorporate a larger number of possible predictors of weaning readiness in a 32-dimensional patient state representation compared with previous research that normally limited features for classification to only a couple of vital signs.[294] Moreover, they make use of existing clinical protocols to inform the design and tuning of a reward function. In this respect it is interesting to observe that machine learning systems, as the one in question, are aimed as decision support tools rather than autonomous machines. However, even such *assisting* medical machine learning tools should be distinguished from conventional assisting medical systems or software. Machine learning medical tools do not only provide information or carry out a simple processing of information but carry out an assessment themselves and provide an opinion that is conventionally reserved for the physician. In other words, the physician would not be simply provided with additional information to aid her assessment but would be presented with an "opinion" as to what should be done. Consequently, the physician would either have to disregard this opinion or execute it. Thus, we might be faced with situations where the physician could be the one who mechanically executes the "opinion" of the machine. This raises interesting questions on medical malpractice, question on consent and products liability law.

Moreover, there are challenges that need to be addressed or at least be aware of when building machine learning tools for an ICU. For example, there are factors that could potentially influence the patient's readiness for extubation, including some that are not noted in ICU chart data, such as patient's inability to protect their airway due to muscle

[291] Prasad, et al., ARXIV PREPRINT ARXIV:1704.06300, (2017).
[292] Id. at, 1.
[293] Id. Reinforcement learning has been also explored in other areas including the sequence of drugs to be administered in HIV therapy or cancer treatment; managing anemia in hemodialysis patients and regulating insulin in diabetics; this process is mainly based on estimating the value in terms of clinical outcomes of various treatment decisions given the state of the patient (id. at, 2.)
[294] Id. at, 3.

weakens.[295] Additionally, there are a great variety of sedatives and ventilator settings that can be leveraged during weaning.[296] Moreover, past treatment and trajectories, in identifying a successful extubation at time *t,* provides us only with an upper bound on the true time to extubate; on the other hand if a breathing trial was conducted and showed that it was unsuccessful, it leads to uncertainty how premature the intervention was.[297] The above, indicate the difficulties both when learning the policy at training stage and when evaluating the policy.[298] It appears that there are multiple interdependent complex factors affecting a decision to be taken at a particular time in an ICU. One the one hand, machine learning technology is particularly suited to consider all these factors, draw patterns and make extremely helpful suggestions. On the other hand, the problem remains, how does a physician properly assess the ML "opinion" which might be even be contrary to her medical opinion on what needs to be done? As we have seen above, a machine learning algorithm irrespective of how accurate has proved when tested still lacks, for example, "common sense". But one may ask, is a physician's common sense always useful? These and other similar questions would arise in such settings where opinions vary and the machine learning algorithm would be providing an additional opinion. While, the physicians would be able to justify their opinions, the machine learning algorithm would not be able to do that. We would of course know how well the ML algorithm performed when tested (i.e. how well it generalized) but still we won't know why the ML reaches a particular prediction. In a setting such as the one at an ICU where medical opinions might vary and factors to be considered might not be clear, the need for some form of ML explanation on the prediction made might be even greater. Regarding the novel interaction between the physician and the AI medical devices, Price makes an interesting comparison between clinical decision-support software[299] that is designed to help physicians diagnose and treat patients with what he calls "black-box medicine."[300]

---

[295] Id. at, 2.
[296] Id.
[297] Id.
[298] Id.
[299] See Randolph A. Miller & Sarah M. Miller, *Legal and regulatory issues related to the use of clinical software in health care delivery*, *in* CLINICAL DECISION SUPPORT 424, (Robert Greenes ed. 2006). As noted by Price, Medical Malpractice and Black-Box Medicine 300. 2018.
[300] By "black-box medicine" he refers to the combination between the exponential wealth of health data and rapid development of machine-learning algorithms that enable this new form of medicine ("black-box medicine"). See Price, Medical Malpractice and Black-Box Medicine 295. 2018.

In this regard, Miller and Miller point out, that clinical decision-support software solely "augments the physician's existing knowledge by providing further information."[301] Consequently, Price argues, that the software provides information but the physician intervenes to make the final choice.[302] He argues that this knowledgeable intervention is precisely what is different about black-box medicine. Because neither the providers nor the developers know the relationships underlying the recommendations of black-box medicine, the physician cannot be at the final step of the process of care. Thus, he continues, once the physician decided to use a particular black-box algorithm then the physician cannot understand and thus verify the algorithms recommendation against the physician's body of substantive expertise; at this stage, the physician has a choice of either accepting or not the algorithm's recommendation.[303] He points out, that this challenge might become even more complex when the algorithm suggests taking an unrelated drug based on previously unknown secondary effects or modifying a drug's dosage or schedule without conforming to existing medical knowledge. In such instances, he concludes, that imposing the same standard of negligence would make little sense.[304] Indeed, it appears that traditional legal frameworks in regulating the interaction between physician, producer of conventional medical devices and patient in certain aspects might not be fit.

The objective of minimizing both the training and test errors in order to avoid illusionary successful prediction capacities poses another interesting challenge. A successful machine learning algorithm should have the ability to generalize well and thus be able to make successful predictions or categorizations for cases that it has never seen before. A learning algorithm might have low training error during training but that does not necessarily mean it would also have to a low-test error when exposed to completely new cases it has never seen before. Generally, models that are highly complex (including those with a huge number of features) may perform better at minimizing training error but tend to generalize poorly as they tend to overfit to the data.[305] In other words there is a balance that needs to be struck. On one the hand, we might have complex models (including the

[301] Miller & Miller, 433. 2006.
[302] Price, Medical Malpractice and Black-Box Medicine 300. 2018.
[303] Id. at, 300-301.
[304] Id. at, 301.
[305] Deo, CIRCULATION, 1922 (2015).

ones with huge number of features) that could be necessary for certain uses. On the other hand, that models generalize well to new data sets.[306] Therefore, it appears that this delicate balance might need to be coupled with some form of interpretability so the physician and patient have at least some idea on how the algorithm assesses a particular case.

Finally, another challenge in deploying ML devices in medicine, not only in the ICU context, concerns the one on the approval of ML medical devices. The approval of new drugs or other new treatments comes in several possible forms.[307] First, the treatment is generally scientifically comprehended.[308] Second, the use of clinical trials is to show the validity of a treatment method. Finally, the validity of the treatment can be confirmed by third parties as well other than the sponsoring company or other post market surveillance mechanisms. In this respect, it is argued that the validation of complex and implicit algorithmic models faces a number of challenges.[309] First, the black-box nature of the algorithms means that they cannot be comprehended on a scientific level. Regarding clinical trials, the implicit and complex nature of algorithms are unlikely to be fit for mechanistic exploration by classic clinical trial methodology;[310] also, some of the benefits of machine learning algorithms in medicine rely on avoiding a slow and costly clinical trial. Another hurdle concerns the role of FDA in validating medical algorithms. The FDA currently lacks the expertise and resources to independently replicate the manufacturers' results: at best it can provide some procedural oversight in ensuring that data collection, consolidation and analysis methods are suitable.[311]

[306] Id. at, 1923.
[307] Price, HARV. J. L. & TECH. , 440 (2015).
[308] Having said that, it should be pointed out that although the mechanisms of action of a drug is generally comprehended there are exceptions as well; for example, aspirin has been commonly available since the beginning of the twentieth century but its mechanism of action was only understood in 1971 (id.); It thus appears that the argument that the mechanism of medical algorithms is not transparent and this should subject them to more stringent rules might not be a strong argument.
[309] Id. at, 441.
[310] Because personalized medicine relies on particular patient profiles, it is difficult to aggregate similar patients (id.).
[311] Id. at, 442.

11. Warnings and machine learning in medicine

a. The relationship between the ML medical algorithm, the physician and the patient

There are a series of legal concerns arising out of the deployment of machine learning algorithms in medicine. Interestingly, we have seen that in the medical domain, unlike the vehicles one, the first fully-autonomous medical devices have already been deployed.[312]

There are at least three main ways that ML algorithms could be interacting with physicians and patients. First, the ML system could be autonomously taking certain decisions and performing certain tasks with regards to a patient; secondly, the ML system could be autonomously providing information to the physician concerning diagnosis, treatment or management of a patient; finally, the autonomous ML system could be providing medical information for example on diagnosis directly to the patient. In all three scenarios, sooner or later, there will be allegations that the ML prediction adversely affected the health of a patient and/or caused the patient's death. The question will then be raised, how should law deal with such issues?

Regarding the first scenario, in the context of autonomous vehicles, Geistfeld notes that to date, scholars reached the share conclusion that elimination of a human driver will shift responsibility onto manufacturers as a matter of products liability and that litigation would particularly focus on design or warning defects.[313] Similarly, it appears that liability issues concerning a fully autonomous ML system that for example takes a decision and then proceeds by itself to administer medication to a patient (scenario one above) would be examined under products liability law.[314]

---

[312] See first approved AI medical device, FDA press release - FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. April 11, 2018.
[313] Geistfeld, CALIF. L. REV., 1619 (2017).
[314] At EU level the products liability framework is governed by the products liability Directive (Council Directive (EEC) No 85/374 [1985] OJ L210/29). See short summary of the underlying principles at Commission Staff Working Document; Liability for emerging digital technologies; Accompanying the document - Communication from the Commission to the European Parliament, the European Council, the

Regarding the other two scenarios, the liability question could get more challenging as arguably information generated by ML algorithms could be assessed under policy considerations governing services or products.[315] Moreover, it has been considered whether algorithms should be subject to new legal frameworks beyond the traditional ones governing products and services.[316] Specifically, it has been considered whether ML systems used for medical decision-making should be given a unique legal status similar to personhood.[317] Be that as it may, if the ML algorithm provides advice to the physician on, for example, what dosage of medication to administer to a patient, such an activity would probably constitute a service. Having said that, it might still be argued that such ML specialized medical predictions given, for example, to a non-specialist physician or a nurse should be examined under similar policy considerations applicable to products.[318] In other words, it may be argued that policy considerations governing products liability are better suited for such ML medical predictions rather than policy considerations governing negligence that is applicable to services.[319]

Council, the European Economic and Social Committee and the Committee of the Regions; Artificial Intelligence for Europe (2018).). The Court of Justice of the EU (CJEU) held that the liability of a service provider when using products that fall within the scope of the Directive, such as in the course of providing treatment to a patient, does not the fall within the scope of the Directive (see *Centre hospitalier universitaire de Besançon*, EU:C:2011:869, paragraph 27). Similarly, neither in the US healthcare providers or healthcare facilities are held strictly liable for defects in the products they sell, use or provide (see Hollander v. Sandoz Pharm. Corp., 289 F. 3d 1193, (10th Cir. 2002). In this respect see Miller & Miller. 2006. who argues against applying strict products liability also for clinical decision-support software to providers and hospitals; as noted by Price, Medical Malpractice and Black-Box Medicine 298. 2018. Regarding the long-time immunity of software to product liability legal actions see Frances E Zollers, et al., *No more soft landings for software: Liability for defects in an industry that has come of age*, 21 SANTA CLARA COMPUTER & HIGH TECH. LJ (2004).

[315] In this respect see for example, Joseph L Reutiman, *Defective Information: Should Information Be a Product Subject to Products Liability Claims*, 22 CORNELL JL & PUB. POL'Y (2012).; Charles E Cantu, *A Continuing Whimsical Search for the True Meaning of the Term Products Liability Litigation*, 35. MARY'S LJ (2003).

[316] Karni A Chagal-Feferkorn, *Am I an Algorithm or a Product: When Products Liability Should Apply to Algorithmic Decision-Makers*, 30 STAN. L. & POL'Y REV. (2019).

[317] Consequently, under this reasoning a wrongful diagnosis would be based on medical malpractice than products liability (see for more details Jason Chung & Amanda Zink, *Hey Watson-Can I Sue You for Malpractice-Examining the Liability of Artificial Intelligence in Medicine*, 11 ASIA PACIFIC J. HEALTH L. & ETHICS (2017).

[318] In this context see *mutatis mutandis* Reutiman, CORNELL JL & PUB. POL'Y, (2012).

[319] See further Powers who examines a number of reasons that are usually raised in arguing that product cases are distinct from services (William C Powers Jr, *Distinguishing Between Products and Services in Strict Liability*, 62 NCL REV. (1983).; see also Alheit who looks at the products-services distinction in terms of causation (K Alheit, *The applicability of the EU Product Liability Directive to software*, 34 COMPARATIVE AND INTERNATIONAL LAW JOURNAL OF SOUTHERN AFRICA, 202 (2001)); Looking at the form of liability in

The question whether policy considerations governing services or products is a better basis to assess liability issues associated with ML medical information is an interesting and challenging question in itself. However, regarding warning defects, which is the subject of this paper, the elements that are considered and balances that are struck in examining the adequacy of warnings under either negligence or products liability law are substantially similar.[320] Therefore, this paper draws inspiration from the doctrine on products liability law.

It aims to identify a framework of warnings that is suitable for ML medical systems when carrying out medical tasks, such as providing medical information on, for example, diagnosis or treatment. The vast majority of states in the 1960s and 1970s have adopted the rule of strict products liability under which design and warning defects could be dealt with.[321] Consequently, the commercial distributor of a product could be subjected to strict liability for the physical harms proximately caused by a defect in the product.[322] Courts have adopted a definition of defect, which is the precondition for strict liability, that distinguishes between manufacturing, design and warning defects.[323]

terms of causation see also Westerdijk who argues that the question of causality in such cases also depends on the *foreseeable use* of that software (RJJ *Westerdijk Produkteaansprakelijkbeid voor software* (1995)243 as noted by K Alheit, *The applicability of the EU Product Liability Directive to software*, 34 COMPARATIVE AND INTERNATIONAL LAW JOURNAL OF SOUTHERN AFRICA, 207 (2001).); Traille argues that information is not a single category to which the same form of liability can be applied, the whole context in which the information is delivered should be considered in order to decide which form of liability should be applied. In Triaille's opinion intelligent software is 'information' that to which products liability law does not apply (J Triaille, "The EEC Directive on product liability and its application to databases and information" (1991) Computer Law and Practice 217 at 222 as noted by id. at, 206.); see *supra* Price who makes an interesting comparison between clinical decision-support software that is designed to help physicians diagnose and treat patients with what he calls "black-box medicine." (Price, Medical Malpractice and Black-Box Medicine 300. 2018.; finally, see Miller & Miller, 424. 2006.

[320] A negligence-based test for a service or a risk-utility test/consumer expectation test for a product would very much consider the same elements and balances. In this context, see early works by Dawn Pelletier, *Is There a Distinction between Strict Liability and Negligence in Failure to Warn Actions*, 15 SUFFOLK UL REV. (1981). and Richard L Cupp Jr & Danielle Polage, *The Rhetoric of Strict Products Liability Versus Negligence: An Empirical Analysis*, 77 NYUL REV. (2002).

[321] Geistfeld, CALIF. L. REV., 1632 (2017).

[322] RESTATEMENT (SECOND OF TORTS) §402A (AM. LAW INST. 1998); id.

[323] Id. Strict liability should be distinguished from absolute liability, as Supreme Court of California held, that "[f]rom its inception…strict liability has never been, and is not now, *absolute* liability" (Daly v. General Motors Corp., 20 Cal. 3d 725, (Cal. 1978).).

Additionally, the findings in this paper could provide inspiration to possible legislative initiatives on warnings. Appropriate warnings for ML medical systems would create a constructive relationship between ML medical systems (ML manufactures), physicians and patients that would provide better healthcare and encourage the speedier development and safe deployment of ML in medicine. Moreover, developing such a constructive relationship would shield the manufacturers of ML algorithms and physicians from uncertainty concerning their legal obligations that could be stifling machine learning innovation.[324] At the same time it would importantly provide a better healthcare serving the best interests of the patients.

Before embarking into a deeper analysis on warnings, it is useful to reiterate the distinction drawn at the outset between conventional products and ML medical systems. As explained above, a conventional medical device in effect mechanically perform tasks largely based on basic rules of physics, chemistry and biology. In contrast a ML system it has an inductive reasoning capability that allows it to generalize from "cases seen to infer information about new cases unseen."[325] Therefore, a ML medical system could, for example, be providing information to a general practitioner that is normally provided by a specialist physician. It is this intelligent character of the AI system that creates a new relationship between the AI system and the physician. This is also why a collaboration between an AI system and a physician was shown in many cases to outperform physicians or AI systems when either acting alone. This is also the reason why AI machines and physicians should not be in competitive but a cooperative relationship. However, there could be also challenges emerging out of this collaboration. On the basis of the

[324] The European Commission stated in its Communication on Artificial intelligence for Europe that in order to fully benefit from the opportunities presented by these emerging new technologies a clear and stable legal framework will stimulate investment and, in combination with research and innovation, will help bring the benefits of these technologies to business and citizens. It is also noted that it is necessary to examine whether the current rules at EU and national level for safety and liability are appropriate and whether for manufacturers and service providers the legal framework continues to deliver an adequate level of legal certainty (see further European Commission, 2. 2018.). Similarly, Geistfeld points out in the context of autonomous vehicles which could be also applicable in the medical domain the rate at which the market converts from conventional autonomous vehicles depends on the price that consumers are requested to pay in order to adopt the new technologies. He indicates two reasons, where systematic legal uncertainty about the manufacturer liability raises the cost of an autonomous vehicle, thereby increasing the price and reducing consumer demand for these new technologies (Geistfeld, CALIF. L. REV., 1617 (2017).).

[325] FINLAY & DIX, 32. 1996.

information provided by the AI system the general practitioner or nurse might not find it necessary to refer the patient to a specialist or proceed to carry out a medical procedure on a patient. This raises a number of issues concerning the liability of the manufacturer of such ML medical systems when these ML systems autonomously perform medical tasks such as diagnosis or prognosis. These ML systems could be affecting the current relationship between physicians, manufactures of ML systems and patients. In this regard, it is also interesting to refer to the distinction drawn by Price between conventional medical software and ML medical algorithms. As we saw, Price explains that a conventional medical software provides information but the physician intervenes to make the final choice.[326] As he points out this knowledgeable intervention is precisely what is different about black-box medicine. Because neither the providers nor the developers know the relationships underlying the recommendations of black-box medicine, the physician cannot be at the final step of the process of care. Thus, he continues, once the physician decided to use a particular black-box algorithm, then the physician cannot understand and thus verify the algorithms recommendation against the physician's body of substantive expertise; at this stage, the physician has a choice of either accepting or not the algorithm's recommendation.[327] He points out that this challenge might become even more complex when the algorithm suggests taking an unrelated drug based on previously unknown secondary effects or modifying a drug's dosage or schedule without conforming to existing medical knowledge.

Consequently, physicians and patients could be alleging that they were not aware of the potentials and limitations in a ML medical prediction due to lack of adequate information (warning). Specifically, the patient could be alleging that she did not give her informed consent to the physician as she did not understand the ML medical prediction on which the physician's actions were based. Therefore, the crux of matter concerns the type of information that a manufacturer should be providing to the physician considering that the physician would be acting as a learned intermediary[328] and considering that the

[326] Price, Medical Malpractice and Black-Box Medicine 300. 2018.
[327] Id. at, 300-301.
[328] Further explained below.

manufacturer would be held to the standard of an expert in the field.[329] Therefore, the manufacturer should be providing information (warning) concerning the ML medical prediction to the physician in a manner appropriate for her expertise that would allow the physician to obtain the patient's informed consent.[330] After the physician is given an adequate warning appropriate for her expertise, she should then "translate" this information for the patient in her verbal explanation in order to obtain the patient's informed consent.[331]

Identifying what should constitute appropriate information (warning) for ML medical devices, as noted above, would create a constructive relationship between ML medical systems (ML manufacturers), physicians and patients that would provide better healthcare and encourage the speedier development and deployment of machine learning in medicine.

b. Warnings and specificities in medicine

Designing warnings even for conventional products could be a hard task. As stated in Restatement (Third) of Torts on products liability (hereinafter Restatement (Third)) §2 comment *i*, "no easy guideline exists for courts to adopt in assessing the adequacy of product warnings and instructions."[332] Be that as it may, Restatement (Third) §2 comment *i* also states that warnings are necessary to:

"…allow the user or consumer to avoid the risk warned against by making an informed decision not to purchase or use the product at all and hence not to encounter the risk…warnings must be provided for inherent risks that reasonably foreseeable product users and consumers would reasonably deem material or significant in deciding whether to use or consume the product. Whether or not many persons would, when warned, nonetheless decide to use or consume the product, warnings are required to protect the

[329] Point raised by Mark Geistfeld (NYU Law School) in a discussion we had on this subject
[330] Id.
[331] Id.
[332] RESTATEMENT (THIRD) OF THE TORTS: PRODCUTS LIABILITY (1998).

interests of those reasonably foreseeable users or consumers who would, based on their own reasonable assessments of the risks and benefits, decline product use or consumption."[333]

The US Court of Appeals for the Ninth Circuit in *Davis* explained that "[w]hen, in a particular case, the risk qualitatively (e.g., of death or major disability) as well as quantitatively, on balance with the end sought to be achieved, is such as to call for a true choice judgment, medical or personal, the warning must be given."[334] In other words, products, such as ML medical devices, that encompass qualitative and quantitative medical and personal risks and hence requiring the exercise of a true choice judgment should be accompanied by adequate warnings. The warning helps to establish how the product will *actually* perform, which can be different from a more demanding expectation of how the product *should* perform.[335] If the design of the product was the cause of the product performing in an unreasonably dangerous manner, the actual performance would frustrate the consumer's reasonable expectation of how the product should have performed and this would constitute a design defect.[336]

---

[333] Id. at, § 2 cmt i. See also Watkins v. Ford Motor Co., 190 F. 3d 1213, (11th Cir. 1999).
[334] Davis v. Wyeth Laboratories, Inc., 399 F.2d 121 (9th Cir. 1968).
[335] Geistfeld, CALIF. L. REV., 1641 (2017). In this context, Geistfeld uses the example of a warning that a vehicle does not have an airbag. In the case of an accident, the airbag will obviously not deploy and this cannot be treated as a malfunction. However, while the consumer does not expect an airbag to deploy in the case of an accident, she still has reasonable expectations that the vehicle would have a functioning airbag if that design feature were necessary for the safe operation of the vehicle. Geistfeld notes that a warning that the vehicle provides no airbags would not defeat this reasonable expectation of safety. Showing that the omission renders the design unreasonably dangerous, the plaintiff would be also showing that this design frustrates the ordinary consumer's reasonable expectations of safe product performance. He indicates that some courts refer to this liability rule as the "modified" consumer expectation test in order to differentiate it from the (ordinary) consumer expectation test applicable to product malfunctions. In this regard he points out that, so formulated, the modified consumer expectation test is substantively equivalent to risk-utility test which is a cost-benefit examination that requires any design change with a disutility (or cost) that is less than the correlated reduction of risk (or safety benefit) (see RESTATEMENT (THIRD) OF THE TORTS: PRODCUTS LIABILITY § 2 cmt. d. 1998. (Geistfeld, CALIF. L. REV., 1642 (2017).). § 2 cmt. d of Restatement (Third) of the Torts provides for a reasonableness ("riskutility balancing") test; specifically, it states that "the test is whether a reasonable alternative design would, at a reasonable cost, have reduced the foreseeable risks of harm posed by the product and, if so, whether the omission of the alternative design by the seller or a predecessor in the distributive chain rendered the product not reasonably safe. (This is the primary, but not the exclusive, test for defective design)."
[336] Geistfeld, CALIF. L. REV., 1642 (2017).

It is therefore evident, that the aim of warning is to provide a clear and unambiguous information to the consumer in order to allow the consumer to make an informed decision whether to purchase or use the product (or in the context of certain ML medical systems, whether to follow the prediction of the ML algorithm). This quest is even more challenging considering that physicians and patients might have unrealistic expectations from ML devices. Even the glamor of the term "artificial intelligence" might be sometimes creating the unrealistic expectations. This risk appeared more than half a century ago, where an intermediate court in Louisiana considered whether mechanical robots driving vehicles (what we now often referred to as to autonomous vehicles) will be held to a higher standard of care than vehicles driven by humans.[337] It is interesting how the court distinguished the capabilities of machines with those of human drivers. It held that:

"A human being, no matter how efficient, is not a mechanical robot and does not possess the ability of a radar machine to discover danger before it becomes manifest. Some allowance, however slight, must be made for human frailties and for reaction, and if any allowance whatever is made for the fact that a human being must require a fraction of a second for reaction and then cannot respond with the mechanical speed and accuracy such as is found in modern mechanical devices, it must be realized that there was nothing that Reuter, a human being, could have done to have avoided the unfortunate result which the negligence of Mrs. Arnold brought upon herself."[338]

Therefore, the challenge concerns, how should the manufacturer of ML medical systems adequately warn physicians and consequently patients about the risks involved? In the context of autonomous vehicles, Geistfeld explains how the systemized driving behavior of autonomous vehicles can resolve the warning challenge.[339] He explains how the aggregate driving performance of the fleet could be providing the necessary data for warnings about the inherent risks of crash. This information would be also of particular use in designing adequate warnings in medicine. In addition to this information, in

[337] David C Vladeck, *Machines without principals: liability rules and artificial intelligence*, 89 WASH. L. REV., 130-131 (2014).
[338] Arnold v. Reuther, 92 So. 2d 593, (La. Ct. App. 1957).
[339] Geistfeld, CALIF. L. REV., 1654 (2017).

certain cases in medicine, it might be also necessary to accompany ML medical predictions with information that addresses specificities in medicine. As we have seen there are distinct challenges in developing and deploying ML systems in healthcare as well as tradeoffs inherent in medical decisions that need to be explained by the physician to the patient in order to obtain the patient's informed consent.

Regarding the challenges in developing ML systems in medicine, we have first seen that diagnosis, treatment and management of patients might not always constitute clear binary classification tasks. Defining the target in medicine when training ML algorithms might prove challenging especially in cases where targets are not clear and encompassing many tradeoffs. Therefore, on the one hand, ML systems might be of particular use in complex medical decisions but on the other, such complex medical decision involve a number of tradeoffs. Additionally, the selection of features in training ML algorithms could be a challenging task in medicine and that could reflect on, for example, the ML recommendation (prediction) for a treatment. For example, we saw that in some instances, clinical staff may have a feeling about the patient's state and preferentially record results, which could be later used as features to train algorithms, that concur with that understanding.[340] Furthermore, as we have seen in many cases there is no agreement on the labeling of the data used in training. Good predictive performance labels are based on certain medical definitions and they are only as good as the underlying criteria.[341] Similarly, it might be tempting to use the actual behavior of clinicians as the correct labels, but they could very well not be.[342] Additionally, the issue of warnings gets more challenging when it comes to medical devices based on deep learning architectures. As we have seen, although these architectures are delivering substantial improvements compared to conventional machine learning algorithms, many scientists and researchers remain skeptical of their application to the medical domain.[343] Deep learning models are used as a black-box without the researchers be able to explain why is in certain cases successful or without the ability to modify them in cases where there are misclassification

[340] Hug, et al., CRIT CARE MED. , (2011).
[341] Ghassemi, et al., ARXIV:1806.00388V2, 4 (2018).
[342] Id.
[343] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 16 (2017).

problems.[344] Also, many deep neural networks could easily reach wrong predictions when some noise is applied to a medical image.[345] Moreover, in order to train effectively deep learning models, it is require to have large sets of training data and thus rare diseases or events might not be well suited to deep learning.[346] Similarly, for many applications, raw data cannot be directly used as training input for deep neural networks.[347] Thus, in addition to the burden of transforming this raw data into appropriate training inputs, as explained above, labeling of these data carries its own separate challenges. Other ML challenges in medicine arise from the objective to make personalized predictions using large volumes of noisy and biased data.[348] We have also seen that human learners can rely on common sense to filter out random meaningless learning conclusions.[349] Whereas, when the task of leaning is carried out by a machine one should provide well defined principles that will shield the program from reaching useless or senseless conclusions.[350] This task carries its own challenges and has its own implications. It was also explained that machine learning in medicine includes helping physicians to choose between a selection of known interventions and even recommend an off-label use of an approved intervention. However, off-label recommendations pose their own challenges when the physician is given such a recommendation and needs to take a decision whether to go ahead with this recommendation. Physicians would have to take a decision in this respect despite that relationships underlying the recommendations of "black-box medicine" are not known.[351] Furthermore, the use of ML in wearable, implantable and ambient sensors are used to continuously monitor certain features related to health and wellbeing.[352] Such devices could have different legal implications on warnings as the physician might not be even present or aware when the ML prediction is given to the patient. In the case where a physician is present the warning implications might change as the learned intermediary doctrine could be applicable that limits recovery against manufacturers where doctors

[344] Id. at, 17.
[345] Id.
[346] Id.
[347] Id.
[348] Ghassemi, et al., ARXIV:1806.00388V2, 1 (2018).
[349] Shalev-Shwartz & Ben-David, 2 (2014).
[350] Id.
[351] Price, Medical Malpractice and Black-Box Medicine 300. 2018.
[352] Ravì, et al., IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS 13 (2017).

prescribe drugs or medical devices to patients.[353] This doctrine will be further examined below. Additionally, we have seen the problem of bias and fairness in designing ML algorithms as well as medical datasets that are not as such biased but imbalanced. Algorithmic bias in medicine, beyond fairness, could have different implications from other domains such as in cases where ML algorithms are used by public administrative bodies in determining, for example, whether certain social benefits should be granted or not to certain applicants. In the case of public bodies using ML systems to take administrative decisions, constitutional and administrative law would play a distinct role when deploying such systems. In healthcare, there could be similar but also other distinct sensitive balances that would need to be struck when deploying ML systems for medical tasks. Finally, in medicine, the physician together with the patient are discussing the benefits and risks for a proposed treatment in order to obtain the patient's informed consent. Patients' expectations from treatment outcomes might differ and different patients might be willing to take different risks.

Therefore, considering the ML medical specificities, how should ML warnings be designed? Providing the classification accuracy of an algorithm (the accuracy of the algorithm for its predictions when tested on data that it has never seen before) would be providing valuable information to the physician and consequently the patient. However, in many cases, this classification accuracy might not be giving adequate information (might too general and more precise information might be needed for the particular patient) for reasons explained above. In addition, testing the performance of an algorithm in the first place carries a number of challenges.[354] There are different metrics to evaluate the performance of ML algorithms and each carries its own drawbacks.[355]

---

[353] Price, Medical Malpractice and Black-Box Medicine 298. 2018. See Timothy S Hall, *Reimagining the Learned Intermediary Rule for the New Pharmeceutical Marketplace*, 35 SETON HALL L. REV. (2004).

[354] See tweet by Krysztof Geras (August 10, 2019); see also Yoshua Bengio & Yves Grandvalet, *No unbiased estimator of the variance of k-fold cross-validation*, 5 JOURNAL OF MACHINE LEARNING RESEARCH 1089(2004).

[355] They include logarithmic loss, area under the curve (AUC) and confusion matrix; see Aditya Mishra, *Metrics to Evaluate your Machine Learning Algorithm*(2018), *available at* https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. The metrics that are chosen to evaluate a machine learning model are very important as they influence how the performance of ML algorithms is measured and compared (see Mohammed Sunasra, *Performance Metrics for Classification problems in Machine Learning*(2017), *available at* https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.).

Sometimes, it is also argued that there should be more transparency in the form of providing, for example, the ML source codes and/or the data used in training. As useful as this information might be, in many cases, it would not be providing the requisite comprehensible information to the physician/nurse and consequently the patient. Neither a general warning about the inherent risks of using ML in medicine would fulfil the legal requirements of an adequate warning. In the context of pharmaceuticals, Geistfeld uses the example of a general warning concerning the inherent risk that a prescription drug will cause adverse side effects to the health of patients.[356] He points out, that providing such general warnings namely, that a particular drug could cause side effects, would not be considered adequate in the case where the manufacturer has precise information about the likelihood and consequences of the side effect.[357] This analogy and reasoning could be also applicable for ML devices used in medicine. Therefore, what additional information would need to be provided in warnings for certain ML medical predictions?

c. Explainable ML and confidence intervals

We have seen above that the aim of warnings is to provide a clear and unambiguous information in order to allow the patient to make an informed decision whether to follow or not the prediction of a ML algorithm. In cases where a physician acts as a learned intermediary the manufacturer should be providing information (warning) to the physician in a manner appropriate for her expertise that would allow the physician to "translate" this information for the patient and obtain the patient's informed consent. Therefore, the issue that needs to be addressed concerns the type of information (warning) that the manufacturer should provide to the physician.

Regarding warnings for prescription drugs, the Court of Appeals of New York held:

[356] Geistfeld, CALIF. L. REV., 1658 (2017).
[357] Id.

"A warning for a prescription drug may be held *adequate* as a matter of law if it provides *specific detailed information* on the risks of the drug…Always bearing in mind that the warning is to be read and *understood by physicians, not laypersons*, the factors to be considered in resolving this question include whether the warning is *accurate, clear, consistent on its face,* and whether it portrays with *sufficient intensity the risk involved in taking the drug*"[358] (citations omitted and emphasis added).

Regarding warnings for other products, the US Court of Appeals for the fifth Circuit held:

"A warning must (1) be designed so it can reasonably be expected to catch the attention of the consumer; (2) be comprehensible and give a *fair indication of the specific risks* involved with the product; and (3) be of an *intensity justified by the magnitude of the risk*"[359] (emphasis added).

Therefore, the warning should be providing information that is accurate, clear, consistent, comprehensible, giving fair indication of the specific risks and be of an intensity justified by the magnitude of the risk. It is useful in this regard to also bear in mind the distinction between failure-to-warn in the negligence context and failure-to-warn in strict liability. The Supreme Court of California held that in contrast to negligence law in a failure-to-warn:

"…Strict liability is not concerned with the standard of due care or the reasonableness of a manufacturer's conduct. The rules of strict liability require a plaintiff to prove only that the defendant did not adequately warn of a particular risk that was *known or knowable* in light of the generally recognized and prevailing best scientific and medical knowledge available *at the time of manufacture and distribution*. Thus, in strict liability, as opposed to negligence, the reasonableness of the defendant's failure to warn is immaterial"[360] (emphasis added).

---

[358] Martin v. Hacker 628 N.E. 2d 1308, (N.Y. 1993).
[359] Pavlides v. Galveston Yacht Basin, 727 F. 2d 330, (5th Cir. 1984).
[360] Anderson v. Owens-Corning Fiberglas Corp., 53 Cal. 3d 987, (Cal. 1991). In this respect, see also Owen who argued that "[W]hile acknowledging that 'strict' liability in design and warning cases is really nothing more than negligence, most courts continue to pretend that it really *is* something more. Thus, even in the

The requirement "known or knowable" creates a number of questions concerning the type of obligations imposed on ML manufacturers. To what extent should a manufacturer continue researching, testing or developing a ML device in order to reveal its true character so that the manufacturer can provide a genuine comprehensible warning to the physician (and consequently the patient) that would enable them to make an informed decision?[361]

The majority of States follow the principle expressed in Restatement (Second) of Torts §402A comment *j* on warnings.[362] Comment *j* provides that "the seller is required to give warning against [a danger], if he has knowledge, or by the application of reasonable, developed human skill and foresight *should have knowledge*, of the presence of the…danger" (emphasis added).[363] The Court in *Vassalo* pointed out that the Restatement (Third) of Torts §2(c) reaffirms the principle expressed in Restatement (Second) of Torts §402A comment *j*.[364] Restatement (Third) of Torts §2(c) states that a product "is defective because of inadequate instructions or warnings when the *foreseeable risks of harm* posed by the product *could have been reduced* or avoided by the provision of reasonable instructions or warnings…and the omission of the instructions or warnings renders the product not reasonably safe." Restatement (Third) of Torts comment *m* explains both the rationale behind this principle as well as the complex nature of foreseeability. Specifically, it states that the "issue of foreseeability of risk is more complex in the case of products such as prescription drugs, medical devices, and toxic chemicals. Risks attendant to use and consumption of these products may, indeed, be *unforeseeable at the time of sale*.

warning context, most courts still call liability 'strict' (DAVID G OWEN, PRODUCTS LIABILITY LAW 109 (2d ed. 2008).

[361] In this regard see also Geistfeld (MARK A GEISTFELD, PRODUCTS LIABILITY LAW 226-227 (Wolters Kluwer ed. 2012).) who argued that "[o]f course manufacturers are subject to negligence liability for failing to discover risks that would have been identified by reasonable research program, but is such a duty practically enforceable?…Without a credible threat of negligence liability, manufacturers do not have an adequate financial incentive for engaging in safety research. That incentive is restored by a rule of strict liability that makes manufacturers legally responsible for risks that have been discovered by the time of trial, even if they were not reasonably foreseeable at the time of sale."

[362] See Vassallo v. Baxter Healthcare Corp., 696 N.E.2d 909, (Mass. 1998).

[363] RESTATEMENT (SECOND) OF TORTS (1965).

[364] *Vassallo v. Baxter Healthcare Corp.*, N.E.2d at 922.

Unforeseeable risks arising from foreseeable product use or consumption by definition *cannot specifically be warned against*" (emphasis added).

Therefore, the crux of the matter is how do we reveal though warnings distinct risks associated with ML medical predictions? Are these distinct ML risks in the medical domain foreseeable risks? How far should the manufacturer (seller) proceed in research to discover these risks? As Restatement (Third) of Torts comment *m* provides "a seller bears responsibility to perform reasonable testing prior to marketing a product and to discover risks and risk-avoidance measures that such testing would reveal." In this regard, Geistfeld also notes, manufacturers under the rule of strict liability are "legally responsible for risks that have been discovered by the time of trial, even if they were not reasonably foreseeable at the time of sale [or use]."[365]

This paper considers whether explainable ML (i.e. ML delivering certain types of explanations for its specific prediction) and/or delivering a confidence interval (i.e. ML indicating its certainty for a specific prediction), in at least certain instances, would be revealing distinct risks associated with ML medical predictions. In other words, should in certain instances, a ML algorithm be providing explanations and/or a confidence interval for its specific prediction in order to provide adequate information (warning) to the physician/patient whether or not to follow the prediction of the ML algorithm?[366] The relationship between explainable ML systems/ML systems providing confidence intervals and warnings would be more extensively developed in further research that will complement this project. This paper sets below the foundations for the correlation between warnings and explainable ML/ML confidence intervals.

As we have seen above, a successful learner should be able to develop specific examples to broader generalization.[367] We indicated that the term generalization denotes how well a machine learning model can apply what it learned during its training to new cases that

[365] GEISTFELD, Products Liability Law 227. 2012.
[366] The point on the confidence interval was raised by Rajesh Ranganath (Courant Institute of Mathematical Sciences, NYU) in a discussion we had on these challenges.
[367] Shalev-Shwartz & Ben-David, 2 (2014).

has never seen before. This generalization ability of the machine learning algorithm could be measured and an accuracy value could then be provided. This accuracy value, in effect, provides a percentage value of the "correct" ML predictions. However, for reasons we have seen throughout this paper, solely providing the accuracy value in medicine might be sometimes giving a false sense of high performance or would not be always providing the "*specific detailed information* on the risks [involved]."[368] In this context, as already noted, the physician should be able to explain to the patient the course of treatment to be followed in order to obtain her informed consent.

Therefore, it appears that, in at least certain instances, a confidence interval would need to be also provided for the *specific* ML prediction made. The confidence interval of the algorithm could be of substantial importance especially for rare and complex cases. A low ML confidence interval for a particular case would be indicating a higher risk if the ML prediction is followed. However, in some instances it might be still argued that the confidence interval, similar to the accuracy value, will not clearly reveal all the risks involved in a ML prediction. The patient might still argue that the accuracy and confidence interval values did not provide the patient with the requisite information for a true choice judgment.[369] This could be particularly the case, for example, in medical instances that do not encompass clear binary choices or the different patients have different expectations from their treatment. Moreover, the confidence interval, like the accuracy value, might be affected by the parameters used in training the ML algorithm such as choice of features and labels used. Therefore, the confidence value might be also giving in some cases a false sense of high performance.

It is in such cases, that the algorithm might be required to also provide some explanations for its *specific* prediction.[370] In other words, an explainable ML algorithm that would be

[368] *Martin v. Hacker* N.E. 2d at 1312.
[369] *Davis v. Wyeth Laboratories, Inc.*, F.2d at 129.
[370] Google in its study entitled "Perspectives on Issues in AI Governance" also appears to support such an assertion. Specifically, it states that having an "explanation for why an AI system behaves in a certain way van be a big help in boosting people's confidence and trust in the accuracy and appropriateness of its prediction"; moreover it states that "[s]ystems being used to influence decisions of life-changing import, such as the choice of medical treatment, warrant much greater effort and depth of explanation than those performing tasks of minor consequence, such as making movie recommendations."

able to provide, for example, some explanation as to why it proposes a specific medical treatment.[371] This type of explanation, similar to the rest of the warnings, would be helping to establish how the product is *actually performing in that instance* when providing its prediction. Solely providing a prior warning on how the ML system is expected to perform, such as the one provided for conventional (mechanically based) medical devices, would not be always providing an adequate warning for ML medical systems. This is particularly useful for ML predictions since, as explained above, ML systems, in contrast to conventional medical devices, encompass an inductive reasoning ability (generalization ability) when making their predictions. This explanation would allow the physician to consider the ML prediction and explanations provided, assess the risks involved, combine it with her medical judgment and "translate" the information to the patient in order to obtain the patient's informed consent concerning the next steps to be followed. Having said that, there is still little agreement on what constitutes explainable machine learning and how interpretability should be measured.[372] Interpretability can be defined in different ways and can take many different forms.[373] For example, Ghassemi et al. argue that models in the medical domain should in certain instances provide justifiability; beyond explaining a specific prediction, models should strive towards justifying the predictive pathway.[374] Other forms of interpretability could include the ML algorithm indicating the area on a medical image that the algorithm is

[371] As Ghassemi et al. argue "[m]odels cannot be deployed 'in the wild' at low cost, and clinical staff must justify deviations in treatment to satisfy both clinical and legal requirements" (Ghassemi, et al., ARXIV:1806.00388V2, 7 (2018).).

[372] Finale Doshi-Velez & Been Kim, *Towards a rigorous science of interpretable machine learning*, ARXIV PREPRINT ARXIV:1702.08608 (2017).

[373] See among others, John Pavlus, *A New Approach to Understanding How Machines Think*, QUANTAMAGAZINE 2019.; Sebastian D Goodfellow, et al., Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings (2018).; Ahmad, et al. 2018.; Mitchell, et al., ARXIV:1810.03993V2, (2019).; Zachary C Lipton, *The mythos of model interpretability*, ARXIV PREPRINT ARXIV:1606.03490 (2016).; Derek Doran, et al., *What does explainable AI really mean? A new conceptualization of perspectives*, ARXIV PREPRINT ARXIV:1710.00794 (2017).; Andreas Holzinger, et al., *What do we need to build explainable AI systems for the medical domain?*, ARXIV PREPRINT ARXIV:1712.09923 (2017);W James Murdoch, et al., *Interpretable machine learning: definitions, methods, and applications*, ARXIV PREPRINT ARXIV:1901.04592 (2019).; Leilani H Gilpin, et al., Explaining explanations: An overview of interpretability of machine learning (IEEE 2018).

[374] Ghassemi, et al., ARXIV:1806.00388V2, 7 (2018). In this context, they also argue that machine learning work in healthcare also provides an opportunity to develop systems that interact and collaborate with human experts. They point out that clinical staff provide more than their expertise; empathy is recognized as an important element of clinical practice (Rita Charon, *Narrative medicine: a model for empathy, reflection, profession, and trust*, 286 JAMA (2001).; MOIRA STEWART, et al., PATIENT-CENTERED MEDICINE: TRANSFORMING THE CLINICAL METHOD (CRC Press. 2013).

looking at in making its prediction; or the algorithm providing more than one option and explaining the pros and cons for each option. When examining ML interpretability, it should be also pointed out that some accurate predictive ML models at the current stage of research might not be highly interpretable. Therefore, there is also a tradeoff in this respect that would need to be considered. In some cases, a balance might need to be struck between system accuracy and system interpretability.[375] These issues will be further examined in another paper.

Therefore, it appears that in certain cases a type of warning obligation that encompasses a form of interpretability and/or confidence interval for the specific prediction made, could be providing the physician (and consequently the patient) with the required "*specific detailed information* on the risks [involved]."[376] The patient would be able to better comprehend the ML prediction by considering the pros and cons of this prediction and consequently allowing her to make an informed choice whether or not to follow the ML prediction. For example, weighing the pros and cons of a ML recommendation to follow an aggressive chemotherapy, it would necessitate, in at least certain cases, some form of explanation why the ML is providing such a recommendation. The ML system providing some explanations for its prediction could be revealing hidden albeit foreseeable risks/benefits. The ordinary consumer (patient) expects from a warning to include disclosures that would materially improve her risk-utility decisions.[377] Especially, in medicine where there are, in many cases, inherent tradeoffs in medical recommendations, a warning that encompasses some ML explanations would improve her risk-utility decision and consequently allow her to provide informed consent for the proposed treatment. Interpretability could make a ML warning in medicine genuinely "comprehensible,"[378] "accurate and clear."[379] It would be providing the patient with an adequate warning particularly in cases when the risk is "death or major disability."[380] Both physicians and patients, obtaining an appropriate ML explanation and/or

[375] See Google, Perspectives on Issues in AI Governance at 9 and 11.
[376] *Martin v. Hacker* N.E. 2d at 1312.
[377] MARK A GEISTFELD, PRINCIPLES OF PRODUCTS LIABILITY 145 (Thomson Reuters/Foundation Press Second ed. 2011).
[378] *Pavlides v. Galveston Yacht Basin,* F. 2d at 338.
[379] *Martin v. Hacker* N.E. 2d at 1312.
[380] *Davis v. Wyeth Laboratories, Inc.*, F.2d at 129.

confidence value for each prediction, could enable them to determine whether or not they wish to follow the ML prediction or obtain another medical opinion, for example, in cases where the physician and the ML algorithm reached conflicting conclusions. In other cases, the explanation provided by the ML algorithm might persuade the physician to follow the ML's prediction even if the physician initially came to a different conclusion. Therefore, it appears that a warning obligation that encompasses a form of explanation (e.g. in some cases even just indicating the part of the medical image been observed by the ML system in reaching its prediction) and/or ML confidence interval would provide better healthcare, more trust in ML medical systems, better protection to the patients' interests and would better shield the manufacturer from liability. It would better protect the manufacturer from the allegation that the warning concerning the ML medical prediction was defective. Explainable ML algorithms and ML algorithms providing confidence intervals would be providing in certain instances the necessary link between warnings and the necessary information needed by the physicians to obtain the patient's informed consent. In this respect, it should be pointed out that the manufacturer could still be liable for design defects. This paper focusses on warnings. Designing defects would be subject to a risk-utility test.[381] Moreover, the patient could have recourse to more claims and/or grounds to base her claims. The patient could also bring legal actions against medical practitioners. Additionally, litigation could involve claims concerning misleading direct-to-consumer advertising, concealment or delay in reporting data or provision of misleading or fraudulent data to the responsible bodies as well as fraud or misrepresentation.[382]

A final point to be raised in this respect emerges from the judgment of the Supreme Judicial Court of Massachusetts in *Vassallo* summarizing the obligations to warn under

[381] Howard Latin, *Good Warnings, Bad Productions, and Cognitive Limitations* 41 UCLA L. REV. 1193(1994). as noted by GEISTFELD, Principles of Products Liability 164. 2011. Restatement (Third) §3(b) adopts a reasonableness ("risk-utility balancing") test as the standard for judging the defectiveness of product designs. As Restatement (Third) comment *d* further explains the "test is whether a reasonable alternative design would, at reasonable costs, have reduced the foreseeable risks of harm posed by the product and, if so, whether the omission of the alternative design by the seller or a predecessor in the distributive chain rendered the product not reasonably safe. (This is the primary, but not exclusive, test for design defective design. See Comment b.)."

[382] See Tamsen Valoir & Shubha Ghosh, *FDA preemption of drug and device labeling: Who should decide what goes on a drug label*, 21 HEALTH MATRIX (2011).

Restatement (Third) of Torts (§2(c) and comment *m*). The Supreme Judicial Court of Massachusetts held that "the defendant will not be held liable…for failure to warn…about risks that were not reasonably foreseeable at the time of sale or could not have been discovered by way of reasonable testing prior to marketing the product."[383] It is argued in this paper that there are risks in ML predictions in medicine that could be hidden but "reasonably foreseeable" and could be revealed if the algorithm provides some explanations and/or a confidence interval for its specific prediction.

d. How detailed should a ML explanation be?

As we explained above, a warning obligation provides the physician (and consequently the patient) with "*specific detailed information* on the risks [involved]."[384] The warning should be genuinely "comprehensible,"[385] "accurate and clear"[386] particularly in cases when the risk is "death or major disability."[387] As noted above, adequate warnings for ML medical applications could consist of a framework that provides, when needed, ML explanations for the ML predictions. However, we also saw that there is still little agreement on what constitutes explainable machine learning and what form interpretability should take in order to constitute an adequate warning.

Regarding warnings for conventional products, Geistfeld notes that an adequate product warning is "a warning which best promotes consumer welfare – is one that enables the ordinary consumer to make the best estimate of the product's net benefit by conveying information about product risk that is not possessed by the consumer but reasonably available to the product seller."[388] Therefore, how do we convey through interpretability the necessary information about a "product risk" that would enable the consumer to make the best estimate of the product's "net benefit"? In *Liriano* the US Court of Appeals, Second Circuit, referring to the obviousness of a risk, held that a "warning can convey at

[383] *Vassallo v. Baxter Healthcare Corp.*, N.E.2d.
[384] *Martin v. Hacker* N.E. 2d at 1312.
[385] *Pavlides v. Galveston Yacht Basin*, F. 2d at 338.
[386] *Martin v. Hacker* N.E. 2d at 1312.
[387] *Davis v. Wyeth Laboratories, Inc.*, F.2d at 129.
[388] GEISTFELD, Products Liability Law 251. 2012.

least two types of message. One states that a particular place, object or activity is dangerous. Another explains that people need not risk the danger posed by such a place, object, or activity in order to achieve the purpose for which they might have taken that risk. Thus, a highway sign that says "Danger – Steep Grade" says less than a sign that says "Steep Grade – Follow Suggested Detour to Avoid Dangerous Areas."[389] As Geistfeld explains, "a fully informed safety decision…requires knowledge of both the risk (the term *PL* in the risk utility test) and the reasonable precautions for avoiding it (described by the term *B*)"[390] In the context of ML in medicine, this could translate to, for example, having interpretable ML medical devices that provide more than one prediction and for each prediction providing different explanations in order to offer a choice. Another option, the ML device could be providing one prediction with explanations that allow the physician/patient to make the best estimate of the product's "net benefit."

As noted above, Restatement (Third) § 2 cmt *i* provides that 'warning must be provided for inherent risks" as such warnings would allow the physician/patient to avoid the "risk warned against by making an informed decision" not to purchase or use the product (in the context of ML medical device not to follow the prediction). It was argued above that the "inherent risks" in certain ML medical systems could be truly revealed through the provision of some explanations on the reasons that the ML is providing that prediction (e.g. the part of the medical image been observed).

Coming back to the issue of foreseeability, it may be wondered what types of risks are foreseeable and what should be precisely revealed through ML interpretability and/or confidence value. The Court of Appeals of Maryland in *Moran* provides some guidance in this respect. It held that "whether foreseeability is being considered from the standpoint of negligence or proximate cause, the pertinent inquiry is not whether the actual harm was of a particular kind which was expectable. Rather, *the question is whether the actual harm fell within the general field of danger which have been anticipated*."[391] Therefore,

[389] Liriano v. Hobart Corp., 170 F.3d 264, (2d Cir. 1999).
[390] GEISTFELD, Products Liability Law 265. 2012. The risk/utility test refers to a risk/benefit test. This formula encompassing the *B* and *PL* parameters will be further explained below.
[391] Moran v. Faberge Inc., 273 Md. 538, (Md. 1974).

foreseeability refers to the general risks that can be contemplated and not necessarily the particular risk. Consequently, in the context of ML, a general risk that can be contemplated, for example, in relation diagnosis by ML based on deep learning architectures, would be treated as a foreseeable risk. However, it should be pointed out that providing an overly general warning, indicating for example, that a particular drug (or a medical device) could cause side effects (or an injury in general) would not be considered adequate.[392] Therefore, the "general field of danger" is the requirement to determine foreseeability but a general vague warning concerning abstract dangers would not fulfil the requirements of an adequate warning. The Court of Appeals of Maryland referred to the "inherent and hidden danger[s]" that need to be warned of.[393] As argued "inherent and hidden danger[s]" in ML predictions could at least in certain instances be revealed through interpretable ML devices and/or ML devices that provide a confidence interval for their predictions.

Considering the above analysis, an issue that still remains concerns the level of detail that is required in the warning. In the context of ML interpretability, the issue concerns the level of detail required in the explanation given by a ML algorithm. An adequate ML explanation could be also providing an adequate warning on the risks/benefits involved in following the ML medical recommendation. In this regard, the US District Court, Northern District of Georgia held in *Jones* that "where a duty to warn arises, the duty may be breached by (1) failing to adequately communicate the warning to the ultimate user or

[392] Geistfeld, CALIF. L. REV., 1658 (2017).
[393] "[B]ased on this negligence law we think that in the products liability domain a duty to warn is imposed on a manufacturer if the item it produces has an *inherent and hidden danger* about which the *producer knows, or should know*, could be a substantial factor in bringing injury to an individual…" (*Moran v. Faberge Inc.*, Md. at 552.).

(2) failing to provide an adequate warning of the product's potential risk."[394] Therefore, the adequacy of the warning rests on both the content and the format of the disclosure.[395]

We have already seen above some of the elements concerning the content of an adequate warning. To recapitulate, a warning should encompass "*specific detailed information* on the risks [involved],"[396] should be "comprehensible,"[397] "accurate and clear,"[398] especially when exercising a "true choice" judgment involving "death or major disability."[399] Regarding the content of warnings, Geistfeld also notes that an "overly general warning can be inadequate as illustrated by an extreme example: WARNING – Product can cause injury. More detail is required, but how much?"[400]

Restatement (Third) of Torts comment *i* provides some help. It states that "[s]ubsection (c) [governing liability for defective warnings] adopts a reasonableness test for judging the adequacy of product instructions and warning. It thus parallels Subsection (b), which adopts a similar standard for judging the safety of product designs." Comment *i* also acknowledges that although the liability standard is formulated in very much identical terms for determining defects in design and in warnings, the defectiveness test is more difficult to apply in the warning context. It is therefore evident that the test for determining the adequacy of warning defects is based on the same type of risk-utility grounds as the ones applied for design defects and this is followed in most jurisdictions.[401] But the challenge in this respect concerns the factors that need to be considered in a risk-utility test and the way to strike the risk-utility balances. Restatement (Third) of Torts

---

[394] Jones v. Amazing Prods., 231 F. Supp. 2d 1228, (N.D. Ga.). Regarding the duty to warn it was held in this case that "[i]n products liability case, whether or not grounded in strict liability or negligence theory, a manufacturer's duty to warn depends on the foreseeability of the use in question, the type of danger involved and the foreseeability of the user's knowledge of the danger" (at 1247). See also *Simonetta* in which case it was held that ""[f]oreseeability does not create a duty but sets limits once a duty is established"…Once this initial determination of legal duty is made, the jury's function is to decide the foreseeable range of danger therefore limiting the scope of that duty." (Simonetta v Viad Corp., 165 Wn.2d 341, (Wash. 2008). See also Geistfeld who explains why the tort duty is justified by the frustration of the ordinary consumer's actual expectations of products safety (GEISTFELD, Products Liability Law 104-117. 2012.)
[395] GEISTFELD, Products Liability Law 273. 2012.
[396] *Martin v. Hacker* N.E. 2d at 1312.
[397] *Pavlides v. Galveston Yacht Basin,* F. 2d at 338.
[398] *Martin v. Hacker* N.E. 2d at 1312.
[399] *Davis v. Wyeth Laboratories, Inc.,* F.2d at 129.
[400] GEISTFELD, Products Liability Law 277. 2012.
[401] Id. at, 283.

comment *i* acknowledges this challenge but also provides some assistant to this quest. It provides that:

"[i]t is impossible to identify anything approaching a perfect level of detail that should be communicated in product disclosures. For example, educated or experienced product users and consumers may benefit from inclusion of more information about the full spectrum of product risks, whereas less-educated or unskilled users may benefit from more concise warnings and instructions stressing only the most crucial risks and safe-handling practices…In some cases, excessive detail may detract from the ability of typical users and consumers to focus in the important aspects of the warnings, whereas in others reasonably full disclosure will be necessary to enable informed, efficient choices by product users…No easy guideline exists for courts to adopt in assessing the adequacy of product warnings and instructions. In making their assessments, courts must focus on various factors, such as comprehensibility, intensity of expression, and the characteristics of expected user groups."

It therefore appears that more information is not necessarily always better. In this respect the US Court of Appeals, fourth Circuit in *Hood* held that "…the price of more detailed warnings is greater than their additional printing fees alone. Some commentators have observed that the proliferation of label detail threatens to undermine the effectiveness of warnings altogether." [402] In the context of ML interpretability, this means that a balance would need to be struck between on the one hand a more detail ML explanation and on the other ensuring the effectiveness of this explanation (warning). This balance would depend on a number of factors that include, the medical context in which the ML medical device is used; the seriousness of the possible injury or possibility of death; the urgency in taking a medical decision for example if the decision is taken in an ICU or an Accident and Emergency department (A&E); whether there is a clear medical target to be achieved; and the confusion that might be created by providing excessive complex explanations and the loss of valuable time in assessing the ML explanations provided for each prediction. As noted above, the different forms of ML interpretability and what would constitute

[402] Hood v. Ryobi Am. Corp., 181 F.3d 608, (4th Cir.).

adequate level of ML interpretability in different medical contexts is the subject of further research that will complement this paper.

e. Physicians acting as learned intermediaries

An issue that was already touched upon above concerns the impact of physicians, acting as intermediaries, on warnings. There are two issues arising in this regard. First, whether warnings directed to patients will be an effective means of communication or whether suppliers of ML devices should also provide warnings to physicians (intermediaries) who will be in the best place to pass this information to the patient.[403] Secondly, whether the supplier escapes liability if she solely gives all the information necessary to the person through who the product is supplied (e.g. the physician).

Regarding the first issue, Restatement (Third) §6(d) provides that a "prescription drug or medical device is not reasonably safe due to inadequate instructions or warnings if reasonable instructions or warnings regarding foreseeable risks of harm are not given to:…(1) prescribing and other health-care providers who are in a position to reduce the risks of harm in accordance with instructions or warnings…" Therefore, applying this reasoning to ML medical devices, warnings might need to be given to the physician. The question explored in this paper is the form that this warning should take.

Regarding the second issue, in the context of prescription drugs and conventional medical devices, a manufacturer of a drug or medical device, based on the *learned intermediary doctrine,* fulfills its obligations to warn by disclosing the appropriate (adequate) information to the prescribing physician who in turn transfers this information to the patient when prescribing the drug or medical device.[404] However, the question remains,

[403] This point regarding the role of intermediaries in general, which can be also applicable to the ML medical context, was raised by GEISTFELD, Products Liability Law 241. 2012.

[404] Id. at, 250. See also Supreme Court of California holding that "[t]he manufacturer cannot be held liable if it has provided appropriate warnings and the doctor fails in his duty to transmit these warnings to the patient or if the patient relies on inaccurate information from others regarding side effects of the drug" (Brown v. Superior Court 751 P.2d 470, (Cal. 1988).). However, changes in the delivery of health care that resulted from direct marketing and managed care could have impacts on the "learned intermediary" doctrine hence a warning might be necessary for both the prescribing physician and the patient (see

what should constitute adequate information? Manufacturers of ML medical devices might be alleging that the physician, as a learned intermediary, should have known of certain risks associated with ML medical applications. In the non-medical context, the sophisticated *user* defense, which in the medical context is the learned intermediary rule, is treated as an exception to the manufacturer's general duty to warn consumers, and therefore, if successfully argued, it will constitute an affirmative defense that negates the manufacturer's duty to warn.[405] Under the sophisticated user defense (the learned intermediary rule in medicine), sophisticated users need not be warned about dangers of which they are "already aware or should be aware."[406] In other words, as the Court explains, if the manufacturer reasonably believes that the "user [the physician in the medical context] will know or should know" about a given "product's risk" the manufacturer need not warn that user of that risk.[407] The "should know" standard might be raising concerns. Different physicians might have different knowledge of the risks involved in, for example, ML diagnosis. This variability of knowledge is also relevant when discussing the "obvious danger rule" recognized by Californian law.[408] In the context of the "obvious danger rule," Geistfeld explains that an objective test is applied and the courts do not investigate the user's subjective knowledge.[409] Therefore, applying an objective test also to the learned intermediary rule would lead to the conclusion that the physician should not be treated as a particularly knowledgeable of ML medical devices who needs no or even minimal warnings.

Furthermore, Restatement (Second) of Torts section 388 cmt. *k* states that:

Centocor, Inc. v. Hamilton, 310 S.W.3d 476, (Tex. App. Ct. 2010). The same reasoning could be also applicable for ML medical devices especially those that would be also used directly by the consumer outside the clinic and not in the physician's presence.

[405] Johnson v. American Standard, Inc., 43 Cal.4th 56, (Cal. 2008).

[406] Id. The Court explained that sophisticated user defense evolved out of the Restatement (Second) of Torts section 388 and the obvious danger rule, an accepted principle and defense in California (Stevens v. Parke, Davis & Co., 9 Cal.3d 51, (Cal. 1973).

[407] *Johnson v. American Standard, Inc.,* Cal.4th at 66.; this was the interpretation given by courts to Restatement (Second) of Torts section 388, subdivision (b); see also Martinez v. Dixie Carriers, Inc., 529 F.2d 457, (5th Cir. 1976).

[408] The obvious danger rule provides that there is no need to warn of known risks under either negligence or strict liability; see *Johnson v. American Standard, Inc.,* Cal.4th at 67.

[409] GEISTFELD, Products Liability Law 233. 2012.

"a condition, although readily observable, may be one which only persons of special experience would realize to be dangerous. In such case, if the supplier, having such special experience, knows that the condition involves danger and has no reason to believe that those who use it will have such special experience as will enable them to perceive the danger, he is required to inform them of the risk of which himself knows and which he has no reason to suppose that they will realize."

In the context of ML medical devices, comment *k* would mean that ML manufacturers, could be "persons of special experience," and hence they could be aware of risks inherent in ML algorithms that are not obvious or even comprehensible to physicians (and patients). Moreover, there may be hidden risks that are not even apparent to the ML engineers but which could be revealed if the ML algorithm is interpretable. Considering that ML in medicine is at its inception and not well understood by physicians, it would be hard to invoke the learned intermediary rule (or obvious danger rule)[410] if no adequate warning is provided. This is another reason why identifying what constitutes an adequate warning for ML applications in medicine is of paramount importance.[411] Once an adequate warning is developed for ML, then like with prescription drugs and conventional medical devices, a manufacturer of a ML medical device would fulfil its obligations to warn by disclosing the information to the physician who in turn transfers this information to the patient to obtain her informed consent. As argued above, a warning for ML medical devices, in at least certain cases, might need to take the form of ML explanations and/or the provision of a confidence interval. Consequently, it appears that, additionally to what was discussed above, providing the physician with an explainable ML device that gives the necessary explanations for the proposed ML prediction and/or the ML algorithm providing a confidence interval for its specific perdition would in the great majority of cases fulfil the manufacturers' obligations regarding warnings.

[410] Accepted principle in defense in California; See *Stevens v. Parke, Davis & Co.,* Cal.3d. and *Johnson v. American Standard, Inc.,* Cal.4th at 65.

[411] See for example, how ML systems could be fooled by surgical skin scars in providing skin melanoma diagnosis at Julia Winkler, et al., *Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition* JAMA Dermatology *at* https://jamanetwork.com/journals/jamadermatology/article-abstract/2740808.

f. Post-sale duty to warn

Another interesting issue that could possibly arise with the deployment of ML devices in medicine concerns the post-sale duty to warn. Restatement (Third) of Torts §10(b) states that a "reasonable person in the seller's position would provide a warning after the time of sale if:... (1) *the seller knows or reasonably should know* that the product poses a substantial risk of harm to persons or property; and (2) those to whom a warning might be provided can be *identified* and can reasonably be assumed to be unaware of the risk of harm; and (3) a *warnings can be effectively communicated* to and acted on by those to whom a waring might be provided; and (4) the risk of harm is sufficiently great to justify the burden of proving a warning" (emphasis added). Concerning the first condition that the "seller knows or reasonably should know" raises the question concerning the extent to which the seller of ML medical devices should be constantly monitoring the product. In the case where the seller "may have known or should have known" of the risk *at the time of sale* then the failure to warn would be a waring defect under §2(c).[412] However, knowledge of a risk may arise *after the time of sale* and this may give a rise to warn at that time. Regarding conventional products Restatement (Third) of Torts §10 comment *c* provides that the "burden of constantly monitoring product performance in the field is usually too burdensome to support a post-sale duty to warn." However, comment *c* adds that when reasonable grounds exist for the seller to suspect that a hitherto unknown risk exists the duty of reasonable care may require investigation. In the context prescription drugs and devices, that can be also applicable to ML medical devices, comment *c* points out that courts traditionally impose a "continuing duty of reasonable care to test and monitor" after sale to discover product related risks. Therefore, in the context of ML medical devices this could include the obligation, for example, to re-test the generalization ability of the algorithm if new data become available. Interpretable ML algorithms could also ensure that part of the obligation concerning post-sale duty to warn is satisfied. It would be satisfied, as the ML device would be continuously providing post-sale explanations (post-sale warnings) for its predictions and this might also reveal risks that were not understood (unforeseeable) at the time of sale but become apparent post-

[412] See Restatement (Third) of Torts §10 comment *c*.

sale. This could therefore also satisfy to some extent the obligation of "continuing duty of reasonable care to test and monitor." As far as, conditions (2) and (3) of §10(b) are concerned, do not seem either to pose much of a problem for ML medical devices. Users of ML medical devices would be relatively easily identifiable and warned accordingly.

g. The correlation between warnings and defective design

Identifying an adequate warning for ML medical devices would also in certain cases be relevant when examining ML design defects and malfunctioning. The correlation between warnings and ML malfunctioning is dealt with in the section that follows. This section deals with the correlation between warnings and defective designs that could be also of particular interest to ML medical devices.

Under comment *k* "unavoidably unsafe products" are exempted from strict liability.[413] Comment *k* states that that the seller of "unavoidably unsafe products" that are "properly prepared and marketed and proper warning is given" is "not to be held to strict liability for unfortunate consequences…merely because he has undertaken to supply the public with an apparently useful and desirable product, attended with a known but apparently reasonable risk." Comment k has been adopted in the vast majority of jurisdictions.[414] Drugs and vaccines are expressly referred to in comment *k* as examples of "unavoidably unsafe products" and they could hence enjoy the protection provided by this exception. The policy considerations behind comment *k* for prescription drugs is explained by the Supreme Court of California in *Brown*.[415] The Supreme Court starts its explanation by going back to the time when Restatement (Second) §402A was contemplated.[416] As it points out, "[d]uring a rather confusing discussion of a draft of what was to become section 402A, a member of the institute proposed that drugs should be exempted from strict liability on the ground that it would be 'against the public interest' to apply the doctrine to such products because of the 'very serious tendency to stifle medical research

[413] There is no provision similar to comment *k* in Restatement (Third); instead, it encompasses a special rule governing defective design of medical products (GEISTFELD, Principles of Products Liability 171. 2011.
[414] *Brown v. Superior Court* P.2d at 476.
[415] Id.
[416] See 38 A.L.I. Proc. 19, 90-92, 98 (1961).

and testing.'"[417] It also explained that "there is an important distinction between prescription drugs and other products such as construction machinery…the products of which were held strictly liable. In the later cases, the product is used to make work easier or to provide pleasure, while in the former it may be necessary to alleviate pain and suffering or to sustain life."[418] Finally, it explained that "[i]f drug manufacturers were subject to strict liability, they might be reluctant to undertake research programs to develop some pharmaceuticals that would prove beneficial or to distribute others that are available to be marketed, because of the fear of large adverse monetary judgments."[419] It appears that the policy considerations applicable for exempting drugs from strict liability would be also applicable to ML medical devices. ML medical devices, like drugs and vaccines, supply the public with an apparently useful and desirable product that promotes health and safety.[420] In particular, ML research in medicine already appears promising in providing unprecedent benefits to healthcare. ML in medicine could revolutionize healthcare hence why circumspection is needed in designing a legal framework for ML medical systems.[421] Similar to drug and vaccine manufacturers, a legal liability framework that subjects manufacturers of ML medical devices to excessive adverse litigation could undermine or even destroy this ML promising revolution in healthcare. There could be such substantial detrimental effects on the ML industry, as a ML medical algorithm might be used simultaneously on thousands or millions of patients. Consequently, class-action lawsuits in cases of injuries could easily drive ML industry into destruction.[422] At the same time, the potential impact of a ML medical algorithm on possibly thousands or millions of patients also calls for circumspection in deploying ML algorithms in medicine. Such deployment should be subject to a legal framework that also ensures the health and safety of patients.

[417] *Brown v. Superior Court* P.2d at 475.
[418] Id. at, 478.
[419] Id. at, 479.
[420] *Mutatis mutandis* see also Geistfeld, CALIF. L. REV., 1670 (2017).
[421] As Geistfeld mentioned about drug manufactures, if they were subject to strict liability, they might be reluctant to undertake research in developing certain types of pharmaceuticals or distribute others that are ready to be used, because of the fear of large monetary judgments (GEISTFELD, Products Liability Law 321. 2012.).
[422] See discussion in relation to contaminated blood by Geistfeld, CALIF. L. REV., 1670-1671 (2017).

Therefore, if there are also strong policy reasons to apply comment *k* on ML medical devices, how would it affect allegations concerning design defects of ML medical devices? The Supreme Court of Nebraska in *Freeman* provides a clear explanation of how comment *k* classically applies in design defect cases. It holds that "[t]he majority of jurisdictions that have adopted comment *k* apply it on a case-by-case basis…Although a variety of tests are employed among jurisdictions that apply comment *k* on a case-by-case basis, the majority apply the comment as an affirmative defense, with the trend toward the use of a risk-utility test in order to determine whether the defense applies…When a risk utility test is applied, the existence of a reasonable alternative design is generally the central factor…Because the application of comment *k* is traditionally viewed as an exception and a defense to strict liability, courts generally place the initial burden of proving the various risk utility factors on the defendant…Thus, under these cases, the plaintiff's burden of proof for his or her prima facie case remains the same as it is in any products liability case in the given jurisdiction."

The Restatement (Third) takes a different approach from Restatement (Second) on this matter. Restatement (Third) §6(c) provides that a 'prescription drug or medical device is not reasonably safe due to defective design if the foreseeable risks of harm posed by the drug or medical device are sufficiently great in relation to foreseeable therapeutic benefits that *reasonable health-care providers*, knowing of such foreseeable risks and therapeutic benefits, would not prescribe the drug or medical device for *any class of patients*" (emphasis added). The liability rule in §6(c) has been heavily criticized.[423] Geistfeld provides an interesting analysis in response to this criticism that is also useful for addressing some of the legal issues arising out of the deployment ML algorithms in the

[423] See James A. Henderson Jr. & Aaron D. Twerski, *Drug Designs Are Different* 111 YALE L.J. 151(2001). as noted by GEISTFELD, Principles of Products Liability 173. 2011. See also Supreme Court of Nebraska in *Freeman* holding that "There are several criticisms of *§ 6(c)*, which will be briefly summarized. First, it does not accurately restate the law. It has been repeatedly stated that there is no support in the case law for the application of a *reasonable physician standard* in which strict liability for a design defect will apply only when a product is not useful for any class of persons…Fourth, the test allows a consumer's claim to be defeated simply by a statement from the defense's expert witness that the drug at issue had *some benefit for any single class of people*. Thus, it is argued that application of *§ 6(c)* will likely shield pharmaceutical companies from a wide variety of suits…" (emphasis added) (Freeman v. Hoffman-La Roche, Inc., 618 N.W.2d 827, (Neb. 2000).).

medical domain.[424] However, before proceeding with Geistfeld's analysis on §6(c), it is useful to explain the correlation between warnings and product design as it will assist in the understanding of the analysis on §6(c).

Regarding the correlation between warnings and product design, the challenge concerns the situations where a warning can substitute for a design change.[425] In other words, as Geistfeld asks, could a warning that states: "WARNING: No airbags in car!" eliminate the need to incorporate airbags into the design of a vehicle as required by the risk-utility test?[426] Similar questions, and even more complex ones, could also arise in the context of ML medical devices. The challenges with regard to ML would be greater as there could be little understanding of the implications of providing information on the technical characteristics of ML algorithms. For example, providing information on type of data that were used, the ML architecture, or the type of features that were used might not clearly indicate the impact, if any, on design defects. Considering Geistfeld's example on the lack of an airbag warning, Restatement (Third) §3 comment d (under sub-title illustrations) states that the "fact that danger is open and obvious does not bar the design claim." In other words, an "open and obvious" danger does not necessarily shield the defendant from liability based on defective design.[427] Extending this reasoning concerning "open and obvious" dangers, it might be logically concluded that a product warning (such as "No airbags") does not either shield the manufacturer from liability concerning a design defect. Designing defects would still need to be subject to the risk-utility test.[428] This conclusion seems logical but at first sight it might be challenged by Restatement (Second) of Torts comment *j*.

[424] GEISTFELD, Principles of Products Liability 174-180. 2011.
[425] Id. at, 164.
[426] Id.
[427] This is accepted in the "strong majority" of jurisdictions (see more detail by Geistfeld id.).
[428] Latin, UCLA L. REV., (1994). as noted by GEISTFELD, Principles of Products Liability 164. 2011. Restatement (Third) §3(b) adopts a reasonableness ("risk-utility balancing") test as the standard for judging the defectiveness of product designs. As Restatement (Third) comment *d* further explains the "test is whether a reasonable alternative design would, at reasonable costs, have reduced the foreseeable risks of harm posed by the product and, if so, whether the omission of the alternative design by the seller or a predecessor in the distributive chain rendered the product not reasonably safe. (This is the primary, but not exclusive, test for design defective design. See Comment b.)."

Comment *j* provides that "[w]here a *warning is given*, the seller may reasonably assume that it will be *read and heeded*; and a product bearing such a warning, which is safe for use if it is followed, is *not in defective condition*, nor is it unreasonably dangerous" (emphasis added). Does comment *j* allow the manufacturer to solely provide a warning instead of a redesigning the product? Geistfeld explains that comment *j* is a source of duty to warn under section 402A.[429] Consequently, comment *j* creates the duty to warn within the strict products liability framework. In other words, the provisions on warnings in comment *j* do not foreclose allegations of a defective design.[430] This approach was followed by the majority of courts hence also reflected in Restatement (Third) §2 comment *l* providing that "[w]arnings are not…a substitute for the provision of a reasonably safe design."

Having provided the above analysis, let us apply it on hypothetical example concerning a warning on the predictions of a ML medical device. Let us assume, that these warnings provide technical details on the data used, labels used and targets set during the training of the algorithm. Let us also assume that the manufacturer correctly argues that this information indicates risks that could result from ML medical predictions. However, suppose, the physician/patient (plaintiff) alleges that it is difficult to follow such a warning. They claim that in order to follow such a warning they would need to consult expert ML engineers and together with other physicians study the risks for each medical prediction that ML makes. As a result, they argue that instead of such a difficult to follow warning, the manufacturer could have eliminated the risks by a reasonable alternative design. In this context, Geistfeld notes, generally, when there is difficulty in following a warning, the cost of redesign (denoted as "$B_{redesign}$") could be less than the consumers' cost of complying with the warning (denoted as "$B_{complying\ with\ warning}$"). Therefore, $B_{redesign} < B_{complying\ with\ warning}$.[431] Under these circumstances, he notes that, the "risk-utility characteristics of the warning do not prevent the design from being defective" (i.e. $B_{redesign}$

[429] GEISTFELD, Principles of Products Liability 165. 2011. Comment *j* provides "[i]n order to prevent the product from being unreasonably dangerous, the seller may be required to give directions or warning, on the container as to its use."
[430] The different forms of liability are also evident from Restatement (Second) §402A cmt. *a* stating that "[t]he rule stated here is not exclusive and does not preclude liability based upon alternative ground of negligence of the seller, where such negligence can be proved."
[431] GEISTFELD, Principles of Products Liability 167. 2011.

< $B_{complying\ with\ warning}$ < PL) (P refers to the probability of a risk materializing and L refers to the cost of injury or loss). So, for our ML hypothetical example, it would mean that the ML manufacturer might need in certain cases to come up with another design that imposes an easier warning to be followed. In other words, the ML medical device is safe for use when the user reads and heeds the product warning but it is still defectively designed.[432] It is at this point where a warning in the form of a ML explanation and/or confidence interval for its specific prediction might provide a shield to such an allegation. An explainable ML, that provides warnings in the form of explanations that are easy to follow, could be also shielding the manufacturer from allegations that the ML medical device has been defectively designed. In other words, an interpretable ML might prevent allegations that the manufacturer could have eliminated the risks by adopting a reasonable alternative design. Therefore, explainable ML medical devices might be fulfilling the requirements of both Restatement (Second) of Torts comment *j* and the risk-utility test in Restatement (Third). In contrast, a conventional warning (i.e. not encompassing interpretability) might leave a door open for the allegation that the ML system has been defectively designed.

Coming back to Restatement (Third) §6(c) that we started considering above, to recapitulate, it refers to prescription drugs and medical devices and states that they are "not reasonably safe due to *defective design*…if *reasonable health-care providers*, knowing of such foreseeable risks and therapeutic benefits, would not prescribe the drug or medical device for *any class of patients*" (emphasis added). As we noted, this rule has been heavily criticized. Geistfeld's analysis contains useful elements which could be also applied to the deployment of ML medical systems. Geistfeld first explains, that a physician who prescribes treatment that is not in the patient's best interest is subject to malpractice liability.[433] Additionally, failing to obtain the patients informed consent is another source of malpractice liability. Therefore, in the medical context, it is the physician that makes the initial risk-utility decision for prescription drugs and medical devices and is then obliged to provide explanations to the patient for that decision in order to obtain the

[432] See in the general context, the reasoning behind this point by Geistfeld at id.
[433] Id. at, 174.

informed consent of the patient.[434] This also the reason why under the learned intermediary rule, the manufacturers of prescription drugs and medical devices satisfy their duty to warn if they provide adequate warnings to the physicians.[435] Restatement (Third) §6(c) provides in effect a risk-utility test to determine whether a prescription drug or medical device is defective.[436] It refers to the "foreseeable risks of harm posed by the drug or medical device" in relation to "foreseeable "therapeutic benefits." It is therefore a risk-utility test albeit, as Geistfeld notes, a "risk-utility test that is modified to account for the manner in which the physician-patient relationship affects the nature of the product transaction."[437] However, Restatement (Third) §6(c) is still criticized. The criticism concerns the claim that a drug would not be considered as defectively designed under Restatement (Third) if it is beneficial for even a small class of users.[438] In this regard, Geistfeld argues that closer analysis shows otherwise.[439] He notes that, if it is unreasonably dangerous to prescribe a drug to a class of users then the manufacturer should warn physicians accordingly. He points out that the fact that a drug should not be prescribed to a class of users does not justify a finding that the drug is defectively designed. However, if the manufacturer does not provide such warnings then she would be subject to liability for the injuries caused by the inadequate warning.[440] In other words, Restatement (Third) §6(c) does not treat the above drug as defectively designed but could consider it as providing inadequate warning. However, such a drug can still be found as defectively designed under Restatement (Third).[441] Therefore, the majority rule determines the alleged defectiveness of prescription drugs and medical devices on a case-by-case basis. Whereas, Restatement (Third) looks at the relevant class of patient to evaluate the design. Under both approaches, Geistfeld argues that it would be hard to find

---

[434] Id.
[435] Id.
[436] Id. at, 175.
[437] Id.
[438] Restatement (Third) §6(c) states that "reasonable health-care providers, knowing of such foreseeable risks and therapeutic benefits, would not prescribe the drug or medical device for *any class of patients*" (emphasis added).
[439] GEISTFELD, Principles of Products Liability 177-178. 2011.
[440] Id. at, 178.
[441] Id.In this regard see, Restatement (Third) §2(b) provides that a product "is defective in design when the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission of the alternative design renders the product not reasonably safe."

liability for defective design, as a drug or medical device would be beneficial to at least one class of patients.[442] However, as he also points out "to make a proper prescription, the physician must have received the requisite risk-utility information from the manufacturer. 'With drugs therefore, the liability game is with the warning candle, not with design.'[443]"[444]

It is clear from the above, that under either approach, majority or Restatement (Third), warnings do play a crucial role in liability challenges concerning drugs and medical devices. Similar type of challenges would also surface regarding ML medical systems. In the same way it could be argued that ML medical devices would still be beneficial to at least one class of patients. As a result, under both the majority approach and Restatement (Third) such ML limitations would not justify a finding that the ML medical device is defectively designed. Therefore, the question will amount to whether an adequate warning was provided in this regard. As argued in this paper, identifying warnings fit for ML algorithms is a challenging task. For example, providing the physicians with warnings/information encompassing technical details concerning the algorithm or its training might not be of much use to the physician. Consequently, the physician would not be enabled to make the required initial risk-utility decision and provide the appropriate explanations to the patient for that decision in order to obtain the informed consent of the patient. As we saw in the analysis of Geistfeld above, Restatement (Third) §6(c) provides in effect a risk-utility test to determine whether a prescription drug or medical device is defective. It refers to the "foreseeable risks of harm posed by the drug or medical device" in relation to "foreseeable "therapeutic benefits." Therefore, if the ML warning does not in practice provide an appropriate risk-utility test to the physician it would render the ML medical device defective. It appears that also in this respect, a ML medical that can provide explanations/confidence intervals for the predictions made could fulfill the requirements of a risk-utility test even under Restatement (Third) §6(c).

[442] Id. at, 180.
[443] Michael D. Green, *Prescription Drugs, Alternative Designs, and the Restatement (Third): Preliminary Reflections* 30 SETON HALL L. REV. 207, 208-209 (1999).
[444] GEISTFELD, Principles of Products Liability 180. 2011.

h. The correlation between warnings and malfunctioning

As we noted above, identifying an adequate warning for ML medical devices would also in certain cases shield the manufacturer from allegations concerning ML malfunctioning. Restatement (Third) of Torts § 3 comment b provides explanations on what constitutes malfunction. It states that:

"…when the incident…is one that ordinarily occurs as a result of product defect, and evidence in the particular case establishes that the harm was not solely the result of causes other than product defect existing at time of sale, it should not be necessary for the plaintiff to incur the cost of proving whether the failure resulted from a manufacturing defect or from a defect in the design of the product…the inference [could be drawn] that the product was defective whether due to a manufacturing defect or design defect. Under those circumstances, the plaintiff need not specify the type of defect responsible for the product malfunction."[445]

In other words, the malfunction is sufficient proof of defect.[446] As indicated in Restatement (Third) of Torts the malfunctioning mainly results from manufacturing defects. From this perspective manufacturing defects "cause the product to fail to perform their manifestly indented functions."[447] Thus, the malfunction doctrine is limited to "situations in which a product fails to perform its manifestly intended function."[448] Consequently, in the case where a ML medical device delivers a misdiagnosis the allegation could be that the product did not perform its "manifestly intended function" constituting a product malfunction that subjects the manufacturer to strict products

[445] RESTATEMENT (THIRD) OF THE TORTS: PRODCUTS LIABILITY § 3 cmt. b. 1998.
[446] Geistfeld, CALIF. L. REV., 1635 (2017). In this respect see also Denny v. Ford Motor Co., 662 N.E. 2d 730, (N.Y. 1995). In this case was held that "[t]he cause of action is one involving true "strict" liability, since recovery may be had upon a showing that the product was not minimally safe for its expected purpose--without regard to the feasibility of alternative designs or the manufacturer's "reasonableness" in marketing it in that unsafe condition."
[447] RESTATEMENT (THIRD) OF THE TORTS: PRODCUTS LIABILITY § 3 cmt. b. 1998.
[448] Id; see also Geistfeld, CALIF. L. REV., 1634 (2017).

liability.[449] However, Geistfeld[450] notes, such conclusions are debatable as the rule adopted by Restatement (Third) of Torts "is not ideal, which reflects the difficulty of formulating a concise, general statement of the principle."[451]

Be that as it may, Geistfeld points out that even if the malfunction doctrine were more rigorously defined, manufacturers would still be subjected to significant uncertainty for a different reason.[452] A considerable majority of states instead of defining malfunction on the basis of the product's "manifestly intended function" they evaluate this issue with the "consumer expectation test."[453] Particularly, in the context of artificial intelligence, the consumer does not know what to expect from AI medical devices hence why the manufacturer should "adequately warn about the associated risks."[454] Geistfeld argues in the context of autonomous vehicles, which could *mutatis mutandis* be also applicable for ML medical devices, once satisfying the obligation to adequately warn about the foreseeable risk of crash that is unavoidable or inherent in a safely designed autonomous vehicle, the manufacturer will also escape liability for such crashes under the malfunction doctrine.[455] The idea behind this reasoning is that an adequate warning about the inherent risks of crash (revealing the true character of the product) cannot frustrate the consumer's expectation in the event that the risk materializes thereby excluding liability under the consumer expectation test and consequently under the malfunction doctrine.[456] Therefore, in the context of ML medical devices, one may argue that providing an adequate warning in the form of appropriate explanations and/or a confidence interval for the prediction made, would avoid frustrating the consumer expectation in the case where an associate medical risk materializes. Consequently, such a warning could in principle also exclude liability under the malfunction doctrine.

---

[449] Point raised in the context of autonomous vehicles by Geistfeld, CALIF. L. REV., 1637 (2017).; this point could be also applicable to ML medical devices.
[450] Id.
[451] David G Owen, *Manufacturing Defects*, 53 SCL REV. 851, 883 (2001).
[452] Geistfeld, CALIF. L. REV., 1637 (2017).
[453] Id. at, 1638.
[454] Id. in the context of autonomous vehicles.
[455] Satisfying the duty to warn does not necessarily satisfy the duty for a non-defective design (see id. at, 1639.).
[456] Id. at, 1639-1640.

However, particularly, when it comes to products that have a primary purpose of *promoting health or safety* (e.g. donated blood or blood products), the sheer fact of injury could be adequate for establishing a *malfunction* subject to strict liability.[457] The argument would be that such blood products or ML medical devices are not marketable with their true character known hence making it defective.[458] Consequently, manufacturers of ML medical devices, like sellers of contaminated blood, could be vulnerable to strict products liability under Restatement (Second) for the inherent risks in ML medical predictions or contaminants in blood accordingly. It is at this point where Restatement (Second) of Torts § 402A cmt. *k* could provide a solution to such cases including malfunctioning ML medical devices. [459]

We have seen the scope and application of comment k when we discussed design defects. To briefly recapitulate, under comment *k* "unavoidably unsafe products" are exempted from strict liability.[460] Comment *k* states that that the seller of "unavoidably unsafe products" that are "properly prepared and marketed and proper warning is given" is "not to be held to strict liability for unfortunate consequences…merely because he has undertaken to supply the public with an apparently useful and desirable product, attended with a known but apparently reasonable risk." Drugs and vaccines are expressly referred to in comment *k* as examples of "unavoidably unsafe products" that could enjoy the protection provided by this exception.

In this regard it should not be forgotten that the manufacturers of ML medical devices would still be subject to strict liability with regard to injuries arising out of products that

---

[457] GEISTFELD, Products Liability Law 339. 2012. See for example in relation to the sale blood products where certain blood-borne diseases cannot be detected at the time of sale and other risks of contaminated blood cannot always be reduced to more ordinary levels (Geistfeld, CALIF. L. REV., 1671 (2017).). See also ruling of the Court of Appeal of California, First Appellate District Division One in *Grinnell* holding that "…it is clearly the law in California that the theory of strict liability in tort is available in cases where the vaccinated individual contracts the disease the vaccine was designed to protect against" (Grinnel v. Charles Pfizer & Co., 274 Cal. App. 2d 424, (Cal. Ct. App.).); see also Restatement (THIRD) of the Torts § 3 cmt. b providing that "manufacturing defects cause the products to fail to perform their manifestly intended functions."

[458] In relation to marketing of blood Geistfeld, CALIF. L. REV., 1671 (2017).

[459] See in this regard *supra* Geistfeld's analysis why the exemption from the rule of strict liability provided by comment *k* presumably concerns *malfunctions* of "unavoidably unsafe products" at id. at, 1670. and GEISTFELD, Products Liability Law 332-340. 2012.

[460] There is no provision similar to comment *k* in Restatement (Third); instead, it encompasses a special rule governing defective design of medical products (GEISTFELD, Principles of Products Liability 171. 2011.

were not "properly prepared" or no "proper warning" was given as in these cases comment *k* is not applicable.[461] In other words, in order to obtain the benefit of comment *k* the defect should be corelating to a risk that cannot be sufficiently reduced by the exercise of reasonable care.[462] These issues will maintain complexity as the allegation could be that comment k should not be applicable as the associated risk could have been sufficiently reduced by the exercise of reasonable care.[463] For example, it may be argued that no reasonable care was exercised due the lack of adequate ML warnings and hence comment *k* should not be applicable to the allegation of ML malfunction. Therefore, if warnings in the form of ML explanations for the predictions made in certain instances are considered as adequate warnings, then such warning would also allow for the application of comment *k*. Therefore, explainable ML/ML providing confidence intervals might be also shielding the manufacturer from allegations on ML malfunctioning.[464] This is so, as comment k requires the product to be accompanied by an "adequate warning."[465]

12. Conclusion

We have seen at the outset of this paper that sometimes, knowledge is incomplete or changing and this requires reasoning methods that can deal with uncertainty.[466] Machine learning could be of particular use to these kinds of problems for which encoding an explicit logic of decision-making performs very poorly.[467]

ML research in medicine already appears promising in providing unprecedented benefits to healthcare. ML medical systems would not only be applicable in the clinic but could soon be also directly used by patients outside the clinic without physician's presence or

[461] For example, in relation to blood products, in cases where there was inadequate sterile environment that contaminated the blood (Geistfeld, CALIF. L. REV., 1673 (2017).).
[462] For example, in the case of blood, HIV was undetectable when it initially contaminated the blood supply (see more detail explanations on these at id.).
[463] See similar question raised in relation to the hacking of autonomous vehicles by Geistfeld (id.).
[464] Where comment *k* is applicable, namely in relation to "unavoidably unsafe" products, manufacturers would not be subject to strict liability but would still be subject to ordinary negligence liability for these malfunctions (Geistfeld id.).
[465] GEISTFELD, Principles of Products Liability 186. 2011.
[466] FINLAY & DIX, 34-43. 1996.
[467] Burrell, BIG DATA & SOCIETY, 6 (2016).

examination. Considering that ML could provide substantial benefits to healthcare, circumspection is needed in choosing a legal framework governing ML in medicine. Similar to drug and vaccine manufacturers, a legal liability framework that subjects manufacturers of ML medical devices to excessive adverse litigation could undermine this ML promising potential in healthcare. At the same time, the potential impact of a ML medical algorithms on thousands or millions of patients also calls for circumspection in deploying ML algorithms in medicine. Such deployment should be subject to a legal framework that also ensures the health and safety of patients.

In the context of drugs and conventional (not based on ML) medical devices, Geistfeld argued that it would be hard to find liability for defective design, as a drug or medical device would be beneficial to at least one class of patients.[468] However, as he also points out "to make a proper prescription, the physician must have received the requisite risk-utility information from the manufacturer. 'With drugs therefore, the liability game is with the warning candle, not with design.'[469]"[470]

Therefore, the crux of matter concerns the type of information that a manufacturer should be providing to the physician considering that the physician would be acting as a learned intermediary and considering that the manufacturer would be held to the standard of an expert in the field.[471] The manufacturer should be providing information (warning) concerning the ML medical prediction to the physician in a manner appropriate for her expertise. After the physician is given an adequate warning appropriate for her expertise, she should then "translate" this information for the patient in her verbal explanation in order to obtain the patient's informed consent.[472]

As we have seen, there are distinct challenges in developing and deploying ML systems in healthcare. Furthermore, there are tradeoffs inherent in medical decisions that need to be explained by the physician to the patient in order to obtain the patient's informed

[468] GEISTFELD, Principles of Products Liability 180. 2011.
[469] Green, SETON HALL L. REV., 208-209 (1999).
[470] GEISTFELD, Principles of Products Liability 180. 2011.
[471] Point raised by Mark Geistfeld (NYU Law School) in a discussion we had on this subject
[472] Id..

consent. Moreover, patients' expectations of treatment outcomes might differ and different patient might be willing to take different risks.

Therefore, identifying what constitutes appropriate warning for ML medical systems would create a constructive relationship between ML medical system (ML manufacturers), physicians and patients that would provide better healthcare and encourage the speedier development and deployment of machine learning in medicine. Additionally, developing this constructive relationship would shield the manufacturers of ML algorithms from uncertainty concerning their legal obligations that could be stifling machine learning innovation.[473] At the same time it would provide better protection to the patient.

It was argued that explainable ML (i.e. ML delivering certain types of explanations for its specific prediction) and/or delivering a confidence interval[474] (i.e. ML indicating its certainty for a specific prediction), in at least certain instances, should constitute elements in warnings. This paper sets the foundations for this correlation between warnings, explainable ML and ML confidence intervals. The relationship between explainable ML systems/ML systems providing confidence intervals and warnings would be more extensively developed in further research that will complement this project. This correlation is based on the distinct manner ML systems learn and generalize. Explanations and confidence intervals for the specific predictions would allow the physician to consider the ML prediction, ML explanation, ML confidence interval and combine them with her medical judgment in order to assess the risks inherent in this ML

[473] The European Commission stated in its Communication on Artificial intelligence for Europe that in order to fully benefit from the opportunities presented by these emerging new technologies a clear and stable legal framework will stimulate investment and, in combination with research and innovation, will help bring the benefits of these technologies to business and citizens. It is also noted that it is necessary to examine whether the current rules at EU and national level for safety and liability are appropriate and whether for manufacturers and service providers the legal framework continues to deliver an adequate level of legal certainty (see further European Commission, 2. 2018.). Similarly, Geistfeld points out in the context of autonomous vehicles which could be also applicable in the medical domain the rate at which the market converts from conventional autonomous vehicles depends on the price that consumers are requested to pay in order to adopt the new technologies. He indicates two reasons, where systematic legal uncertainty about the manufacturer liability raises the cost of an autonomous vehicle, thereby increasing the price and reducing consumer demand for these new technologies (Geistfeld, CALIF. L. REV., 1617 (2017).).

[474] The point on the confidence interval was raised by Rajesh Ranganath (Courant Institute of Mathematical Sciences, NYU) in a discussion we had on these challenges.

prediction. Thereafter, the physician would be able to "translate" this combined information to the patient. The patient would then be able to better comprehend the ML prediction allowing her to consider the inherent risks involved and make an informed choice whether or not to follow this prediction. As explained above, the ML device providing explanations for its prediction could be revealing hidden albeit foreseeable risks. Explainable ML and/or delivering confidence intervals could in certain cases make a ML warning in medicine genuinely "comprehensible,"[475] "accurate and clear."[476]

[475] *Pavlides v. Galveston Yacht Basin*, F. 2d at 338.
[476] *Martin v. Hacker* N.E. 2d at 1312.