Penultimate draft. Final version published *open access* at
https://link.springer.com/article/10.1007/s43681-022-00141-z

# EXPLAINABLE MACHINE LEARNING PRACTICES: OPENING AN ADDITIONAL BLACK-BOX FOR RELIABLE MEDICAL AI

Emanuele Ratti[1]

Mark Graves[2]

**Abstract.** In the past few years, machine learning (ML) tools have been implemented with success in the medical context. However, several practitioners have raised concerns about the lack of transparency – at the algorithmic level – of many of these tools; and solutions from the field of Explainable AI (XAI) have been seen as a way to open the 'black-box' and make the tools more trustworthy. Recently, Alex London has argued that in the medical context we do not need machine learning tools to be interpretable at the algorithmic level to make them trustworthy, as long as they meet some strict empirical desiderata. In this paper, we analyse and develop London's position. In particular, we make two claims. First, we claim that London's solution to the problem of trust can potentially address another problem, which is how to evaluate the reliability of ML tools in medicine for regulatory purposes. Second, we claim that in order to deal with this problem, we need to develop London's views by shifting the focus from the opacity of algorithmic details to the opacity of the way in which ML tools are trained and built. We claim that in order to regulate AI tools and evaluate their reliability, agencies need an explanation of how ML tools have been built, which requires documenting and justifying the technical choices that practitioners have made in designing such tools. This is because different algorithmic designs may lead to different outcomes, and to the realization of different purposes. But given that technical choices underlying algorithmic design are shaped by value-laden considerations, opening the black-box of the design process means also making transparent and motivating (technical and ethical) values and preferences behind such choices. By using tools from philosophy of technology and philosophy of science, we elaborate a framework showing how an explanation of the training processes of ML tools in medicine should look like.

**Keywords**: black-box; machine learning; medical AI; reliable AI; values; trustworthiness

## 1 INTRODUCTION

In the past few years, machine learning (ML) tools have been implemented with enthusiasm in medicine. Diagnostic ML tools can be as accurate as human experts, sometimes more (Topol 2019; Chockley and Emanuel 2016). Moreover, ML tools can highlight interesting connections between heterogenous data (Akkus et al. 2017). But despite these stories of success, physicians and

[1] mnl.ratti@gmail.com , Johannes Kepler University Linz, and Technion Israel Institute of Technology

[2] Parexel AI Labs, San Francisco, CA, USA

clinicians are worried because many predictions are generated by algorithms that are opaque, namely that it is not clear how the algorithm has arrived at a particular output. This lack of understanding is problematic, because opacity does "not foster trust and acceptance and most of all responsibility" (Holzinger et al. 2020, p 194). Despite these widespread worries, London (2019) in a recent article has argued that we do not need explanations of opaque algorithms, and all we need to trust ML tools in medicine are ways to measure precisely and reliably their performance within a certain context. In this paper, we analyze London's position and make two claims.

First, we say that London's solution to the problem of trust does not really address such a problem. Rather, London's strategy is the first step towards clarifying another problem, namely how to evaluate the reliability of ML tools in medicine. Establishing whether ML tools are reliable and fulfill their intended function is fundamental to properly regulate their introduction in medical settings. In other words, individual practitioners lack the resources to fully evaluate medical ML systems sufficiently to trust them. Establishing whether ML tools are trustworthy requires resources usually associated with regulatory bodies.

Our second claim is that this problem of regulation can be addressed by shifting the attention from the opacity of algorithms, to another type of opacity, which requires another type of explanation, different from the ones sought to explain the internal functioning of algorithms. This opacity refers to the procedures to train algorithms, which are replete with choices that have limited documentation (Mitchell 2019; Gebru 2018). Although training is usually considered only in the context of supervised ML, this opacity arises in unsupervised and hybrid approaches as well.[3] Given that choices shaping training make a difference for the final output (over and above details of how algorithms optimize functions), for the performances London refers to, and for the tools to fulfill their intended purpose, we argue that to regulate ML tools in medicine another type of explanation is needed. In particular, we need to document the algorithm training process by describing the technical decisions made from problem selection to model deployment; but this is not enough: given that these technical choices are necessarily shaped by value-laden considerations (Ratti 2020), motivating both cognitive and non-cognitive values influencing the technical decisions behind training is necessary. Therefore, the explanation we seek is

[3] To a certain extent, all of the sociotechnical, dataset and hyperparameter selection concerns of supervised ML are also relevant for unsupervised ML. The main difference is only the externally generated "correct" responses are missing from unsupervised approaches. Selection of the dataset for unsupervised ML can be particularly important, as that may capture implicit bias (Dev et al., 2020).

documentation *plus* motivation for why certain technical choices have been made (Kroll 2018). This explanation would show that the AI tool has been designed for an intended use, and it realizes this purpose in a well-defined context. We formulate this proposal by developing US Food and Drugs Administration's (FDA) reflections (2019) on the regulation of AI tools in medicine into a framework based on works in philosophy of technology and philosophy of science[4].

The structure of the paper is as follows. In Section 2, we discuss London's position, and we show how his ideas are promising in tackling the problem of how to establish the reliability of ML tools, which can be useful in a regulatory context. In Section 3, we describe more in detail the problem of regulation of ML tools, by describing FDA suggestions. For the purpose of establishing the reliability of ML tools in medicine, the FDA taskforce emphasizes not only the importance of meeting empirical standards (as London does), but also that these tools should be designed in the right way. This means that we should document the way the devices have been built, and the important technical choices made throughout the process. We provide a framework based on works in philosophy of technology and science to interpret FDA claims. In Section 4, we elaborate further our framework to show how explanations meeting FDA desiderata should look like, and we specify in detail the role of values in it. Finally, we identify some limitations of our approach.

## 2. LONDON'S POSITION

In (2019), London argues against explainability and interpretability. These are usually interpreted as solutions to address the problem of how physicians can trust ML tools when they are opaque. However, London claims, opacity and trust should be addressed differently.

According to his analysis, ML in medicine *seems* to suffer from three limitations. First, ML systems are theory-agnostic, given that they 'learn' a model from vast data sets, without using domain-knowledge information. But domain knowledge seems required in medicine, given that medical decisions are shaped by that knowledge (Zihni et al. 2020). Second, (with the exception of a few tools) ML does not track causal relationships, but merely patterns and regularities in datasets. But this seems problematic, because medical experts justify decisions by appealing to explanations structured as logical arguments filled with causal content (Shortliffe and Sepulveda

[4] Please note that we are not formulating a dilemma. Addressing reliability for regulatory purposes does not exclude addressing issues of trustworthiness caused by algorithmic opacity. However, in this article we are agnostic with respect to the problem of algorithmic opacity and trustworthiness.

2018; Zihni et al. 2020). Finally, algorithmic systems can be black-boxes. However, opacity diminishes trust between the system and the medical expert, and can result in resisting the use of ML tools (Diprose et al. 2020). For these reasons, it is generally accepted that we should aim for less inscrutable tools that can explain why they have generated certain outputs rather than others, in a way that medical experts can understand. Explainable AI tools (XAI) has been developed to deal with these problems for some years.

Despite these convincing points, London claims that in medicine we do not need algorithmic systems to be more interpretable. Far from being flaws, theory-agnosticism, opacity, and associationism characterize the practice of modern medicine. As long as algorithmic systems are validated by means of rigorous empirical testing, we do not need explanations of why they work. London's position is compelling, but it suffers from two problems that, when raised, lead to the identification of another type of opacity. This opacity requires a different approach than the one envisaged by classical literature in XAI.

### 2.1 Regulatory practices

The first problem has to do with the solution offered by London to the issue of trust.

It is true that the importance of XAI has been overemphasized, and that opening the black-box may not always be necessary for trustworthiness. However, London's solution to the problem of opacity is, in our interpretation, a proposal apt to address another problem. In order to describe this point, let us consider London's discussion of the case of the algorithm ranking asthmatic patients. In this popular case (Caruana et al 2015), an algorithm ranked asthmatic patients as having a low probability of dying from pneumonia (because it ignored the extensive treatments asthmatics currently receive). The typical solution to this problem is to opt for less accurate and more interpretable models. But London argues that we do not need less accurate and more explainable models; rather, we need thorough empirical validation, a precise description of the tool's intended use, and how its characteristics make it likely to robustly support the intended use[5]. In this way, we will also know what the validation is actually measuring, thereby resulting in a more reliable tool. In our opinion, what is implicit in London's considerations is that in order to prove the

[5] "[G]reat emphasis should be placed on ensuring that data sets and analytical approaches are aligned with the decisions and uses they are intended to facilitate" (2019, p 20)

reliability of a ML tool in the medical context, we need show that its characteristics are desirable and are the best for achieving a certain purpose in a specific deployment context.

This is a valuable input, but London does not specify in detail *who* should make sure that the tool, from this perspective, functions well. This should not depend upon a single doctor: although a thorough description of the intended use and performance may increase trust, it will be impractical to appeal to in every case. At the end, London mentions 'regulatory practices'[6]. Rather than the single doctor, we agree that this should be the task of agencies supervising the regulation of these tools. But if this is true, then London's solution does not necessarily address the problem of trust between doctors and AI tools. Rather, it speaks to the problem of how we regulate the introduction of ML tools in the medical context by making sure that they function well.

### 2.2 Another type of opacity

The second problem we see is the implicit use of an analogy that obfuscates the issues at stake**.**

Throughout his article, London employs an analogy between ML and pharmaceuticals in order to motivate the focus on empirical metrics and intended use. He uses examples of drugs efficacious without the mechanism of action being known. He describes procedures where theoretical/mechanistic reasoning caused harm until considerable empirical evidence justified changing treatment options. If all of this is true for pharmaceuticals, then the same applies to AI tools as well.

But by questioning the analogy, we discover that more is needed for regulating the tools form the point of view describe in 2.1. Clinicians use pharmaceuticals to directly treat patients, while AI tools are used as instruments to assist clinicians and physicians. In other words, AI tools are devices, like stethoscopes or fMRI machines. Looking at ML tools either as medical devices or as pharmaceuticals suggest different ways of evaluating their fit for regulatory purposes.

If we consider AI tools as devices, then assessing whether they function well for regulatory purposes implies looking not only at empirical metrics (which remain fundamental), but also at *how these tools are built*, and which criteria we use to determine if they function properly. But by starting to investigate how algorithmic systems are built, we realize that the choices made by practitioners have limited documentation - *the ML practice of training algorithmic systems is*

[6] "[R]egulatory practices should establish procedures that limit the use of machine learning systems to specific tasks" (2019, p 20)

*opaque, and for regulatory purposes it needs to be transparent*. In this article, we develop this point: algorithmic systems are tools, and tools are designed for specific purposes. To claim that your tool is the right tool for the purpose, you have to show (1) how its design and training process does facilitate the realization of the purpose, and (2) that in fact it realizes the purpose itself. In other words, it means motivating why the characteristics of a ML tool are desirable to achieve a specific purpose in a specific deployment context.

## 3. ML TOOLS AS MEDICAL DEVICES

What does it mean concretely to treat ML tools as medical devices, and evaluate them from this point of view for regulatory purposes? One prominent example comes from the FDA (2019).

### 3.1 FDA on regulating ML

In 2019, the FDA published a discussion paper to elaborate a series of recommendations to regulate ML in the medical context. In general, ML tools in medicine fall under the category of Software as Medical Devices (SaMD). The FDA relies on the International Medical Device Regulators Forum (IMDRF) definition of SaMD as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device", where medical purposes are defined "as those purposes that are intended to treat, diagnose, cure, mitigate, or prevent disease or other conditions" (2019, p 2). The challenge for regulators is that up until now any substantial changes to SaMD must go through a submission procedure to evaluate the risk posed by the software changes, especially if changes affect performance, safety, intended use of the device, or other clinical functionalities. But ML systems, by design, may change their behavior without human intervention. FDA explains this problem by drawing a distinction between 'locked' and 'unlocked' algorithms. A locked algorithm is defined as "an algorithm that provides the same result each time the same input is applied to it and does not change with use" (2019, p 3). In unlocked algorithms, which includes many ML algorithms, the 'adaptation' can happen after the specific SaMD-ML is put on the market and they may "provide a different output in comparison to the output initially cleared for a given set of inputs" (p 3). The FDA argues for a new framework to regulate this novel dynamic and make sure that they are reliable.

While the focus on modifications is important, we think that algorithm modification is just an instance of a more fundamental issue surrounding modifications. Given other aspects of the FDA proposal, rather than a narrow focus on modifications of unlocked algorithms, a better angle is to characterize the functions of a tool for specific practices and contexts and how either the tool changes keeping the same purpose or the purpose may change. In some cases, we require the tool to maintain those functions for which it was designed, while in others one must explain how the purpose can 'automatically' change in a reliable way. This in turn needs to specify much more than just the possible modifications; it requires a specification of the details of the design of the tools, and how such a design can facilitate a specific purpose. This implies that practitioners should motivate why the characteristics of their tools are desirable and facilitate reliability, safety, and effectiveness. This is possible through the description of a general approach towards the design of ML tools.

The FDA already goes in this general direction as it proposes requirements for ML-based SaMD (from now on SaMD-ML) by opting for a *Total Product Lifecycle Approach*. This lifecycle is divided into AI Model Development, AI Production Model, and AI Device Modifications. In order for this lifecycle to generate a safe and effective tool that will achieve a pre-determined purpose and will stay on it despite adaptation, this approach is based especially on what the FDA calls *good machine learning practices* (GMLP), which are "those AI/ML best practices (i.e. data management, feature extraction, training and evaluation) that are akin to good software engineering practices or system practices" (2019, p 9). In other words, if one goes through the process of development, production, and modification by following GMLP, then the tool will be safe, reliable, and effective (pending meeting some empirical requirements). Following GMLPs will result in a tool with desirable and exemplar characteristics, given a specific purpose and a specific deployment context. However, other than general considerations on data and proper training, GMLP are left largely unspecified.

We get more details once FDA specifies the practices that invest the modifications of SaMD-ML. Given the problem of unlocked algorithms, FDA requires the submission of the types of anticipated modifications (i.e. SaMD Pre-Specifications) and, most important, the Algorithm-Change Protocol (ACP), which is defined as a "step-by-step delineation of the data and procedures to be followed so that the modification achieves its goals" (p 10). There are four categories of ACP, and it is important to notice that these apply more broadly not only to the modifications, but

also to the way the algorithmic system is in general built. Therefore, rather than being strictly about ACP, we interpret these categories as covering the domain of what FDA has previously referred to as GMLP. The categories are data management (e.g. collection protocols, reference standard determination, etc), training (e.g. objectives, choice of ML method, data pre-processing), performance, and update procedures. FDA also mentions transparency, even though here transparency is about GMLP and ACP for users. To sum up, FDA proposes that, in order to introduce SaMD-ML in the medical context that are safe, reliable, and effective, we need to envision the total product lifecycle of how SaMD-ML systems are developed, produced, and modified. The way safety, reliability, and effectiveness are evaluated is not just by verifying that the tool meets some empirical requirements (though these remain fundamental), but also that the design of the tool meets GMLP.

But, as we have said, GMLP remains largely unspecified in the document. In our view, this is because the burden of showing that the design facilitates the purpose reliably is left to practitioners. Practitioners should not merely document the design, but they should also motivate how it is the best among current alternatives, given the intended use/purpose. This means that developers should provide reasons *motivating* the technical choices made throughout the training process. Motivating design choices means showing that the choices are desirable and result in good ML tools. We should aim for those choices that promote valuable characteristics of ML tools that are seen instrumental to achieve a certain purpose well. As we will see, establishing which are 'the valuable characteristics' of ML tools is a value judgement. Going back to the general point, we claim that documenting design and motivations is akin to *explain* the design of ML tools. But what does this mean exactly? In the next section, we introduce a philosophical framework to bring clarity to these considerations.

### 3.2 Situating FDA claims between philosophies of science and technology

#### *3.2.1 Role functions*

There is a rich literature in philosophy of technology debating how to explain the work of artifacts. Explaining how tools work can mean a variety of things. For instance, it can be describing the physical operations performed by tools to produce a certain behavior; but it can also mean describing which design choices have been made to ensure tools produce a certain behavior. Here,

for ML tools we focus on this second meaning. One popular strategy is to extend the mechanistic account of explanation to engineering (van Eck 2015).

The mechanistic account of explanation has received a lot of attention and it has been developed especially in the life sciences (Craver and Darden 2013). Mechanistic explanations are descriptions of how systems produce certain outcomes or how systems are maintained. A popular characterization argues that, after the phenomenon to explain is identified, scientists decompose it in terms of activities and entities deemed causally relevant. The final step is to identify how entities and activities are organized to produce/maintain the explanandum.

As van Eck shows (2015), there is a connection between this mechanistic account and Cummins' framework for functional analysis (1975). In Cummins' view, functional ascriptions describe how an item or an activity contributes to the complex capacity of a system. An entity *x* has a function *f*, if *x* performs *f* in order to contribute to a capacity *c* of a system *y*. If we want to explain the respiratory system, we ascribe a specific function to the heart, which is to pump blood – the fact that the heart emits a noise will not count as a valid functional ascription in this context. Functional ascriptions situate entities or activities in the organization of a system. In mechanistic accounts, the system becomes the phenomenon we want to explain, and functional ascription becomes a matter of understanding entities' roles within a mechanism. These are *mechanistic role functions*, which depend on the *general mechanistic organization*. This account is extremely popular in molecular biology, and metaphors appealing to technical artifacts abound. It is natural to think that the account can be extended to technical artifacts themselves – and the account has indeed been extended to artifacts. We can explain how a car works by showing how its parts contribute to the functioning of the car itself. We can explain this also in terms of design choices, by showing that parts have been designed and arranged in a way that fulfill optimally their role functions, thereby resulting in an excellent car.

In engineering, the mechanistic account has one important distinguished mark (van Eck 2015): we can distinguish between different types of role functions, and each will map into a specific functional decomposition. In particular, there are:

- *Behavior function,* i.e. the desired behavior of a tool
- *Effect function*, i.e. the desired effect of behavior of a tool
- *Purpose function*, i.e. the purpose for which a tool is designed

While behavior functions include the details underpinning the physical behavior of a tool, in the case of effect functions the description refers to the "technologically relevant effects of the physical behaviors of technical artifacts" (2015, p 354). Van Eck makes the example of electric screwdrivers: while the effect function is to loosen or tighten screws, the behavior function is much more detailed, comprising "a conversion of flows of materials, energy, and signals" (2015, p 354). Purpose functions refer to states of affairs that we want to effect in the real world by having the tool perform certain functions via certain behaviors and effects. In the case of the screwdriver, it is to loosen or tighten screws of specific tools in a particular context.

From these role functions, different models of decomposition can be derived. The functional decomposition – be it behavioral, effect, or purpose – explains the particular role function as a whole by describing how the function is produced by the interacting parts of different 'modules'. For instance, effect function decomposition is a model of an organized set of effect functions, delineated by specifying how "the component functions realise the module functions and how the module functions realise the overall function" (van Eck 2011, p 841). In the case of electric screwdrivers, it will be to explain how the different parts of the tool interact in a way that produces the desired effect (i.e. loosen/tighten). Most important, effect function decompositions of artifacts do not merely describe how desired effects are obtained, but it can also include reasons why the parts of the tools have been designed in a way that the desired effect obtained is optimal. Consider the design of a personal computer: we can motivate the choice of a processor by saying that, given the desired effect of excellent graphics performance, the processor was the best available. The effect is connected to the purpose, in the sense that effect functions may be necessary or instrumental to purposes. In the case of the computer, we can motivate the desired effect (i.e. excellent graphics performances) by saying that this is instrumental to the purpose (i.e. using the computer for gaming).

### *3.2.2. SaMD-ML and role functions*

How do these role functions and decompositions map into documenting and explaining design choices of SaMD-ML? What type of role function do we need to ensure the reliability, and hence the effectiveness and safety of SaMD-ML? Which decomposition do we use in order to explain the desired role function?

First, in SaMD-ML there is little emphasis on behavioral role functions. Behavioral role functions include the hardware underlying the ML system. While its materiality makes it (allegedly) less interesting for the context of SaMD-ML, it has been noted that hardware components play a significant role in issues about ML reproducibility (Heil et al 2021). Given the lack of space and the lack of emphasis of FDA on these issues, we leave considerations of behavioral role function for another work.

What is at stake here are *effect* and *purpose functions*. In order to assess SaMD-ML from the FDA's perspective, we need to explain those role functions, and how the overall effect function is instrumental for the purpose function. This is another way of expressing London's views, namely the idea of specifying meaningful endpoints (i.e. effect functions) in specific deployment contexts (i.e. facilitating purpose function in specific contexts).

The overall purpose function is described as the intended use of SaMD-ML, of which the FDA suggests three: to treat and diagnose; drive clinical management; inform clinical management. These are general categories listed by FDA, and more nuanced distinctions are possible (Liu et al. 2019). FDA wants to make sure that the characteristics of the tool allow the user to realize the purpose function reliably. In our understanding, a necessary condition for this is to achieve a *desired effect*, conceptualized by the FDA as a *positive clinical evaluation*. This is a combination of the estimates of valid clinical association (i.e. valid clinical association between ML output and the targeted clinical condition), analytical validation (reliably processing input data to generate reliable outputs), and clinical validation (ML outputs targeting the right population in the context of clinical care). To connect the dots more precisely, if we want to ensure that the purpose function is achieved by the user, an assessment of the overall effect function is necessary. Evaluating the effect function includes both an effect function decomposition as well as quantitative evaluation of the performances of the tool (as London correctly argues).

Therefore, SaMD-MLs are evaluated by assessing whether the effect function makes possible the purpose function. Different positions can possibly emerge on this topic (Figure 1). In what we characterize as Option A, one can say that to assess how effect functions reliably facilitate purposes, we need an effect function decomposition intuitively characterized as *a thorough mechanistic description of how the algorithm has produced an output*: the relevant aspects needed to explain the overall effect function are the details of how algorithms would make predictions from data, which is a position that can be popular among XAI's proponents. In Option B, one may

claim that we do not need a (algorithmic) decomposition of the effect function, because we only need to measure empirical performances (this partially overlaps with London's position), usually with metrics as discrimination and calibration (Chen et al 2019). Discrimination, for instance, is the ability of SaMD-ML to distinguish different conditions (such as the presence of breast cancer or not in a biopsy), and typical metrics used for calibration are sensitivity, specificity, etc.

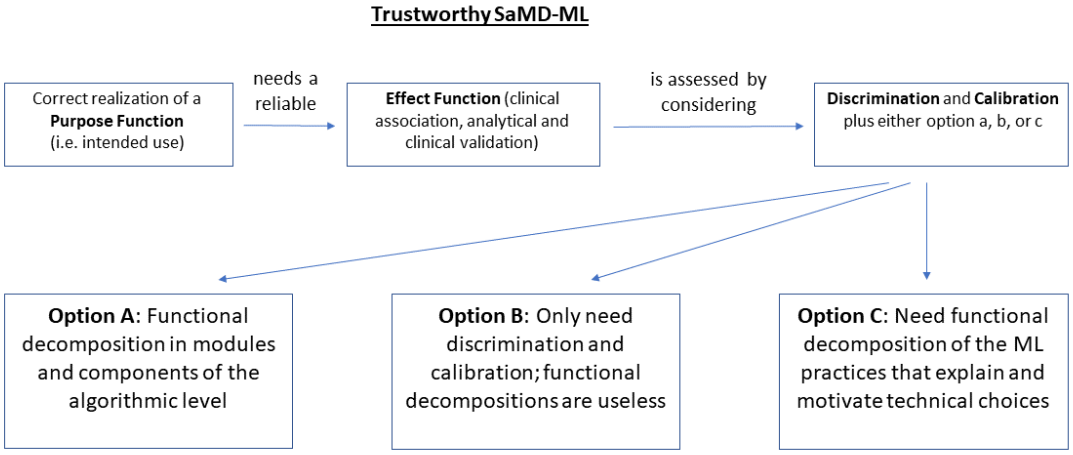


**Figure 1**. Purpose function, and different ways of decomposing the effect function of SaMD-ML

### 3.3. Alternative ways of decomposing the effect function

In this article we argue for Option C. Algorithmic details are not the only 'modules' through which the overall effect function of SaMD-ML can be decomposed. The FDA in its paper refers to a description of how the algorithm has been trained, and why it has been trained in one way and not another. Explaining design choices of the algorithmic training means not only describing different aspects of SaMD-ML and why the tool behaves in such and such a way, but it also means motivating why technical choices of the design are the best for the overall effect function and the purpose function (Option C, Figure 1).

Different training choices would produce different overall effect functions, which have consequences for the purpose function. As an example illustrating the point (and showing the distinction between effect and purpose), consider London's interpretation of the case of the neural net and asthmatic patients. The goal of such a tool may be to facilitate a reliable allocation of resources "against a baseline risk of death that is independent of current medical practice" (2019,

p 19) – this is the purpose. The effect is to provide a risk score for patients. The problem of the tools (i.e. its unreliability in providing a desired effect instrumental to the purpose) was that training data reflected the probability of death *given* standard medical practice – asthmatic patients are treated more aggressively, and hence the probability of them dying from pneumonia is low given standard care. The only way to find out about this problem is to document thoroughly the process of building the algorithmic system, and to motivate the choices behind it. If the FDA has to evaluate whether the overall effect function of the neural net (e.g. provide a risk score for patients independently of current medical practice) is instrumental for realizing the purpose function (e.g. prioritize resources in an ICU), one important aspect will be considering whether training data can actually be used to train the algorithm in a way that realize the overall effect function. We cannot identify the problem unless there is explicit attention to the different aspects of the construction of the tool. In this case, using different data sets will lead to different effect functions, which in turn will be more or less instrumental for the purpose function.

It is here that technical practices - GMLP - become relevant. For ML systems that adapt within a regulated environment (called "unlocked" by FDA), the FDA suggests that practices must be "good". We interpret this as saying that effect functions can change as long as the purpose remains the same, and the changes take place under GMLP which will ensure the reliability of the tool. This shifts the focus away from algorithmic decomposition to showing how technical practices result in a better design that will facilitate the purpose. Empirical requirements are satisfied here: it is within functional decomposition with respect to purpose and practices that discrimination and calibration would be necessary (option B) to show agreement with the purpose.

But what do we mean by 'technical practice'? We refer to concrete technical choices. In the FDA document, choices constituting GMLP are only vaguely listed. Here we claim that ML practitioners should decompose the overall effect function from a procedural angle, by dividing the training process into 'modules' or 'phases' (based upon varying technical practices), and they should motivate fundamental technical choices that they have made, where fundamental choices are the ones making a difference to the overall effect function and, as a consequence, to the purpose function. We claim that documenting and motivating choices to make sure that effect and purpose are aligned is *explaining the design of SaMD-ML*.

## 4 EXPLAINABLE AI WITHOUT XAI: OPENING THE BLACK-BOX OF ML PRACTICE

With all the information from Section 3, we claim that explaining ML practice has two components:

(a) a description of the way SaMD-ML has been developed and constructed
(b) a motivation for the fundamental technical choices made by the ML practitioners

Complying with a) only will provide a long list of technical requirements and specifications, spelled out as neutral, step-by-step recipes. Formulating b) means also providing reasons why the technical choices made are best and result in the overall effect really allowing the purpose function. It has been show that documenting technical choices is a neglected practice (Mitchell 2019; Gebru 2018), and motivating choices even more. Moreover, the nature of technical choices reveals that the practice of ML is replete with value judgements. How we motivate technical choices is shaped by technical constraints, but it is not limited to them – here we claim that value-laden judgements are inevitable in ML practice, given that technical choices are underdetermined (Ratti 2020). In this section we describe first a pipeline to accommodate how a) can be formulated (4.1), then clarify in which sense technical choices are necessarily shaped by values (4.2), describe b) in detail (4.3), and finally characterize limitations of our approach (4.4).

### 4.1 Documenting ML Practices

Decomposing the overall effect functions by identifying modules based upon the practices used to develop SaMD-ML for a specific purpose, can be done in different ways. Chen et al (2019) characterizes a pipeline for healthcare, and we extend that with best practices from data science, e.g., to distinguish data understanding and preparation from data collection and model development. This builds upon the FDA's AI Model Development and gives a functional decomposition of the aspects of ML that can explain an overall effect function of SaMD-ML from the point of view of design choices. We characterize the design and development of SaMD-ML as six stages, which we identify as 'modules of the overall effect function' (Figure 2).

First, there is what Chen et al (2019) call problem selection (i.e. understanding and definition). In developing SaMD-ML, one has to choose carefully which prediction or other ML tasks the SaMD-ML will perform in meeting the purpose. Although driven by the sociotechnical

processes constraining the ML task, the first step of the process is to understand those needs and constraints and formally define the problem to be addressed, i.e., one defines the purpose and overall effect function. This choice has important ramifications for the other phases. As a working example, consider the problem of building a ML system to extract diagnostic information from Electronic Health Records (EHR). The purpose could be to extract an explicit diagnosis; to extract information sufficient for diagnosis (such as test results), even if not previously coded; to create cohorts or subpopulations for targeted treatment; to select individuals (or medical institutions) for clinical trials or retrospective investigations; etc. The effect function specifies a diagnosis given the data, typically with a predictive model.[7]

The second module is data acquisition. Acquiring training data for ML adds complexity beyond the task taken in isolation, depending on the problem one has selected. Even if data exists in some form within organizations, additional steps are needed to select the data and ensure it is fit for the purpose. For instance, if one needs data to train a model for identifying the possibility of a rare disease from medical records, then one must train on many examples of records with the disease, much more than proportionate to the population. Moreover, depending upon the purpose, one may need different kinds of negative examples, e.g., the occurrence and results of different tests would vary in their relevance depending upon whether the system is extracting a diagnosis or predicting a disease without formal diagnoses.

The third stage/module is data understanding and preparation, which include characterizing the data, especially issues of quality (e.g. cleaning, transforming, reducing) to prepare data for analysis and modeling. Although glossed over by Chen et al, this stage is well characterized in data science and data mining (Martínez-Plumed et al. 2019). Data preparation is also constrained by requirements of the modeling algorithms to be used and the need to divide the set into developmental (e.g., training and tuning) and a testing (or clinical validation) data sets for supervised approaches. In unsupervised ML, the data scientist may need to understand the biases that may occur, such as gender or race imbalances, and adjust or augment datasets.

The fourth module is model development, which is the phase where the algorithm is run on the data to create the model or other ML construct, e.g., trained and tuned with the

[7] Clustering approaches might also be used to create a treatment group not otherwise differentiated by a formal diagnosis (e.g., if a disease could have different subtypes not clinically well-distinguished with respect to the proposed treatment).

developmental data sets. This is the module/stage where most of the choices for the 'model architecture design' take place. In designing the architecture, parameters and hyperparameters are fundamental. Parameters are usually defined as the internal values of the model we are generating via the training of the algorithm. The number of parameters that can be estimated from data in the model would be a hyperparameter and should also be appropriate to the dataset size and complexity as they influence how parameters are learned during training. Examples include k in k-nearest neighbors or k-means, number and size of hidden layers in NN, etc. Opening up the black box of training requires explaining why certain hyperparameters were chosen, e.g., was it a convenient default or the result of substantial tuning efforts with a specific aim.

Next, there is the stage of validation of performances and interpretation for its intended use (testing, or clinical validation), where metrics similar to the ones of tuning are evaluated on the test set, but additional emphasis is placed on clinical results. Chen et al. clarify the distinction between validation in ML (what they and we call tuning) and validation in a clinical setting, though we would add the need for those clinical criteria to inform the selection of tuning metrics used in model development. Thus, clinical validation criteria can also affect the training, especially via model tuning.

The last module/stage is assessment of model impact and deployment and monitoring. As Chen et al. notice, "a performant model alone is insufficient to create clinical impact" (p 413). From a data science perspective, deployment is a prerequisite for assessing impact. Reconciling the clinical and ML perspectives illuminates the need to recognize that the model should be deployed in as realistic setting as possible to assess impact, and that many follow up steps may be required for regulatory approval. What we have to do here is show how design choices of, e.g., a visualization or user interface for a 'dashboard', allow a range of usage modalities. Moreover, design choices can facilitate or impede the integration of the tool in specific clinical workflows.

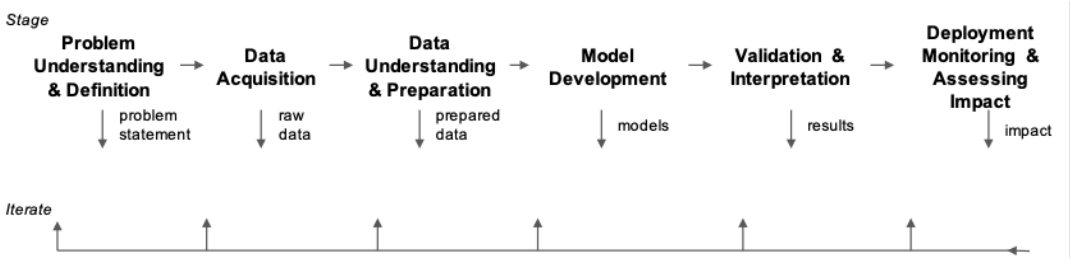


***Figure 2:*** *Modules of the Machine Learning practices*

### 4.2 Value judgement in training

Section 4.1 shows that there are a lot of aspects that have to be documented during ML practices. We have mentioned throughout this article that some choices taken in the process are pervaded by value judgements. What does this mean exactly, and why does this matter? In order to explain this, we draw a parallel between the context of this article and the problem of theory choice in philosophy of science. The context of this article is understanding the reliability of ML tools in medicine. We have translated this problem into the problem of establishing what is the best design for a SaMD-ML given a specific context of deployment and a given intended use. This problem, in turn, can be addressed by understanding which are the best technical choices that practitioners have to make in order to design the best SaMD-ML. We think that these ideas can be approached by thinking about design/technical choices under the lens of the problem of theory choice in philosophy of science, and this will show that in which senses technical choices are value-laden.

How scientists choose among theories/hypotheses is highly contentious. Ideally, theory choice is determined by criteria that scientific theories should meet in order to be considered 'good scientific theories'. Various lists are present in the literature, but they are all variations on Kuhn's seminal list (1977), which includes predictive accuracy, internal coherence, external consistency, unifying power, and fertility. When we say that theory choice is ideally determined by these criteria, we mean that it would be very convenient if these functioned as rules: a theory/hypothesis with more unifying power is better than one with less, etc. If theory choice functioned as an algorithmic procedure, one would be able to apply those criteria unambiguously. However, Kuhn argues that this is a misleading idealization, because criteria are imprecise (i.e. individuals disagree on how to apply them in concrete cases), and they conflict with one another. Therefore, Kuhn concludes that "the criteria of theory choice with which I began function not as rules, which determine choice, but as values which influence it" (1977, p 362). McMullin in a seminal paper (1983) makes similar considerations: theory choice is a procedure close to value judgement, meaning it is not an unambiguous procedure determining which choice is the best, but a propensity to consider certain characteristics as more desirable. In addition to the difficulties of deciding which epistemic value is the best for a theory in a given context, it has also been shown that non-epistemic values, such as social and moral values, shape theory/hypothesis choice. In particular, because of a gap between hypothesis and evidence, one faces inductive risk, namely

accepting/rejecting a hypothesis with the risk that it will turn out to be false/true; and usually deciding to accept or reject the hypothesis is a function of the seriousness of making that mistake, where 'seriousness' can be evaluated from the point of view of non-epistemic values (Rudner 1953; Hempel 1966; Douglas 2009; Elliott and Richards 2017).

Both the arguments from epistemic and non-epistemic values point to an underdetermination: any dataset is insufficient to determine which theories are best, and we need to resort to pragmatic considerations (informed by various values, both epistemic and non-epistemic) to justify theory choice convincingly. But there is evidence that ML faces analogous problems (Ratti 2020): throughout the ML pipeline, technical choices are underdetermined and practitioners resort to value-laden considerations. Thus, a step-by-step description will not suffice to document the development of SaMD-ML. In order to understand what is really behind a tool, we must understand the values shaping and constraining the development of these tools, because any design necessarily is shaped by technical choices that are influenced by values.

But what is exactly a value here? McMullin argues that something counts as a value in a specific entity if "it is desirable for an entity of that kind" (p 5). Why something is desirable can vary a lot; in the case of epistemic values in science, those are values because they are conducive to truth. But here we are in a different context. Technical choices are value-laden when the reasons to take them are based on the expectation that they will promote certain valuable characteristics of that process or, in our case, of the SaMD-ML. There are two valuable characteristics of the practices that we think are relevant:

- there are characteristics of ML practices that simply make SaMD-ML a more reliable tool assuming certain goal performances, some of which will be subordinated to the purpose function (e.g. usability, as we will see). But others will be independent of the purpose function (e.g. the metrics London refers to). In other words, there will be technical choices that will allow SaMD-ML to meet some performance metrics better than others. We call these values *performance-centered values*
- there are things that we find desirable because they result in an overall effect function that does not harm data subjects, or that data subjects may possibly benefit from. Characteristics

that we can find desirable from this perspective are shaped by *social, political, and moral values.*[8]

In the next section we will identify some of these values in the different modules of the ML practice (see Table 1 for a summary).

### 4.3. Identifying values in ML practices

Let us now see in detail how values and technical choices influence each other in every single phase (i.e. module) described in 4.1.

The first module (i.e., *problem understanding and definition*), requires seeking understanding of what is needed and making choices in how to define the problem. The overall effect function aims to solve the problem statement (or answer the posed questions) for the stated purpose. Performance-centered values include consistency between the constructs of the problem statement and external needs as well as its internal coherence. Moreover, choosing the problem to address incorporates a number of human biases and assumptions (including well-reasoned clinical ones) that need to be made explicit, especially as the project most likely requires cross-discipline collaboration.

In *data acquisition*, data availability is a performance-centered value playing a key role. But we want to ensure not only that there is enough data to train properly, but also that the data is representative enough, given the particular goal. Going back to the example of asthmatic patients, it may be difficult to find data sets about patients that are independent of medical care received, and one may need to redefine the problem to account for diagnostic information and resulting treatments actually available. How much value is placed on available data versus acquiring more

[8] A common, though controversial, way of talking about values is by distinguishing direct vs indirect roles for values. Here, disentangling direct and indirect roles is difficult. Admittedly, sometimes this is straightforward: values play an indirect role especially when we are measuring the performance of the trained models - in particular, when we have to interpret the results in terms of false positive and false negative. But in other contexts, the distinction between direct and indirect is more complex. For instance, if we have to decide whether data available are enough, or are of enough quality, then we have to make a decision about thresholds, and sometimes these thresholds will be determined by values playing an indirect role. But it can also be that the very notion of what constitutes representativeness or good quality of data is shaped by values, and in this case the values will play a direct role, because it is the value itself that acts as "stand-alone reasons to motivate our choice" (Douglas 2009, p 96). We think that talking about values in the way we do - by referring to the characteristics deemed desirable that the values promote - is a much more straightforward way of understanding the relation between values and technical choices.

data will determine how SaMD-ML will be developed to address a problem. In addition, racial and ethnic diversity of training data often fails to match the diversity of the population—a problem in both ML and clinical trials (Gianfrancesco et al., 2018, Knepper & McLeod 2018). Moreover, given that there are different ways of thinking about representativeness that cuts across different social factors, sometimes the values of *availability of data* and *representativeness/inclusiveness* (understood as an ethical and social value) can stand in an odd relation. Although data may appear sufficiently representative, it might only come from well-served and wealthy areas. But collecting data coming from underserved areas may mean processing poor data, possibly optical character recognition (OCR) of scanned records, with lots of gaps. We will have to justify the use of this problematic data by saying that we want to promote inclusiveness that cuts across not only different ethnic groups, but also groups with different incomes. Moreover, in the example of rare diseases in Section 4.1, sometimes rather than just listing the dataset used, one should explain the decision to *oversample* from a disease group (or other underrepresented subpopulation) in terms of the desired effect on the model. Finally, data acquisition requires identifying legal issues around ownership, *ethical concerns* about privacy and implicit bias, and *technical challenges* for balancing data representation and identifying underlying changes over time.

In *data understanding and preparation*, conflicting values may lead to different choices. The value of *inclusion/representativeness* can stand in an odd relation with '*data quality'* (a performance-centered value). Varying the threshold separating acceptable versus low quality data may lead to more or less inclusive SaMD-ML. Another important aspect is how we define, especially in the medical context, a reference truth for our tools; this, according to Chen et al., requires clinical judgement, and we should make explicit the performance-centered values of what makes a *good reference truth*. Eliminating statistical outliers may lead to cleaner data, and more useful models, but related outliers in the overall population might indicate subpopulations that should be considered separately.

In model development, we face similar conundrums. In choosing an algorithm, we may favor one that generates complex models, but if we do not have enough data, then we may want to opt for a less complex algorithm. Here the performance-centered values of *data availability* and *complexity* constrain one another. But choosing an algorithm depends also on data modalities (2D images, 3D volumes, lab measurements, texts, etc). For data including simple or engineered features such as clinical characteristics, then linear/logistic regression, support vector machines,

and decision tree are considered accurate, while for images convolutional neural networks seems to be the norm (Chen et al. 2019). Which are the characteristics that we are looking for in an algorithm given a particular data modality? Other values central in developing a model are *similarity* (i.e. the model should accurately reflect the phenomena of interest) or *generality* (i.e. the model performs robustly with respect to novel data), and sometimes these can stand in a tradeoff relation. Because the output of the model contributes directly to the overall effect function, the adaptability and opacity of ML models are central to what needs explaining. Our approach differs significantly in this respect from London and other approaches to transparency or XAI. Rather than decompose how the model generates the effect function, or calibrate it to empirical clinical results, we emphasize justifications for the inputs to the training, good practices in selecting and tuning the algorithm, justifying that the model is appropriate for the purpose, and ML and clinical validation of the model outputs.

The stage of validation is replete with ethical and social values shaping performance-centered values, especially considerations of inductive risks. Metrics will change depending on how the SaMD-ML deals with conditions and level of risk. One may choose metrics minimizing either false positives or false negatives depending on both the purpose function, as well as ethical and social values – e.g. either to minimize false positives with diagnosing non-threatening diseases and costly follow-up or to minimize false negatives with effective early treatment of serious or contagious disease. It is frequently necessary to trade off either reducing false positives or false negatives, as it may be not possible to reduce both simultaneously, and these tradeoffs can be captured by metric pairs precision-recall and sensitivity-specificity. The two tradeoffs are similar, with sensitivity mathematically identical to recall, and both increase with a reduction in false negatives. The difference between tradeoffs is that while specificity and precision both increase with fewer false positives, specificity takes true negatives into account while precision takes into account true positives. Precision works well in information retrieval to measure relevance when retrieving a few documents out of potentially millions, but specificity works better when modeling high-prevalence disease to identify healthy individuals (i.e., true negatives for the disease). There is no technical rationale for choosing among these metrics, and how to adjust the tradeoffs may depend heavily upon performance-centered as well as social/moral values. If the purpose of the SAMD-ML involves diagnosing a disease, then sensitivity-specificity may be a better tradeoff, while if the system is retrieving disease information from patient records, then precision-recall

might be better. Design of the SAMD-ML involves not only selecting one of the tradeoffs, it involves justifying the reasons one tradeoff was selected for model validation over the other one.

Finally, there is model impact and deployment and monitoring. In this module, the user interface of the tool is mostly driven by the performance-centered value of *usability*. In other words, designers should motivate the interface by describing how it affords usability. But usability is not everything. We also want to make sure that the interface does not mislead users into thinking that they have obtained a particular piece of information, while in fact they have not, e.g, *graphical integrity*. When presenting a relative quantity, such as a relative difference, ratio, or percentage, it is important that the reference value reflect the user's expectation rather than a modeling artifact. This is because in this case the purpose function will be impacted. Therefore, motivating why the interface does not mislead users is fundamental.

***Table 1***. *Examples of values in the training process. As shown in the text, values can stand in various relations with each other (complementarity, tradeoff, etc)*

| MODULES OF TRAINING | PERFORMANCE-CENTERED VALUES | SOCIAL/ETHICAL/POLITICAL VALUES |
|---|---|---|
| Problem Understanding and Definition | External consistency, internal coherence | Identification of human biases in choosing the problem |
| Data Acquisition | Availability of data | Data representativeness; privacy; implicit biases |
| Data Understanding and Preparation | Data quality; reference truth; data cleanness | Data representativeness |
| Model Development | Complexity; data availability; similarity; generality | Values characterized by ML ethics literature |
| Validation and Interpretation | Precision-recall; sensitivity-specificity | Values related to the inductive risk literature |
| Impact, Deployment, and Monitoring | Usability | Graphical integrity |

### 4.4 Limitations of the account

This overview of how values can influence different choices must not be thought as complete and definite. There are at least three limitations to what we have described in this section. First, we have characterized the relation between values mostly in tradeoffs terms. However, values can stand in many different types of relation. Let us indulge for a moment in an analogy; let us think about values as mathematical variables: as there are different types of relationships between mathematical variables (e.g. linearity, non-linearity, discontinuity, non-monotonicity, etc), so we can make similar considerations with values. The second limitation is that is that it is not clear how the values we described have been selected. We base the identification of values on our professional and research experience, but we acknowledge that introspection is not a substitute for a real methodology, at least not long-term. Finally, our characterization of the interplay between values and technical choices is an idealization, especially because we have represented the data-aware machine learning pipeline as a process where one person takes all decisions; however, the practice of ML is a social practice, involving different actors and stakeholders (Lowrie 2017; Anthony 2021), and this can mean that there will be negotiations of values among different individuals. But these limitations can be overcome in future works. This is a philosophical framework for a much bigger, empirical project. We hope later to use qualitative methods (e.g. interviews, ethnographies) and enrich our list of values, relationships among values, and characterization of negotiations among stakeholders.

Despite these limitations, we think that our framework has substantial advantages over existing frameworks that aims to provide tools to identify values in the practice of data science. Here we mention only a few of these frameworks, with no presumption of completeness. For instance, Loi and colleagues (2020) provide a contribution similar to ours, in particular with the idea of 'design explanation', which is an explanation of the goals of the system, in conjunction with information about why the design of the system is the way it is and the norms and values guiding it. Another example is Selbst and Barocas (2018). They argue that we should understand "values and constraints that shape the conceptualization of the problem (..) how these (...) inform the development of machine learning models (...) how the outputs of models inform final decisions" (p 47). They propose, as others do (Lehr and Ohm 2017; Mulligan et al 2019), to consider 'documentation' of the training process as a way to explain and regulate ML tools. These are all valuable contributions, but when they address values, it seems to us that they refer especially to social/ethical values, while technical/performance-centered values are neglected. What

generally count as 'technical preferences' are seen as less problematic than ethical/social values, even though in principle they are preference and hence values. This assumes that practitioners largely agree on 'technical values', while social/ethical values are somehow arbitrary and hence need to be discussed more thoroughly. However, this is largely misleading: 'technical values' (i.e. performance-centered values) are no less values than ethical/social values (4.2), and they pose the same problems of ethical/social values. They are preferences, they can be seen as arbitrary, they can be endorsed without awareness, etc. The only exception to this trend that we could identify is the work by Birhane and colleagues (2021), that provided a list of 67 values which influence to various degrees ML research, including technical values, and van de Poel (2020) who formalizes technical norms within sociotechnical AI systems. While their work shows the possibilities implied by more 'empirical' approaches, we think that it does not engage enough with the connections between specific technical choices and the values they identify.

## 5 CONCLUSION

By developing some intuitions that London has in (2019) and preliminary work of the FDA (2019) to stimulate ideas on AI regulation in the medical context, we have made the following proposals. In regulating medical AI, we should address not only algorithmic opacity, but also on other black-boxes plaguing these tools. In particular, there are many opaque choices that are made in the training process and in the way algorithmic systems are built, which can potentially impact SaMD-MLs performances, and hence their reliability. Second, we have said that opening this alternative black-box means explaining the training process. This type of explanation is in part documenting the technical choices made from problem selection to model deployment, but it is also motivating those choices by being transparent about the values shaping the choices themselves – in particular, performance-centered values and ethical/social/political values. Overall, our framework can be considered as a starting point to investigate which aspects of the design of AI tools should be made explicit in medicine, in order to inform discussions on the characteristics of reliable AI tools, and how we should regulate them. We have also highlighted some limitations, and we have claimed that in the future it will be necessary to empirically investigate the practice of machine learning in light of our framework, and to identify more nuances in the values shaping ML training.

We want to end this article by repeating that the problem of explaining opaque technical choices is not an alternative to explain the opacity lying at the algorithmic level. Unlike London, we think that the worries about algorithmic opacity in medicine are more than justified. However, we leave any consideration on how the two opacities are connected to each other for future works.

**Conflict of interest**: On behalf of all authors, the corresponding authors states that there is no conflict of interests

**REFERENCES**


Akkus, Z., Ali, I., Sedlář, J., Agrawal, J. P., Parney, I. F., Giannini, C., & Erickson, B. J. (2017). Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence. *Journal of Digital Imaging*, *30*(4), 469–476. https://doi.org/10.1007/s10278-017-9984-3

Anthony, C. (2021). When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies. *Administrative Science Quarterly*, *66*(4), 1173–1212. https://doi.org/10.1177/00018392211016755

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2021). *The Values Encoded in Machine Learning Research*. http://arxiv.org/abs/2106.15590

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015-August, 1721–1730. https://doi.org/10.1145/2783258.2788613

Chen C, Liu Y, and Peng L. 2019. "How to Develop Machine Learning Models for Healthcare." *Nature Materials* 18 (5): 410–14. doi:10.1038/s41563-019-0345-0.

Chockley K, and Emanuel E. 2016. "The End of Radiology? Three Threats to the Future Practice of Radiology." *Journal of the American College of Radiology* 13 (12). Elsevier Inc: 1415–20. doi:10.1016/j.jacr.2016.07.010.

Cummins, Robert. 1975. “Functional Analysis.” *The Journal of Philosophy* 72 (20): 741–65.

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. The University of Chicago Press.

Dev, S., Li, T., Phillips, J. M., & Srikumar, V. (2020). On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 7659–7666. https://doi.org/10.1609/aaai.v34i05.6267

Diprose WK., Buist N, et al. 2020. “Physician Understanding, Explainability, and Trust in a Hypothetical Machine Learning Risk Calculator.” *Journal of the American Medical Informatics Association* 27 (4): 592–600. doi:10.1093/jamia/ocz229.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin, and Ted Richards, eds. 2017. *Exploring Inductive Risk - Case Studies of Values and Science*. Oxford University Press.

FDA. 2019. “Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning ( AI / ML ) -Based Software as a Medical Device ( SaMD ) - Discussion Paper and Request for Feedback.” *U.S Food & Drug Administration*, 1–20.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. 2018. Datasheets for datasets. *ArXiv Preprint ArXiv:1803.09010*.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, *178*(11), 1544. https://doi.org/10.1001/jamainternmed.2018.3763

Heil, B., Hoffman, M., Markowetz, F., Lee, S.-I., Greene, C., & Hicks, S. (2021). Reproducibility standards for machine learning in the life sciences. In Nature Methods (Vol. 18, Issue 10, pp. 1122–1127). Nature Research. https://doi.org/10.1038/s41592-021-01205-4

Hempel, C. (1966). *Philosophy of Natural Science*. Prentice-Hall.

Holzinger A, Carrington A, and Müller H. 2020. "Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations." *KI - Kunstliche Intelligenz* 34 (2). Springer Berlin Heidelberg: 193–98. doi:10.1007/s13218-020-00636-z.

Knepper, T. C., & McLeod, H. L. 2018. When will clinical trials finally reflect diversity? *Nature*, *557*(7704), 157–159. https://doi.org/10.1038/d41586-018-05049-5

Kroll, Joshua A. 2018. "The Fallacy of Inscrutability." *Philosophical Transactions of the Royal Society A*.

Kuhn, Thomas. 1977. "Rationality, Value Judgment, and Theory Choice." In *The Essential Tension*, 320–39. Chicago: Chicago University Press.

Lehr, D., & Ohm, P. (2017). *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*.

Liu Yun, Chen C, et al. 2019. "How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature." *JAMA - Journal of the American Medical Association* 322 (18): 1806–16. doi:10.1001/jama.2019.16489.

London, Alex John. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report* 49 (1): 15–21. doi:10.1002/hast.973.

Loi, M., Ferrario, A., & Viganò, E. (2020). Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-020-09564-w

Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data and Society*, *4*(1). https://doi.org/10.1177/2053951717700925

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández Orallo, J., Kull, M., Lachiche, N., Ramírez Quintana, M. J., & Flach, P. A. 2019. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories." *IEEE Transactions on Knowledge and Data Engineering*. doi:10.1109/TKDE.2019.2962680

Mcmullin, Ernan. 1983. "Values in Science." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2: 686–709.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Mulligan, D. K., Kluttz, D. N., & Kohli, N. (2019). *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*. https://ssrn.com/abstract=3311894

Ratti, Emanuele. 2020. "Phronesis and Automated Science: The Case of Machine Learning and Biology." In *A Critical Reflection on Automated Science - Will Science Remain Human?*, edited by Marta Bertolaso and Fabio Sterpetti. Springer.

Rudner, R. 1953. "The Scientist qua Scientist Makes Value Judgement." *Philosophy of Science* 20 (1): 1–6.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*(3), 1085–1139. https://doi.org/10.2139/ssrn.3126971

Shortliffe, Edward H., and Martin J. Sepúlveda. 2018. "Clinical Decision Support in the Era of Artificial Intelligence." *JAMA - Journal of the American Medical Association* 320 (21): 2199–2200. doi:10.1001/jama.2018.17163.

Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/tnnls.2020.3027314

Topol, EJ. 2019. *Deep Medicine - How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. Minds and Machines, 30(3), 385–409. doi:10.1007/s11023-020-09537-4

van Eck, Dingmar. 2011. "Supporting Design Knowledge Exchange by Converting Models of Functional Decomposition." *Journal of Engineering Design* 22 (11-12): 839–58. doi:10.1080/09544828.2011.603692.

van Eck, Dingmar. 2015. "Mechanistic Explanation in Engineering Science." *European Journal for Philosophy of Science* 5 (3): 349–75. doi:10.1007/s13194-015-0111-3.

Zihni E, Madai VI, et al. 2020. "Opening the Black Box of Artificial Intelligence for Clinical Decision Support: A Study Predicting Stroke Outcome." *PLoS ONE* 15 (4): 1–15. doi:10.1371/journal.pone.0231166.