Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Technical Challenges and Solutions

Eike Petersen*[1], Yannik Potdevin[2], Esfandiar Mohammadi[1], Stephan Zidowitz[3], Sabrina Breyer[1], Dirk Nowotka[2], Sandra Henn[1], Ludwig Pechmann[4], Martin Leucker[1,4], Philipp Rostalski[1,5], and Christian Herzog[1]

[1]*Universität zu Lübeck, Lübeck, Germany*
[2]*Christian-Albrechts-Universität zu Kiel, Kiel, Germany*
[3]*Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany*
[4]*UniTransferKlinik Lübeck, Lübeck, Germany*
[5]*Fraunhofer Research Institution for Individualized and Cell-Based Medical Engineering (IMTE), Lübeck, Germany*

July 19, 2021

Abstract

Machine learning is expected to fuel significant improvements in medical care. To ensure that fundamental principles such as beneficence, respect for human autonomy, prevention of harm, justice, privacy, and transparency are respected, medical machine learning applications must be developed responsibly. A large number of high-level declarations of ethical principles have been put forth for this purpose, but there is a severe lack of technical guidelines explicating the practical consequences for medical machine learning. Similarly, there is currently considerable uncertainty regarding the exact regulatory requirements placed upon medical machine learning systems. In this paper, we survey the technical challenges involved in creating medical machine learning systems responsibly and in conformity with existing regulations, as well as possible solutions to address these challenges. We begin by providing a brief overview of existing regulations affecting medical machine learning, showing that properties such as safety, robustness, reliability, privacy, security, transparency, explainability, and nondiscrimination are all demanded *already* by existing law and regulations — albeit, in many cases, to an uncertain degree. Next, we discuss the key technical obstacles to achieving these desirable properties, and important techniques to overcome those barriers in the medical context. Since most of the technical challenges are very young and new problems frequently emerge, the scientific discourse is rapidly evolving and has not yet converged on clear best-practice solutions. Nevertheless, we aim to illuminate the underlying technical challenges, possible ways for addressing them, and their respective merits and drawbacks. In particular, we notice that distribution shift, spurious correlations, model underspecification, and data scarcity represent severe challenges in the medical context (and others) that are very difficult to solve with classical black-box deep neural networks. Important measures that may help to address these challenges include the use of large and representative datasets and federated learning as a means to that end, the careful exploitation of domain knowledge wherever feasible, the use of inherently transparent models, comprehensive model testing and verification, as well as stakeholder inclusion.

Contents

1 Introduction

The potential of modern machine learning techniques to significantly improve clinical diagnosis and treatment, and to unlock previously infeasible healthcare applications, has been thoroughly demonstrated [1, 2, 3, 4, 5, 6, 7, 8, 9]. At the same time, the risks posed by the application of medical machine learning (MML) have become apparent. *Non-robust models* that do not generalize to broader patient cohorts or to recordings in different settings or using different measurement devices, a *lack of transparency* regarding the reasoning behind model predictions, *privacy* challenges and unintended *discriminatory biases* against certain patient groups have been identified as critical obstacles to overcome if widespread clinical adoption of MML is to be achieved [10, 5, 11]. Currently, the Partnership on AI's "Artificial Intelligence Incident Database"[1] (which is neither limited to medical systems nor machine learning-based systems) lists more than 1200 reported "situations in which AI systems caused, or nearly caused, real-world harm", illustrating that these technical challenges have real-world consequences [12]. Notable examples of reported problems with *medical* machine learning systems include

- a system diagnosing hip fractures mainly based on patient traits and hospital process variables, almost regardless of the patient's X-ray recording [13],
- poor generalization performance of a chest radiograph-based pneumonia detection system from one recording site to others [14],
- surgical skin markings influencing the decision-making of a system designed for melanoma recognition [15],
- disparities in true positive rates of chest X-ray diagnosis systems between patients of different sex, age, race, or insurance types [16], and
- racial bias in an algorithm that computes a health risk score and is used to assign healthcare system resources in the US [17].

Figure 1 shows a schematic overview of some of the many places in the MML workflow in which risks arise, as further discussed in this article. Besides these technical challenges, creating a regulatory environment that does not stifle innovation yet addresses potential sources of harm will be crucial to facilitate a prosperous future for MML [10, 18, 19, 20, 11].

1.1 Regulations and responsibility

Medical products based on machine learning algorithms are already receiving regulatory approval based on current medical device regulations [21] which impose substantial requirements on the development process, the final MML product, and post-market surveillance. Furthermore, broad regulations such as the EU's general data protection regulation (GDPR) or nondiscrimination law equally apply to MML, of course. The exact consequences of these existing regulations for MML applications are far from fully developed. At the same time, the regulatory landscape concerning general AI/ML systems (thus including MML systems) is evolving rapidly, further exacerbating regulatory uncertainty. Dozens of countries and organizations have published general principles for ethical AI [22], and regulatory approaches, as well as technical standards, are currently under heavy development [23, 24, 25, 26, 27, 28, 29, 11]. Due to the preliminary, fast-changing nature of these efforts, it is currently uncertain which standards and regulations companies will have to comply with in order to obtain regulatory approval for MML systems in years to come. It appears to be clear, however, that besides the already existing medical device regulations (MDR, FDA 21 CFR and IEC 62304, to name just a few examples), MML will need to follow general ethical principles for the use of AI such as those put forth by the EU expert commission on AI [30, 31], the OECD [32], or the Chinese ministry of science and technology's national new generation AI governance expert committee. Although there is currently an abundance of proposed guidelines for ethical AI, most of these proposals converge on a small set of core values: beneficence, respect for human autonomy and dignity, prevention of harm, justice and fairness, privacy, transparency and explicability, and responsibility and accountability [33, 34, 35, 36]. Recently, the WHO has released a comprehensive guidance document concerning the governance and ethics of AI for health [11], which also agrees with the above key ethical principles. An MML application that implements these core values to a satisfactory degree can thus be considered subject to low regulatory risks with regard to future regulatory changes.

Responsibility, of course, demands more than just compliance with existing law and regulations. Firstly, regarding those properties that *are* subject to existing regulations, such as, e.g., patient safety and nondiscrimination, it may be perfectly feasible to design a system that successfully receives regulatory approval yet is not as safe or non-discriminatory as it could (and maybe should) be. This discrepancy may be due to the rapidly evolving body of knowledge about the underlying technology and its associated risks (leaving regulatory processes trailing behind), a lack of technical precision in regulations, less than rigorous compliance evaluations, or the influence of lobbying on regulations and standards [37]. Thus, from the perspective of a responsible developer or manufacturer, it may not be morally sufficient to aim for the minimum level of technical and procedural safeguards that

[1] See https://incidentdatabase.ai/.

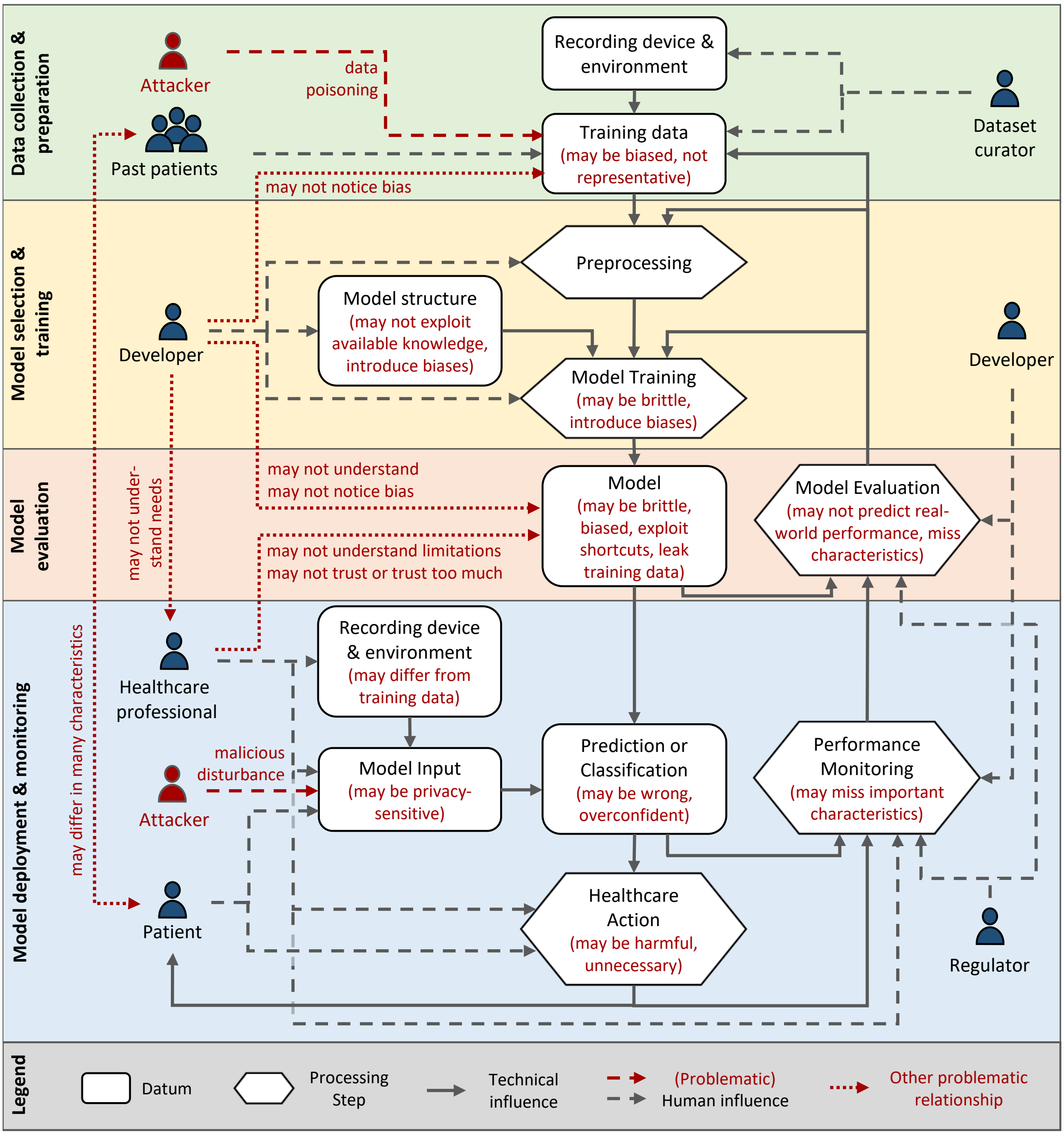
Data collection & preparation
Model selection & training
Model evaluation
Model deployment & monitoring
Legend
Attacker
data poisoning
Recording device & environment
Training data (may be biased, not representative)
Dataset curator
Past patients
may not notice bias
Preprocessing
Model structure (may not exploit available knowledge, introduce biases)
Developer
Model Training (may be brittle, introduce biases)
Developer
may not understand
may not notice bias
may not under-stand needs
Model (may be brittle, biased, exploit shortcuts, leak training data)
Model Evaluation (may not predict real-world performance, miss characteristics)
may not understand limitations
may not trust or trust too much
may differ in many characteristics
Healthcare professional
Recording device & environment (may differ from training data)
malicious disturbance
Attacker
Model Input (may be privacy-sensitive)
Prediction or Classification (may be wrong, overconfident)
Performance Monitoring (may miss important characteristics)
Patient
Healthcare Action (may be harmful, unnecessary)
Regulator
Datum
Processing Step
Technical influence
(Problematic) Human influence
Other problematic relationship

Figure 1: A schematic overview of the medical machine learning workflow throughout the model's life cycle, some of its stakeholders, and some of its risks. The risks and possible solutions to mitigate them are discussed throughout this document.

enables receiving regulatory approval.

Secondly, responsibility encompasses more dimensions than are covered by current regulations, most importantly *social* and *moral* dimensions [38, 11]. Quoting Dignum [39], responsibility is "not just about making rules to govern intelligent machines; it is about the whole socio-technical system in which the system operates, and which encompasses people, machines and institutions." In the context of machine learning for medicine, this includes the wider implications of a new ML-based product for the patient and his or her relatives, the hospital and its stakeholders, the healthcare system, and the society at large. As an example, Elish and Watkins [40] study the impact of deploying "Sepsis Watch", a deep learning-based system for predicting a patient's risk of developing sepsis, on the daily work of hospital nurses. They find that the introduction of the MML system presents new challenges for these nurses and significantly changes their responsibilities and daily work. As a *disruptive* innovation, the new system does not fit into established hospital procedures, hierarchies, and communication rituals, thus requiring a significant amount of "repair work" (mainly by the nurses tasked with implementing the system) to be integrated into the sociotechnical hospital system successfully [40]. Moreover, it challenges physician autonomy (thereby complicating successful communication of the computed risk scores) and demands the integration of a new type of information (an inscrutable risk score computed by the system) into established clinical decision-making processes [40]. Stepping back from this particular example, there are many ethical discussions to be led about whether some particular functionality should be realized using an artificially intelligent system *at all*, or under which circumstances [41, 42, 43, 44, 45, 11]. To address these and similar challenges, various frameworks for algorithmic impact assessments have been proposed [46, 47, 48, 49, 50, 11]. On the other hand, one may also argue that there is a moral responsibility to make improved treatments available, if possible and if there are no opposing moral reasons not to do so. We will *not* address these questions in our survey. Instead, we will assume that a decision has already been made (responsibly) to realize a particular healthcare functionality using a machine learning system.

1.2 A principles–to–practices gap

There is currently a significant gap between high-level ethical and regulatory requirements on the one hand and their practical implementation on the other hand [36, 42, 22]; this has been called the "principles–to–practices gap" [51]. As an example, the FDA's proposed regulatory framework for modifications to AI/ML-based software as a medical device [52] mentions the requirement to follow "good machine learning practices (GMLP)" — which are, however, neither provided nor referenced in the document (a need also discussed by the documents' authors themselves). To the authors' knowledge, the only currently available practical technical guideline that aids engineers in responsibly implementing (general) ML applications is the recent guidance by the German Fraunhofer IAIS institute [53]. The guidance targets developers and auditors of general ML systems and discusses practical measures throughout the lifecycle of an ML application to achieve *trustworthy* AI. The authors distinguish six dimensions of trustworthiness, namely, fairness, autonomy and human control, transparency, reliability, safety and security, and privacy, and list potential risks, risk mitigation measures, and performance indicators for each of these six dimensions. Part of the reason for the relative lack of practical guidance is, of course, the fact that — despite a flurry of research activity in these areas in recent years — the research community is only beginning to propose solutions to many of the relevant technical challenges, with new technical problems surfacing frequently. Moreover, it appears unlikely that there will ever be a simple, universal procedure to successfully implement a safe, robust, reliable, privacy-preserving, transparent, and fair machine learning system for a given application. This is, of course, not to diminish the immense progress made in recent years in areas such as adversarial robustness, explainability, algorithmic fairness, differential privacy, and uncertainty quantification, among many other relevant fields. None of these challenges should be considered solved, however. Additionally, a synthesis of many of these diverse technical advances into practical technical guidance, particularly concerning the unique requirements of medical ML, is lacking. Finally, it has also been noted [42] that most AI white papers are "top-down", i.e., written by governmental institutions and policymakers who have little contact with the actual development and use of AI systems. There is a perceived lack of "bottom-up" initiatives on practical AI ethics, driven, e.g., by hands-on AI developers and the people affected by the AI system [42], although there are notable counter-examples [47]. In this document, we aim to provide such a bottom-up perspective on the technical aspects of responsible and regulatory conform medical machine learning.

1.3 Scope and outline of this document

The aim of this document is to provide practical guidance for researchers and companies developing ML healthcare applications, taking into account the current and likely future regulatory landscape on the one hand and the current state of the art in technical best practices on the other hand. The following section 2 will provide a brief overview of regulations concerning medical ML, as well as the main challenges to overcome to obtain regulatory approval for a medical ML system. From this overview, four key requirements for the realization of responsible

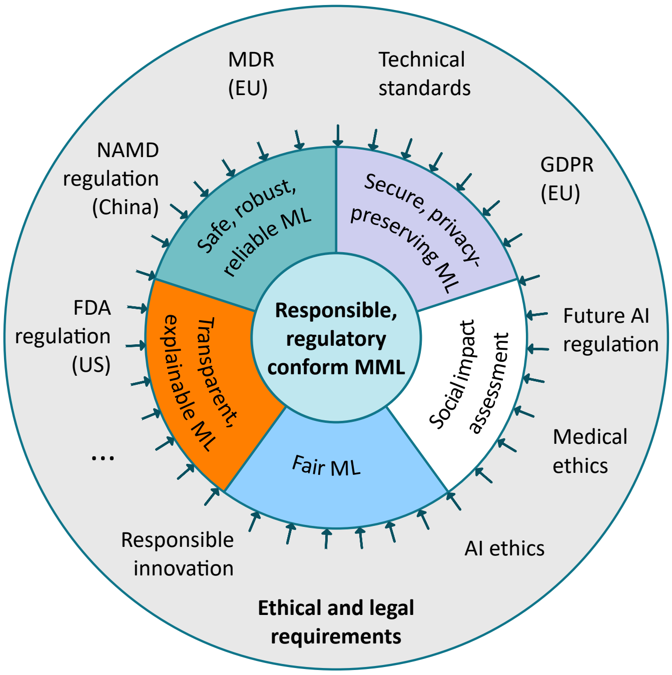

Figure 2: Medical machine learning lies at the intersection of medicine and machine learning and is subject to regulations and ethical considerations regarding both fields. The key requirements that emerge from all of these are safety, robustness, reliability, privacy, security, transparency, explainability, and fairness. (In order to perform *responsible* MML, social impact assessment is another key requirement, which is not addressed in this document.)

and regulatory conform medical ML emerge:

1. safety, robustness & reliability,
2. privacy & security,
3. transparency & explainability, and
4. algorithmic fairness & nondiscrimination.

Figure 2 illustrates the wealth of regulations and ethical fields affecting MML, as well as the aforementioned key requirements for responsible and regulatory conform MML. In section 3 to section 6, we then proceed to discuss the current state of the art regarding solutions to these four challenges, always taking particular account of the peculiarities of medical ML applications. In section 7, we return from the technical details to discuss the high-level themes that emerge, as well as important open problems and future challenges (both technical and regulatory).

Lastly, several notes on the scope of this document. First, we do *not* aim to provide precise technical best-practice recommendations, simply because research on these challenges is still young and evolving quickly. Instead, we provide an expository overview of each technical challenge, why it arises, how it may instantiate in a medical context, and currently promising approaches to its solution. Second, we will not address *dynamical* MML systems here, by which we refer to both continuously learning MML systems as well as ML-based closed-loop control of physiological systems (without any human decision-making in the loop), such as automatic insulin delivery systems [54, 55] and closed-loop neural interfaces [56]. These systems and the technical and regulatory challenges that arise from them will be discussed at length in a future publications of ours. Third, as was already mentioned above, we will *not* address the social implications of developing and deploying an MML system, the question of whether such a system should be built at all, questions of data governance, nor the implementation of inclusive and participatory development processes. A broad overview of these important ethical, societal and governmental challenges, as well as proposed solutions, can be found in recent guidance by the WHO [11]. Here, we will focus on the remaining *technical challenges* and assume that a decision has already been made to realize a particular healthcare functionality using a machine learning system. Thus, the question that remains is this: how to develop such a system responsibly and in a way that ensures regulatory approval?

2 The regulatory landscape

A brief review of laws, standards, and guidelines influencing MML development as of today (cf. fig. 2) shows that there is very little ML-specific regulation as of today. Several jurisdictions have issued regulation of specific ML-enabled technologies such as facial recognition and autonomous driving, but neither general regulation of ML systems nor of medical ML systems [57]. Thus, currently, ML-based medical devices are regulated based on classical, non-ML-specific medical device regulation such as the MDR in the EU and the Code of Federal Regulation Title 21 in the US [21, 20, 58, 59, 60]. Additionally, broad, non-medicine-specific regulations such as the EU's general data protection regulation (GDPR) [61], and anti-discrimination law, of course, apply. In the context of medical device regulations, MML systems are generally interpreted as software (the model and its implementation), of which large parts are automatically generated by other software (the training procedure) in a data-dependent way [52, 62]. As all three components (training data, training procedure, final model implementation) affect the resulting MML system, all of them are subject to regulatory scrutiny. Currently, the path to regulatory approval for medical ML systems strongly depends on the particular interpretations of existing law pursued by the auditing institution, and, as a consequence, evaluations are performed with varying methodologies and rigor [58, 61, 63]. Nevertheless, a considerable number of devices have already received regulatory approval based on these regulations [21, 58]. Thus, the question arises: which requirements on MML systems can be derived from these classical medical device regulations?

Medical device regulations require manufacturers to demonstrate the beneficence, functionality and safety of their products, demanding, among many other requirements, precise specification of an *intended use*, *lifecycle management* including comprehensive *post-market surveillance*, *risk management*, and *verification and validation* [20, 60, 63]. For medical device software, software as a medical device (SaMD) or health software, IEC 62304 (*Medical device software – Software life cycle processes*) and IEC 82304 (*Health software*) specify the most critical requirements and best practices to follow, in addition to the rather general IEC 60601-1 (*Medical electrical equipment - Part 1: General requirements for basic safety and essential performance*) [64, 65, 66, 67]. One essential requirement concerns the deployment of a *quality management system (QMS)* (following, e.g., the ISO 13485 [68] in Europe or FDA 21 CFR Part 820 in the U.S. [69]) as an essential step to ensure that the final product is safe, effective, and efficient [70]. Such a QMS requires the documentation of the whole product lifecycle, including the product planning stage, design and development, verification and validation, as well as deployment, maintenance, and disposal [70]. In addition to the use of a QMS, proper *risk management* (typically following ISO 14971 [71]) is another core requirement that must be pursued throughout the whole lifecycle, and that affects every stage of the development process [72, 60, 64, 71, 63]. A system's intended use must be specified precisely, including, e.g., the medical indication, the treated body part, the target patient population, the intended user (clinical doctor, nurse, patient, patient's family?), and the intended usage environment [71, 73, 74]. Based on the intended use, the product is assigned a *safety class* that determines the level of further safety requirements [75, 20, 69, 74, 71]. (Notably, any system that is intended to monitor physiological processes is assigned risk class IIa or higher according to the MDR, indicating that most MML systems will likely fall into this or a higher risk class [20].) Potential risks — arising during the intended use as well as during "reasonably foreseeable misuse" [74] — must be identified and mitigated, either by reducing their likelihood of occurrence or their severity [72, 71, 73, 74]. Risk mitigation measures may include technical improvements to the system itself, usability engineering measures (such as described by IEC 62366-1 [73]) like mandatory user training and careful UI design, as well as modifying the system's intended use [67, 71, 74, 73]. Finally, extensive *post-market surveillance* activities are also required by the EU's MDR [74, chapter VII], the FDA [76], and various risk management standards [68, 71, 77]. As an example, the EU's MDR demands post-market surveillance by the manufacturer to continuously monitor and report on the safety, usability, and beneficence of the product, identify and report on an unexpectedly increased incidence or severity of faults, analyze any severe incidents, and perform appropriate field safety corrective actions.

Risk management is, thus, a central paradigm of medical device regulation. Interestingly, many of the well-known technical challenges with machine learning systems can be understood as risks within a risk minimization framework: non-robust models that react to spurious input changes, susceptibility to adversarial attacks, a lack of transparency to the users, and biased decision-making all pose risks to the patient, the hospital, and, potentially, the healthcare system as a whole. Accordingly, technical measures to ensure the safety, robustness, reliability, privacy, security, transparency, and fairness of the MML system all represent risk mitigation strategies that are (to some degree, at least) required by already existing regulations [78]. To further substantiate this claim, in the proposed regulatory framework on AI recently put forth by the EU commission [29], two of the four stated objectives of the policy proposal relate to the effective governance and enforcement of "existing law on fundamental rights and safety requirements applicable to AI systems". (The other two objectives concern legal certainty for manufacturers and the development of a single, harmonized market.) In this context, it is also relevant to note that the MDR broadly requires "taking account of the generally acknowledged state of the art" [74] regarding risk control measures and safety principles. In current practice, "state of the art" is interpreted conservatively as "applicable standards and guidance documents, information relating to the medical condition managed with the device and its natural course, benchmark devices, other devices and medical alternatives available to the target population" [79].

Regarding the *verification and validation* of MML systems, the regulations distinguish between the clinical applicability of the product, the validation of the software used in the medical product itself, and the validation of the *tools* used to develop that software. Based on the safety class of the designated product, the final product (including any machine learning model) must pass various stages of clinical validation [75]. Most of the classical machine learning validation activities fall under this category. In addition, the software used to implement the trained model (often a subset of a public ML library) must be tested for its correctness as it is considered a *software of unknown provenance (SOUP)* according to IEC 62304 [62]. What many ML practitioners may be unaware of, however, is that the tools used to develop that software must *also* be validated to obtain regulatory approval — this includes, in particular, any machine learning libraries or frameworks that have been used to label or preprocess training data or to perform the actual model training [62]. This *tool validation* follows a classical software testing scheme: the desired software functionality is specified, and test cases are defined or automatically generated that validate the correctness of the software.[2] In particular, it must be demonstrated that, given the training dataset and the selected model structure and hyperparameters, the training process has correctly identified an "optimal" set of model parameters [62]. This demonstration may involve classical estimation error metrics. Again, notice that this tool validation, and the requirements placed thereupon, differs from a validation of the performance of the trained model: these two elements of the overall validation efforts serve complementary purposes. Finally, note that *transfer learning*, i.e., the use of pre-trained models (often pursued to reduce the necessary amount of domain-specific training data), entails complex requirements on the validation procedure [80].

Since 2018, the EU's general data protection regulation (GDPR) imposes strong additional constraints on data processing systems, including (medical) machine learning applications [81, 78, 82, 61]. (While the GDPR is EU law, it famously has affected companies worldwide [83].) Firstly, the "right to an explanation" has been the subject of intense academic debate [81, 84, 85]. It has spurred a sprint of research in explainability methodology [86] and has been translated into national legislation in several EU countries [87]. GDPR article 13, among other relevant paragraphs, requires a user to be provided with "meaningful information about the logic involved" in automated decision-making. Based on the GDPR, the UK Information Commissioner's Office (ICO) has recently issued comprehensive guidance on explaining decisions made by AI systems [88]. While the exact extent to which the GDPR mandates a "right to an explanation" is currently unclear, it appears evident that black-box AI without any further explanation is *not* GDPR-compliant. Secondly, GDPR article 22(3) grants a (closely related) "right to contest the decision", which opens up questions regarding the interplay of contestability, model explainability, and transparency [82]. Thirdly, GDPR article 22(4) grants a "right to nondiscrimination" with regards to automated data processing and profiling, explicitly forbidding discriminatory use of sensitive attributes such as, among others, racial or ethnic origin, religion, and political opinion [81, 89, 78]. Again — the exact requirements this places on machine learning applications are currently unclear [81], but completely ignoring bias and discrimination issues is undoubtedly not GDPR-compliant [78]. Fourthly, GDPR article 35(7) requires a Data Protection Impact Assessment (DPIA), which should contain "1. a systematic description of the envisaged processing operations and the purposes of the processing, [...]; 2. an assessment of the necessity and proportionality of the processing operations in relation to the purposes; 3. an assessment of the risks to the rights and freedoms of data subjects [...]; and 4. the measures envisaged to address the risks." It has been argued that such a DPIA in the sense of the GDPR can serve as a form of algorithmic impact assessment, although it lacks essential elements such as mandatory public disclosure of the DPIA and explicit mandatory stakeholder involvement [50]. Finally, another rather obvious consequence of the GDPR concerns the maintained privacy of the training data, which may necessitate the use of privacy-preserving ML techniques as more and more information extraction attacks become known in the literature [90, 91, 92, 93, 78].

Based on their previous white paper [94, 95] and the report of the high-level expert group on AI [30], in April 2021 the EU commission has published the first concrete proposal on regulating AI-based applications [29]. The proposed regulation is derived from four objectives: to ensure that AI systems are safe and respect existing law on fundamental rights and values (including, e.g., nondiscrimination law), to ensure legal certainty, to enhance the effective enforcement of already-existing law on fundamental rights and safety requirements, and to prevent market fragmentation due to opposing regulations. Following the proposal, applications would be classified as falling into different risk categories, with most medical applications being classified as "high-risk AI systems". A high-risk classification would entail a list of legal requirements "in relation to data and data governance, documentation and recording keeping [sic], transparency and provision of information to users, human oversight, robustness, accuracy and security" [29]. Notable requirements placed on high-risk applications include

- the use of a risk management system that "shall consist of a continuous iterative process run throughout the entire lifecycle" (article 9) as well as an appropriate quality management system (article 17),
- quality criteria regarding the training data as well as their governance and management processes (article 10),
- that "their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" (article 13), and
- that they achieve "an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle" (article 15).

To retain flexibility, the precise technical solutions to achieve compliance with these requirements are left to technical standards (to be developed) and other guidelines. Outside the EU, essentially all larger regulatory bodies have issued white papers indicating their inclination to set up AI/ML-specific regulation in the near

[2] Importantly, only those functions of a tool or library that are used during the development process must be validated.

future [57]. These include the FDA's "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan" [23], emphasizing a need for improved methods to evaluate and address algorithmic bias and a need for transparency regarding the training data, the model structure, and evidence of the model's performance. Similarly, the Chinese State Council has proposed a "New Generation Artificial Intelligence Development Plan" (AIDP) [96]. Aside from these coming hard law changes, many technical standards and guidelines on subjects such as ML robustness, transparency, and bias mitigation are currently under development and will be available soon. Notable examples include the work of the ISO/IEC JTC 1/SC 42 [27, 26], the IEEE P7000 family of standards [25] and the P2801 and P2802 standards [97, 98], and the UK information commissioner's office (ICO) guidance on auditing AI algorithms [78] and explaining decisions made by AI systems [88].

To summarize, current regulation requires comprehensive risk assessment and mitigation, quality management, data privacy, nondiscrimination, and some degree of transparency about the reasoning behind an automated decision, for any medical product. Within the context of medical ML, the pursuit of properties such as safety, robustness, privacy, security, transparency, and freedom from discrimination against patient groups can thus be framed as *risk mitigation strategies* [78, 94, 99]. These requirements are legally binding already *today*, with future standards and regulatory changes only aiming to further clarify and substantiate them [78, 29, 63]. Considering these requirements during development today, although their current evaluation by regulatory bodies appears non-comprehensive, thus seems advisable not only from an ethical responsibility perspective but also from a legal point of view. Moreover, it aids in building expertise in these crucial topics that will continue to rise in importance swiftly. To this end, the following sections of this document will provide brief overviews of the technical challenges and possible solutions to achieve safe, robust, reliable, secure, privacy-preserving, transparent, and fair medical ML.

3 Safety, robustness & reliability

Medical technology directly or indirectly affects patient health and therefore must adhere to rigorous safety requirements. The classical approach to developing a safe medical device is to follow a *risk-based approach*, cf. section 2—this requires the identification of possible faults and the quantification of their likelihood of occurrence, as well as the risks that may arise from them [72]. If unacceptable risks are identified, actions must be taken to prevent and control (mitigate) the faults causing these risks [72]. For classical medical hardware devices, developers must consider safety measures such as radiation protection, fire protection, or protection against mechanical hazards. The safety aspect that is most significantly affected by the use of machine learning in a medical product, however, is *functional safety*.

A functionally safe system reliably performs its intended function: if faults with a hazardous effect occur, a functionally safe system is able to recover a safe state nevertheless [100].[3] To achieve functional safety, a system must thus i) prevent any faults in advance or, where this is infeasible, ii) reliably detect, and iii) control them. Two types of faults are usually differentiated [100]: systematic faults and random faults. A *systematic fault* is defined as a "failure, related in a deterministic way to a certain cause, which can only be eliminated by a modification of the design or of the manufacturing process, operational procedures, documentation or other relevant factors" [100]. Systematic faults may arise from construction errors, software bugs, or human operator mistakes. A *random fault*, on the other hand, is defined as "a failure occurring at a random time, which results from one or more degradation mechanisms" [100]. In classical hardware components, these are often a consequence of aging processes that unavoidably lead to hardware failures. Their rate of occurrence can be quantified and judged regarding its acceptability. The IEC 61508 [100] generally assumes that software faults are systematic and not random, which would signify that faults of an MML system are all to be considered systematic. However, the standard is currently not explicit in its characterization of machine learning systems and their failures, and neither are other standards concerning functional safety [101].[4] Concerning medical ML systems, problems such as biased datasets and fragile models represent examples of systematic errors: they result from an imperfect model design process, akin to classical software bugs. Transparency and explainability, discussed in depth in section 5, serve to reduce the likelihood of systematic errors occurring by enabling the developer to inspect and analyze the system in detail. Once the system is deployed, transparency and explainability *to the user* facilitate the early detection of faults of an MML system by a human operator. Section 6 is concerned exclusively with methods to prevent systematic biases occurring in the learned model.

In the following discussion of barriers and measures to achieve functional safety of MML systems, we distinguish between the three closely related properties of *robustness*, *reliability* and *safety*, for which we employ the definitions also used by Borg et al. [102]:

Robustness denotes "the degree to which a component can function correctly in the presence of invalid inputs or stressful environmental conditions" [103].

[3] The following discussion of functional safety is based on IEC 61508:2010 [100].

[4] Standardization efforts are currently underway, e.g., at the ISO/IEC JTC 1/SC 42: `https://www.iso.org/standard/81283.html`.

Reliability denotes "the probability that a component performs its required functions for a desired period of time without failure in specified environments with a desired confidence" [104].

Safety denotes "freedom from unacceptable risk of physical injury or of damage to the health of people, either directly or indirectly, as a result of damage to property or to the environment" [100].

We (informally) call a machine learning method *robust* if it yields similar models when applied to training data drawn from similar distributions.[5] Consider, for example, an ML method producing an artificial neural network (ANN) to detect lung cancer based on X-ray images. Now suppose that two training sets of X-ray images of human lungs have been recorded from the same subjects with X-ray scanners from two different manufacturers A and B, with the scanner of manufacturer A yielding noisier images than the scanner of manufacturer B. We call the ML method *robust* if the two classifiers resulting from training on the two datasets are similar. A closely related yet subtly different notion is the one of robust ML *models* (as opposed to robust ML *methods*, comprising the whole ML tool-chain which yields an ML model). In accordance with various formal definitions of (local) robustness proposed in the literature [106], we informally denote a machine learning model to be robust if it provides similar predictions for similar inputs. Returning to the above example, we would call a classifier robust if it classified images taken from the same patient similarly, regardless of the scanner used for the recording. It is well known (and intuitively makes sense) that the robustness of a model against variations along a certain dimension is closely linked with the model's response surface being *smooth* along that dimension [107, 106]. Thus, robustness-enforcing methods can be understood as biasing the model towards smoothness along certain dimensions. Notice, however, that the maximally robust model is the one that returns a constant output regardless of the input — thus, it immediately becomes apparent that robustness on its own cannot serve as an optimization target. Robustness can therefore only be considered beneficial insofar as it serves to increase the *reliability* of the model, i.e., its tendency to perform well for most of the target data [108].

The following section discusses the most critical barriers to achieving safety, robustness, and reliability in medical machine learning systems. Section 3.2 then describes technical measures to increase model robustness and reliability, which can both be seen as necessary requirements to *prevent* the occurrence of failures. Finally, section 3.3 discusses methods for fault *detection* and *mitigation*.

3.1 Obstacles to achieving safety, robustness, and reliability

Possibly the most salient challenge to overcome for achieving safe, robust, and reliable MML concerns the lack of large, representative, high-quality datasets. The high dimensionality of the input space in most application scenarios means that it is usually impossible to sample the input space sufficiently densely without relying on vast amounts of training data. Moreover, training data collection may be limited to specific countries, hospitals, medical devices, participant acquisition channels, or further circumstances that may constrain the types of data observable *in principle* using a specific data collection process. Consequently, models are often used to make "out of distribution" (o.o.d.) predictions, i.e., extrapolate to previously unseen regions of the data space. Unfortunately, existing medical databases are typically confined to small, inaccessible data *silos* [109], and riddled with labeling errors [110, 111, 112] and biases [113, 114, 115, 116, 117, 118]. There are various valid reasons for this, including legal constraints,[6] concerns over patient privacy, the low prevalence of rare diseases, and the high cost of performing studies [119]. Data labeling, segmentation, and annotation are major sources of errors and require extensive effort, often by medical experts [109]. Inter-observer variability, the inherent uncertainty in the medical domain, and plain labeling errors all contribute to label noise [111]. Moreover, since an exact and accurate ground truth label is often unavailable, data labeling often depends on interpretations of feature relevance and labeling styles [109]. For all of these reasons, the gathering of large, high-quality, representative datasets is widely recognized as a decisive challenge for the successful application of machine learning in the medical context [119, 6, 1, 120, 121].

On a more conceptual level, the traditional statistical learning framework assumes that the data used for training the algorithm and the real-world target data to which the classifier will ultimately be applied are drawn from the same data generating distribution. This assumption is rarely met in practice [122, 123]: usually, data observed "in the wild" (the *target domain*) differ in many characteristics from those in the training data[7]; a phenomenon called *distribution shift* or *dataset shift* [122, 123]. Distribution shift may occur for many different reasons, including sampling biases and different instrumental or environmental noise in the target data and the training data. In the context of medical image analysis, potential causes include differing viewing angles, movement of the cap-

[5] This informal definition is in the spirit of classical robust statistical estimation theory, see, e.g., Daszykowski et al. [105].

[6] Important regulations in this regard include the general data privacy regulation (GDPR) in Europe and the health insurance portability and accountability act (HIPAA) in the US.

[7] By *training data*, we refer to the union of the training, validation, and test datasets. If it is necessary to refer to the first of those three exclusively, we will refer to it as the training data*set*. (But this distinction will rarely be necessary.)

tured objects, lighting conditions, or different optical sensors [124]. In particular, it is a well-known challenge to train ML models that work across different scanners or imaging protocols [125]. A robust ML model would overcome such rather superficial distribution shifts and would perform as expected. Another type of distribution shift may arise when the training data are *inherently* different from the target data. Using images of parts of dead bodies to train a model to be applied to living patients is one example from the medical domain; the use of electronic health records gathered from young, healthy persons to train a model that will (also) be applied to sick and older patients is another. In these cases, again, the essential information required for solving the task is assumed to be present in the training data nevertheless, but a non-robust machine learning model may fail to distinguish relevant and irrelevant features and may thus perform poorly on the target data. Unfortunately, modern deep learning methods such as those used in medical imaging analyses are particularly brittle with respect to even minimal dataset shifts [126, 127, 128, 129] — in other words, they are not robust. In the medical domain, dataset shift has thus been identified as a critical challenge preventing the widespread application of MML methods because, e.g., trained models are not robust against differing recording environments or devices, patient demographics, hospital types, healthcare systems, or healthcare policy shifts [125, 130, 14, 131, 132, 133]. Notice, however, that not all dataset shifts necessarily pose a problem, and some may even be beneficial: intentionally over-representing minority groups in the training data is one example that will be discussed in detail further below.

Spurious correlations represent a related yet different challenge: it is well known that irrelevant features often correlate with the prediction label [134]. These spurious correlations are usually not a problem when using a white-box model structure that exploits the available prior knowledge about relevant and irrelevant factors. Highly flexible deep learning models, however, incorporate very little — if any — prior knowledge, and are thus prone to exploit these spurious correlations for improving their predictive accuracy on the training data; a phenomenon often called *shortcut learning* [134]. Examples of confounding factors that have been found to (inadvertently) strongly influence model predictions include patient and hospital process data [13], surgical skin markings [15], hospital system and department within a hospital [14] and various other factors including age and gender [135]. Recently, Zhang et al. [136] have shown that adversarial examples can be understood as intentional exploits of spurious correlations learned by the model.

Both their susceptibility to spurious correlations and their non-robustness to distribution shift are consequences of modern deep neural network (DNN) model's vast amount of parameters and neglect of any prior knowledge. These features grant them the expressiveness to capture every statistical peculiarity present in the training data. Moreover, modern DNNs are sufficiently expressive for many different models to achieve very similar predictive performance on data drawn from the same distribution as the training data; a situation called *model underspecification* [137]. D'Amour et al. [137] show that this situation is omnipresent in modern (deep) ML pipelines and that it results in highly variable real-world performance: a subset of the large class of models achieving similar performance on the training distribution may perform well on (inevitably slightly different) real-world data [134], but whether a model from this subset is actually learned — and not one that transfers badly to the real world — is more or less random [137]. Thus, in addition to the trained models, the *training process* is typically also non-robust: the resulting models may differ widely if supposedly unimportant details such as the random seed, the initialization method, or the particular implementation of the algorithm are changed [137]. Figure 3 illustrates the relationships between dataset shift, shortcut learning, and model underspecification.

3.2 Measures to achieve robustness & reliability

In the following, we will discuss various technical measures for achieving more robust and reliable MML systems, proceeding stage by stage through the lifecycle of an MML application.

Problem specification At the beginning of the problem specification phase, the target distribution should be carefully described. This distribution depends upon many factors, including the clinical use case, the patient population, target measurement devices, and recording environments. This first step is crucial for selecting appropriate robustification strategies against expected distribution shifts and for later identifying an expected or unexpected distribution shift during deployment. If it is difficult to obtain a model that is robust and reliable across the originally intended target domain, it may be necessary to restrict the target domain. This may mean limiting the application to be used within a particular patient group,[8] only with a small, specific set of sensors, or with a certain camera angle and lighting, to give some examples. Moreover, it needs to be determined whether the introduction of adversarial examples by malicious attackers is of concern; refer to the following section 4 for a discussion of adversarial robustness. Besides its technical necessity, a precise specification of the MML system's *intended use* is demanded by existing regulations.

[8] Of course, applicability to a large and diverse group of patients should always be the goal. Developing products only for certain patient groups can be seen as unfairly discriminating against other groups, depending on the circumstances and reasons for doing so.

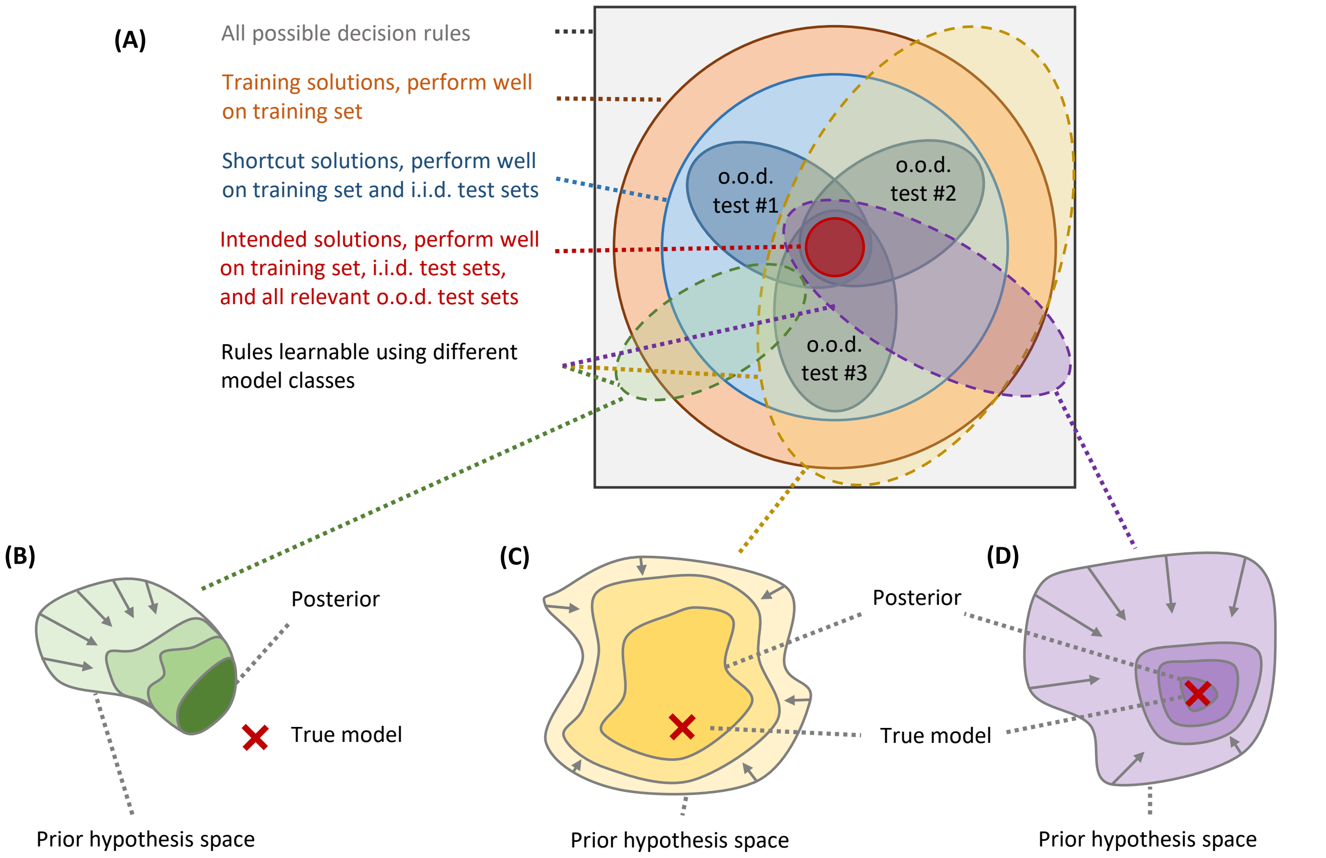
(A)
All possible decision rules
Training solutions, perform well on training set
Shortcut solutions, perform well on training set and i.i.d. test sets
Intended solutions, perform well on training set, i.i.d. test sets, and all relevant o.o.d. test sets
Rules learnable using different model classes
o.o.d. test #1
o.o.d. test #2
o.o.d. test #3
(B)
Posterior
True model
Prior hypothesis space
(C)
Posterior
True model
Prior hypothesis space
(D)
Prior hypothesis space

Figure 3: **(A)** Models that perform well on the training dataset need not perform well on the test dataset, and those performing well on the test dataset still need not perform well in the real world, for reasons including dataset shift, spurious correlations, and model underspecification. Figure closely modeled after Geirhos et al. [134]. Independent identically distributed (i.i.d.), out of distribution (o.o.d.). **(B)** A model class with a too restrictive hypothesis space prevents convergence towards the true model. **(C)** A model class with inadequate inductive biases renders convergence towards the true model inefficient, requiring lots of data. **(D)** A model class must cover a sufficiently large hypothesis space that includes the true model, it must be equipped with adequate inductive biases *and* there must be sufficient data to enable convergence towards the true model. Figures (B)–(D) closely modeled after Wilson and Izmailov [138].

Data collection and preparation It is well recognized in the ML community that the availability of a large, representative and diverse dataset is an essential ingredient to train a robust and reliable ML model [139, 140]. As a first and foremost requirement, a dataset should contain sufficiently many examples from each relevant target group. In practice, this is usually not realistic to achieve due to the high dimensionality of the input space in most application scenarios, as was discussed in section 3.1. Thus, as out of distribution (o.o.d.) prediction is generally unavoidable, testing the model's performance in such o.o.d. prediction settings (in other words, its generalization capability) is crucial. One strategy to assess a model's generalization capabilities is to intentionally introduce o.o.d. examples in the *test* dataset, i.e., examples that differ in some characteristic from the examples included in the training dataset (cf. fig. 3). In practice, training, validation, and test datasets are often carefully (and iteratively) crafted to include a sufficient amount of examples from all regions of the target data distribution [139]. To achieve this, it is often useful to significantly over-represent minority groups, e.g., patients with rare diseases or ethnic minorities, in the training data (as compared to the target data distribution), because otherwise such groups would only be sparsely represented, and the resulting model might perform poorly on such rare groups [141, 122]. In this sense, it may be helpful to use a training distribution differing widely from the target distribution: as an example, it is often beneficial to employ a uniform training distribution in instances where the target distribution is strongly imbalanced [141, 142, 143]. Such a reweighting then corresponds to over-representing minority groups and under-representing majority groups in the training dataset. Dataset weighting methods such as propensity score weighting [144] or importance weighting [145] are often employed to implement such a rebalancing. Aside from dataset balancing strategies, the intentional introduction of training examples that contradict potential spurious correlations represents another valuable dataset curation strategy that may help prevent shortcut learning [134]. Notice that such dataset curation strategies actively induce data shifts — which are, however, hoped to exert a beneficial effect on the training process and outcome — and also compare section 6 for a discussion of the effect of dataset curation on the fairness of the resulting model. In this regard, it is also worth mentioning the "All of Us" initiative, which "plans to enroll a diverse group of at least 1 million persons in the United States" [146], with most of the participants stemming from groups that were previously underrepresented in biomedical datasets. Similarly, the proposed "European health data space" may represent an essential step in the right direction if executed well.[9]

To alleviate the (difficult to fulfill) need for gathering large-scale, application-specific, centralized databases containing sensitive patient health data, distributed learning methods such as *federated* learning [147, 148] and *split* learning [149] have been proposed. These methods allow training a model using data from different sites without requiring them to share their sensitive data with any other site or a central entity. This is achieved by running partial model update steps locally at each site and then sharing the result of these computations (but not the data they are based on) with other sites or a central entity. Federated learning and split learning are closely related and differ mainly in the way the partial model computations are split across the participants. Each method has been found to be preferable performance-wise under different circumstances [150], and hybrid methods have been proposed to combine the benefits of both [151, 152]. Gao et al. [150] provide a comprehensive performance evaluation of both approaches for varying types of datasets, models, and (computing or communication) performance constraints. Due to the particular challenges related to the gathering of large medical datasets, distributed learning has been prominently put forth as an essential ingredient for solving the data silo problem in medical machine learning [91, 92, 93, 153, 154]. Various distributed learning toolboxes have been developed [155, 156, 157], and the IEEE has recently published a guidance document on the development of federated learning frameworks and applications [158]. Owing to their distributed nature, however, these methods raise new challenges regarding privacy and security, which will be discussed in detail in section 4.

Semi-supervised learning represents an alternative avenue for reducing the effort necessary to create large datasets by leveraging unlabeled data in addition to a smaller set of labeled data [159]. Gathering sufficiently large labeled datasets requires an amount of manual labeling effort by medical experts that may simply not be admissible in many cases. As an example, Bai et al. [160] propose a semi-supervised learning method for cardiac MR image segmentation. Semi-supervised learning methods depend on some basic assumption about the underlying data distribution, such as smoothness of the response surface (similar input data have similar labels), that decision boundaries pass through low-density regions of the data space, or that data sharing a label lie on low-dimensional manifolds [159]. The utility of using additional *unlabeled* images in the context of medical imaging has recently been questioned [133] because additional unlabeled examples can only provide further information about the input data distribution but *not* about the mapping between input data and their target labels. A slightly different approach to solving the labeling problem is *automatic labeling*. These methods typically exploit more domain-specific knowledge than general semi-supervised learning methods. Trivedi et al. [161] have proposed a machine learning model for automatically extracting mammogram labels from their accompanying free-text clinical records. The authors found their method to perform well in the case of short text records. Similar methods have been proposed for extracting information on pulmonary emboli from free-text radiological clinical records [162]. In a slightly different vein, Yi et al. [163] trained a deep neural network to label chest mammograms semantically with their mammographic view and breast laterality.

Transfer learning in general and *domain adaptation* (which is a sub-field of transfer learning) in particular represent other alternative approaches for increasing the amount and representativeness of the available training data. The fundamental idea is to employ available data from different (but in some sense similar) domains for improving model performance. There are various approaches within this field, from using models pretrained on general-purpose labeled image databases and specializing them to a medical image recognition problem [164] to using data recorded from different patient groups or using different recording devices.[10] One instance of the latter approach is called *domain representation learning*; its aim is to identify a set of feature transformations that accurately represents data recorded from different domains, e.g., histological images from different labs [165].[11] The actual classification or regression model is then trained on the data in this identified domain representation, thus making it applicable to data from many different domains and enabling the use of data from superficially different domains in the training phase. Notice that standard preprocessing techniques like normalization and motion correction [167] can be understood as

[9] See https://ec.europa.eu/health/ehealth/dataspace_en.

[10] As was mentioned briefly in section 2, there are complex regulatory questions associated with the use of transfer learning [80].

[11] Specific representation learning methods have also been advocated as tools to increase the interpretability [166] or fairness of the resulting MML model, cf. section 6 for details.

enforcing a particular kind of domain representation. For recent reviews and applications to biomedical imaging, refer to, e.g., [168, 169, 170].

If a sufficiently large and representative dataset is unavailable, (partially) *synthetic data* may be used to augment or completely replace real data [171]. Many methods fall into this broad category, from image data augmentation [172, 133] and synthetic resampling methods [143, 173, 143, 142, 174] to (patho)physiological simulation models [175, 176] and GAN-based synthetic medical data generation methods [177, 178]. As an example from general-purpose image recognition, Geirhos et al. [179] demonstrate that training a model on a *stylized* version of the ImageNet database, which equips objects with unusual textures, biases the trained model towards recognizing shapes instead of textures, thereby significantly increasing robustness. When using synthetic data, a crucial question regards the proper choice of the model parameters for generating a synthetic dataset that optimally supports learning a robust and reliable model. *Domain randomization* is one approach that has been proposed for solving this problem by incentivizing successful concept learning [180, 181]. However, many questions remain unanswered: what are the requirements on the simulator and the parameter selection method for training a robust and reliable model using synthetic data? Should one use only synthetic data or mixed datasets? Moreover, just like above: how does one synthesize a *balanced* dataset? Recent evidence against the utility of synthetic dataset augmentation has been provided by Taori et al. [129], who found that adding various types of synthetic distribution shift — e.g., artificial noise or image rotations — did *not* increase the resulting model's robustness against naturally occurring distribution shift. This observation is in contrast to the results of Michaelis et al. [182], who found simple training image stylizing [183] to significantly increase model robustness. Many groups are currently investigating the use of synthetic data for machine learning; thus, it appears likely that answers will emerge soon. For a recent review regarding the use of synthetic data in deep learning, refer to Nikolenko [171].

A final aspect regarding the employed datasets concerns their preprocessing. Application-specific preprocessing steps that exploit relevant domain knowledge may significantly simplify the subsequent estimation problem. As one recent example, Li et al. [184] demonstrate improved chest X-ray classification accuracy when using a specific preprocessing step that suppresses bones in the recorded images. Concerning the handling of errors in the employed datasets, one needs to distinguish between different types of errors. If *measurement outliers* frequently occur in the target data, then they should also be represented in the training data — in this way, a prediction system (potentially including a preprocessing step that discards such outliers) can be designed that is robust to these outliers. *Labeling errors* (or, in the case of regression, outliers in the target signal), on the other hand, should be avoided in the training data, wherever possible. Note, however, that a *robust* learning method — as discussed above — should yield a similar model even in the presence of a few labeling errors. It has been claimed in the literature that deep learning models are inherently robust to label noise in the training dataset [185]. In light of the recent analyses of D'Amour et al. [137], which question the validity of test set performance (for underspecified models) as an indicator of real-world performance, it is unclear, however, whether these results translate into practice. In a similar vein, Northcutt et al. [186] recently reported that labeling errors in the *test* dataset (as opposed to the *training* dataset, as analyzed by, e.g., Rolnick et al. [185]) crucially affect real-world performance because they influence model selection. The obvious countermeasure against label errors in training and test datasets is an increased investment into the collection of high-quality data. More generally, it is crucial to perform an in-depth data quality assessment, for which various data quality indicators have been proposed and (health data specific) toolboxes are available, cf., e.g., Schmidt et al. [187]. Efforts for the standardization of data quality assessments are underway, both medicine-specific [97] and concerning general ML [188].

Model selection & training The canonical method to achieve robustness of the training process is to take some kind of prior knowledge about the class of reasonable models into account. This will bias the training process towards models that concur with these prior assumptions. In the simplest case, one can incorporate prior knowledge by assuming a simple, parametric model structure. While the use of simple model structures may, in some cases, increase model robustness, this may come at the expense of overall prediction performance. However, prior knowledge can be injected into the estimation process in a myriad of different ways, including choosing a specific (grey box) model structure, the specification of Bayesian priors over various model parameters, or imposing hard constraints on model parameters and properties.[12] One prominent line of research in this field concerns the training of models (including, in particular, deep neural networks) that satisfy some prediction monotonicity constraint with respect to one or multiple input characteristics [189, 190]. Such a monotonicity constraint, which represents a reasonable yet very mild assumption in many application scenarios, may already significantly improve model robustness. Similarly, network structures have been proposed that exploit symmetries and invariance properties, such as invariances with respect to image rotation, translation, and reflection [191], and Evans and Grefen-

[12] Such hard constraints can, of course, also be formulated as probabilistic priors.

stette [192] propose a method to combine the flexibility of neural networks with the data efficiency of inductive logic programming (ILP). In another branch of research, Chen et al. [193], Barnett et al. [194] have proposed slightly modified deep learning architectures that utilize a special *prototype layer* which characterizes the similarity of the input data with prototypical training dataset examples. While this modification was mostly introduced to increase the interpretability of the model (cf. section 5), this also represents an inductive bias that may reduce the likelihood of learning spurious correlations during training. Importantly, as one would expect, recent research demonstrates that the incorporation of such mild constraints on the learned models (if they are well-justified) does *not* diminish predictive model performance [190, 193, 194]. Such constraints may, on the contrary, help alleviate the underspecification problem by constraining the feasible model space [137].

Another branch of research investigates causality-aware models that impose constraints on the relationships between causes and effects, and thereby increasing robustness to dataset bias, distribution shift, spurious correlations, and adversarial examples [195, 196, 197, 131, 198, 136]. It has been argued that classical ML models' inherent unawareness of the possible interactions of causes and effects imposes a fundamental limitation to the level of performance they can achieve [196, 198]. In the medical domain, Richens et al. [199] have recently shown causal inference to significantly outperform purely associative inference, as performed by usual ML methods, on a differential diagnosis task. Subbaswamy and Saria [131] have discussed the merits of using causal models in healthcare for achieving robustness to distribution shift. Similarly, Castro et al. [133] have argued that a causal perspective is crucial for machine learning for medical imaging and have provided a framework for doing so. Moreover, Holzinger et al. [200] have argued that causality is a necessary prerequisite to achieve real explainability, which will be the subject of section 5. Section 6 will discuss the beneficial properties of causal models for achieving algorithmic fairness.

Finally, notice that classical regularization techniques such as weight decay or smoothness regularization also increase model robustness by incorporating some prior knowledge about the system [201], thereby reducing the chance of overfitting the available data.

Model evaluation Just like a large and representative *training* dataset is crucial for training a robust and reliable model, a large and representative *test* dataset is crucial to enable an accurate assessment of the trained model's performance in the real world. Many considerations regarding the generation of good datasets have already been discussed in detail above. Here, it shall suffice to emphasize a crucial difference between MML applications and many other ML applications. In many practical ML applications, it is current practice to iteratively refine the test and training datasets based on field failures: it is observed that the current model performs poorly on some groups; thus, more examples of that group are added to the training and test datasets. In this way, the representativeness of these datasets is iteratively improved long after the initial deployment. While this may be a good practice in a low-stakes environment such as online advertising, it is, of course, limited in its applicability to medical ML: in the healthcare setting (as in many other physical settings), developers cannot afford to deploy poorly working prototypes and learn from their mistakes; the first deployed version of an MML system must *already* perform safely, robustly, and reliably [202]. Thus, utmost care must be taken to ensure the representativeness of — in particular — the initial test dataset.

In addition to data-based validation, various formal verification methods have been proposed for machine learning models, akin to classical software verification techniques. One benefit of formal verification methods is that they can enable assessing compliance with lawful requirements (e.g., nondiscrimination), as may be necessary to enable users to *contest* a decision made by the system [82]. While computational complexity is high, novel efficient relaxation techniques have been proposed to enable the application of Satisfiability Modulo Theory (SMT) solvers to deep learning models [203, 204]. These solvers can be employed to formally verify various properties of the input-output behavior of a learned model. Katz et al. [204] employed their proposed techniques to verify application-specific desirable properties of a real-world airborne collision avoidance system for unmanned aircraft. In the medical domain, Guidotti et al. [205] have employed an SMT-based technique for formally verifying a desirable input-output property of a prosthesis myocontroller. Because such SMT-based techniques are computationally very demanding and not applicable to arbitrarily large networks, Pei et al. [206] have proposed an alternative method (VeriVis) using input space reduction techniques. They utilize this method to verify multiple practically relevant safety properties of the input-output behavior of (then) state-of-the-art computer vision systems, thereby discovering thousands of violations of these safety constraints by various competitive academic and commercially available vision systems. These detected safety violations can then of course be incorporated into the training dataset, thereby hopefully increasing the robustness of the learned model. Michaelis et al. [182] discuss a range of additional tools that serve similar verification purposes for deep learning models. Of course, similar formal verification techniques have also been proposed for other classes of machine learning models, including, in particular, various kinds of tree ensembles [207, 208, 209]. Several of these formal verification methods can also be used to iteratively "repair" a model that does not (yet) satisfy the specified constraints until it does [205]. Finally, there is a large body of work regarding the formal verification of (e.g. ℓ_∞) adversarial robustness, which — as mentioned above — is closely related to the model's smoothness [204, 210, 211, 212]. In particular, the field of *semantic adversarial deep learning* is concerned with verifying the robustness of the learned model to semantic variations, i.e., variations that are of application-specific interest as opposed to purely synthetic noise [213]. Similar to the verification techniques discussed above, such semantic adversarial examples that have been found to be violated by a trained model can then be included in the training dataset to achieve model robustness towards this type of variation.

As a final remark on model evaluation, interpretable models are obviously easier to evaluate thoroughly than black-box algorithms since the developer and other stakeholders can (manually or in an automated fashion) assess the model's reasoning behind a particular decision, thereby increasing the likelihood of detecting erroneous behavior before model deployment [53, 214]. This represents one of the principal arguments for employing interpretable models, which will be the subject of section 5.

Model deployment and monitoring During the final stage of an MML model's lifecycle, continuous monitoring for possible distribution shift is a permanent and essential task. Is the model being employed in an environment it was not trained for, using sensors that have not been part of the training data, or on unforeseen patient groups? Besides gathering general statistics about the target data distribution, one may want to regularly draw random samples from the target distribution and re-evaluate the model accordingly. Aslansefat et al. [215] have recently proposed a formal method for monitoring and quantifying distribution shift, potentially serving to recognize when a model is used outside of its safe application area. Crucially, feedback from the end-users — patients or healthcare workers — regarding the performance of the model and its potential failures should also be gathered and incorporated continuously. Many, if not all, of the above-mentioned measures are required by current regulations regarding post-market surveillance activities for medical devices (cf. the discussion in section 2).

3.3 Measures to achieve functional safety

Model robustness and reliability are necessary but certainly not sufficient requirements to achieve an MML application's functional safety. These two properties both serve to *prevent* the occurrence of failures. There are, however, more strategies that can be used to this end. For one, *transparency* regarding safe operating conditions is crucial: Under which circumstances and on which dataset was the model trained? Which are the assumed conditions on the operating environment and the patient characteristics? These requirements must be clearly stated and communicated to the user, and deviations from these required conditions should be detected automatically where possible, requiring manual human decision-making in these cases. Similarly, automated rejection techniques which aim to recognize situations in which the model's predictions are unreliable because it is performing "out-of-distribution prediction" can be used [216, 217, 218, 219]. Besides being transparent about the required operating environment, another important systematic risk mitigation strategy is to *ensure* that these operating conditions are met by means of *procedural safeguards* [220] such as, e.g., installing a sensor at a fixed position, designing a system that incorporates an ideal lighting solution, only allowing sensors by particular manufacturers, or designing a user interface that requires the doctor to check that the patient belongs to a supported patient group. On an organizational level, akin to classical software security measures, *red teaming*[13] and *bias and safety bounties* have recently been proposed as effective methods for increasing the safety of ML systems [221]. Finally, proper operator training is essential [222, 223, 23]: users — be it patients or healthcare workers — must be educated about the capabilities and limitations of a machine learning system in order to prevent incorrect and potentially unsafe usage. Safety measures such as the ones mentioned in this paragraph fall under the umbrella term of *usability engineering* (as described, e.g., by IEC 62366-1 [73]) and are an essential requirement for regulatory conform MML.

If an error occurs despite all countermeasures, how can its negative impact be mitigated? Firstly, the error should be *detected*. Automated rejection techniques, which detect inputs for which the model is particularly uncertain, or for which only a few similar examples exist in the training data, have already been mentioned above. On the human side, again, interpretability of the model (cf. section 5) is highly beneficial: if the user can understand the model's reasoning behind a decision, the assumptions that went into it, as well as its uncertainty, this may serve to recognize faults that might otherwise remain unnoticed. To leverage this human capability to detect system faults, preventing *automation bias* and *automation complacency* is crucial [224, 225, 226]. It has been demonstrated [226, 225] that in many practical scenarios, human operators — especially when under time pressure and multi-tasking — quickly place too much trust in an automated system which has performed reliably in a large number of cases, thereby increasing the likelihood of a system fault not being detected by the operator. Proper user training and emphasizing the user's responsibility have been demonstrated to reduce the prevalence

[13] See, e.g., https://www.wired.com/story/facebooks-red-team-hacks-ai-programs/.

of automation bias in healthcare workers operating clinical decision-making systems, alongside careful user interface design and reporting of confidence ratings [225]. Both proper user training and deliberate user interface design are key elements of medical device development processes [73].

Finally, as a last resort, medical ML systems should react robustly to any failures that might occur despite all prevention measures, i.e., perform failure *mitigation*. This becomes especially important once the ML system automatically interacts with the patient in a closed loop. For this reason, this challenge will be discussed in a future publication concerning *dynamic* MML systems.

3.4 Summary

To achieve functional safety of ML-based medical systems, failures must be *prevented* where possible, *detected* if they occur nevertheless, and *mitigated* to ensure that the system remains in a safe state at all times. In MML systems, model robustness and reliability represent essential requirements for failure prevention. Unfortunately, these properties are inherently difficult to assess with regards to neural network models because good test set accuracy often does not translate into good real-world performance due to distribution shift [122, 129], model underspecification [137], and test set label errors [186], among other challenges. The use of large, representative datasets (which may be aided by employing federated learning methods, synthetic data augmentation, or automated labeling schemes), the incorporation of prior knowledge about the class of reasonable models, and the use of formal verification techniques are key ingredients to achieve models that are robust and reliable. However, to ensure the safety of an MML system, one should not stop at technical considerations about the model but also consider measures to increase the safety of the whole system, such as operator training, proper user interface design, and physical enforcement or automatic verification of safe operating conditions. Such *usability engineering* methods are required by current medical device regulations as a risk mitigation strategy, and MML developers can draw on decades of experiences and standards in this field [73]. However, important transfer work in developing functional safety concepts (and standards) for MML applications remains to be done [101]. A unique role is played by transparency — regarding, e.g., the training data composition, model structure, training process, and final trained model — which increases robustness and reliability throughout all lifecycle stages of the MML system, and which will be the exclusive subject of section 5. Finally, section 6 will discuss the problem of fairness and nondiscrimination, which is closely related to reliability: if a model performs equally well across all relevant patient groups, it is both reliable and fair (at least concerning this particular definition of fairness).

4 Privacy & security

Medical machine learning applications demand an exceptionally high degree of data privacy and application security. Beyond classical privacy and security measures, this necessitates guaranteeing the privacy and security of the employed ML methods. While providing (possibly restricted) public access to the MML system is unavoidable if it is to be used in practice — users must at least be able to provide inputs to the system and receive model predictions in return — it should at no point during the system's lifecycle be possible for an outsider to extract sensitive patient data. Moreover, malicious actors must be prevented from negatively influencing the outcome of the training process or a particular prediction. Spurred by the urgent practical and theoretical challenges this poses to the realization of privacy-preserving and secure machine learning processes, recent years have seen a proliferation of research on subjects such as privacy-preserving ML and federated learning. In this context, an important role is played by different communication schemes that can be pursued, and that determine who has access to the training data, the trained model, and a patient's query data that serve as inputs to the trained model. Figure 4 depicts four commonly used communication schemes that may be used in practice. The following sections will discuss the vulnerabilities of each of these setups, as well as possible defenses against malicious attackers.

4.1 Obstacles to achieving privacy & security

The challenges of achieving privacy and security in ML healthcare applications are (at least) sixfold:

1. classical cybersecurity,
2. confidentiality of the training data (models shall protect their training data),
3. confidentiality of the query data (patients that use a model shall not expose their patient data),
4. confidentiality of the model (preventing the extraction of the model from a cloud service),
5. integrity of the training process (malicious actors must not be able to influence the final model negatively or extract training data), and
6. integrity and robustness of the model (adversarially crafted inputs shall not confuse the model).

Classical cybersecurity Classical cybersecurity challenges include bugs and vulnerabilities of the ML healthcare application's software stack. Such vulnerabilities can lead to a completely compromised system and massive incidents, as a growing number of security incidents

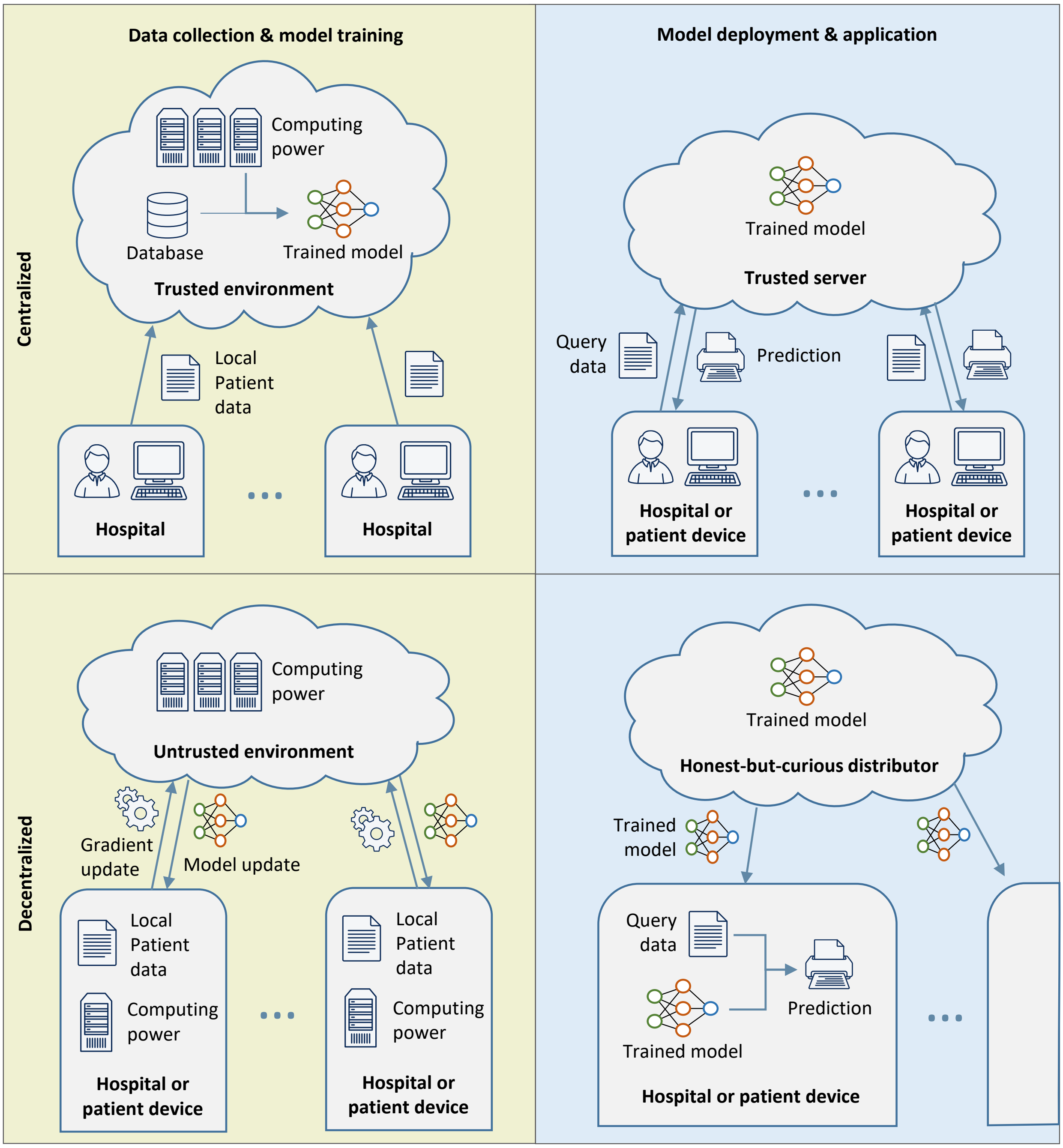

Figure 4: Many different communication schemes have been proposed for data collection, model training, and model deployment. All of them have merits and drawbacks — many of which will be discussed in section 4.1 — and are preferable under different circumstances. Mechanisms for preserving privacy and security in these different schemes against various attacks have been proposed; they are discussed in section 4.2. Centralized training and deployment is the current standard scheme in ML. The depicted decentralized training scheme corresponds to (horizontal) federated learning; the depicted decentralized deployment scheme is often called "Edge AI". (For decentralized, e.g., federated, training, a central coordinator is not strictly necessary: completely decentralized approaches have been proposed as well.)

show [227, 228]. As medical devices are increasingly interconnected, they become easier targets, as reports on the vulnerability of medical devices show [229]. As a result, cybersecurity now plays a major role in the approval process of medical devices [230].

Confidentiality of the training data Recent work [231] shows attack strategies for extracting information about training data from ML models; thus, the ML model itself breaks the confidentiality of the training data. For models that are solely used by authorized personnel (e.g., hospital personnel), such extraction attacks are not a concern. In the context of Edge AI, however, the ML model is deployed to end-users' smartphones, and any malicious party can query the model and extract information about the training data. Even ML models hosted in a protected environment and query-frequency-limited seem to be vulnerable to privacy attacks. So far, there is no indication that restricting the frequency of queries would alleviate these attacks. These privacy attacks do, however, have varied success on applications with different degrees of complexity. Recent work has shown [232] that it is harder to protect applications with high-dimensional input spaces, such as images or signals, compared to applications with low-dimensional input spaces, such as structured data, questionnaire responses, or tables with a small number of attributes. Aside from the *model* potentially leaking sensitive patient data, it is also becoming increasingly apparent that publishing medical datasets in a matter that prevents the re-identification of supposedly anonymized health data is a highly challenging endeavor. To provide two recent examples, Rocher et al. [90] demonstrate that using publicly available, supposedly anonymized datasets, a large fraction of the U.S. population could be re-identified (thereby defeating the purpose of anonymization). Similarly, Packhäuser et al. [233] demonstrate that given a chest X-ray recording of a patient, other recordings of the same patient can be identified among more than 100,000 recordings in a publicly available dataset with very high accuracy, thereby enabling re-identification of the supposedly anonymized recordings in the dataset. Both authors thus call for a reconsideration of the classical anonymization strategies, i.e., de-identification and population sampling.

Confidentiality of the query data Sensitive patient data are threatened not only during model training but also when querying an ML model. In Edge AI applications, where the model remains on the patient's side while querying it, the patient's query data — serving as inputs to the model — are protected. At the same time, however, any potentially malicious party can freely query the model and thus extract potentially sensitive information about the training data (see above). If the ML model itself is an asset or simply too large for a client device (such as GPT-3 [234]), a patient would remotely access the model, and the model might even run on a cloud service. In this case, a naïve implementation could leak sensitive patient data during each query.

Confidentiality of the model The model itself can be of high value. From a business perspective, it might be desirable to keep the precise model private. While classic security measures may prevent direct model access, a recent line of work [235] shows that by using a sufficiently large amount of queries, attackers can extract a model from inference only. This attack scenario becomes relevant once an attacker has either a device on which the valuable ML model is deployed or unlimited remote access to a valuable model.

Integrity of the training process For privacy and regulatory reasons, it may be desirable to train an ML model in a distributed manner. Federated and distributed learning both are techniques using which several parties can keep their (confidential) data local yet train a global model together [148, 236], and many groups have advocated for the use of such techniques for healthcare data [237, 153, 91, 154]. In these distributed learning settings and, more generally, if the data collection process cannot be fully trusted, the integrity of the training process has to be protected. Hitaj et al. [238] have demonstrated that malicious participants to the training process can extract the data of other participants, and recent work has shown [239] that malicious training parties can also manipulate the resulting global model. These manipulations can even result in backdoors: for specific inputs, the globally trained model may return maliciously crafted undesirable results, rendering the model unreliable and potentially unsafe.

Integrity and robustness of the model Beyond the integrity of the training process, ML applications face the challenge that classical ML models can — infamously — be confused by maliciously crafted inputs, called adversarial examples [240, 241, 107, 242]. Adversarial examples can be crafted from legitimate inputs with minuscule modifications that are practically invisible for a human but significantly alter an ML model's output. Naturally, in protected environments — such as a hospital — where only authorized personnel uses an ML model, adversarially crafted inputs are not a concern. For Edge AI applications and remotely accessible ML models, however, adversarial examples can render an ML model unreliable. This might concern, e.g., MML systems intended for home use. The vulnerability to adversarial examples is higher in applications with high-dimensional input spaces [243], in particular those applications where the inputs have a natural source of perturbation, such as camera images or physiological measurement signals. Humans have a high tolerance for signal variation in such settings, whereas non-robust ML models can drastically change their output if the input is only slightly modified. As an example, minimal modifications in the lighting of an image will result in a technically different image that appears indiscernible from the original image for humans but may be classified differently by a non-robust ML model. Changes in *low*-dimensional input spaces (e.g., structured data or tables with a small number of attributes) on the other hand can be spotted easier by humans and leave less freedom for unobservable modifications; an effect sometimes referred to as the curse of dimensionality in adversarial examples [243]. Finally, notice that a model's vulnerability to adversarial examples is closely related to its *robustness*, also cf. the corresponding discussion in section 3.

4.2 Measures to achieve privacy and security

For classical *cybersecurity*, there is a rich body of literature regarding best practices, such as running known attack strategies against the software stack (so-called penetration testing), formal verification, and architecture design that isolates as many system parts as is feasible, and organizational measures, such as incidence response teams and red teaming. For an actionable summary of practical best practices for facing classical cybersecurity challenges, refer to, e.g., the NIST[14] and BSI guidelines.[15]

For ensuring that a model does not leak information about the data it was trained on, a young body of literature [244, 245] on *privacy-preserving machine learning* (PPML) presents methods to train machine learning models in a way that ensures that no information about individual patients can be extracted; an approach called *differential privacy*. As such PPML methods try to approximate the original learning task in a well-generalizable and privacy-preserving manner, these methods inherently lead to a loss in accuracy. This loss of accuracy is manageable for convex optimization problems, such as linear or logistic regression, SVMs, or rank pooling. For non-convex methods such as neural networks, however, current (differential privacy) approaches lead to a significant loss of performance [244], although this loss can be circumvented under specific circumstances [246]. In some cases, pre-training an ML model on public data and then fine-tuning this pre-trained model via privacy-preserving methods — a variant of *transfer learning*, which was already discussed in more detail in section 3.2 — might represent an attractive alternative from both a privacy and an accuracy perspective. To validate an employed technique, penetration testing frameworks such as the adversarial robustness toolbox (ART) [247][16] can be utilized for assessing the vulnerability of the trained model to extraction attacks, among other threats.

Distributed learning methods (see the bottom left panel in fig. 4) such as *federated learning* [147, 148] and *split learning* [149] have already been discussed in some detail in section 3. They represent attractive methods for increasing the pool of available training data without having to share these data between different institutions. This is especially relevant in the face of increasingly prominent de-anonymization attacks on publicly available datasets [90, 233]. It is, however, important to note that they still suffer from the problem of information leakage from the trained *models*. To solve this problem, distributed learning methods can be combined with, e.g., differential privacy schemes, as discussed above, or homomorphic encryption [92, 93, 237, 148]. Protecting the integrity of the *training process* during a distributed learning process against malicious actors is a very young challenge that is most relevant in scenarios in which there is a large number of potentially unknown participants. The best automated techniques use different kinds of outlier detection methods that look for suspicious behavior of single training parties [248], similarly to the outlier rejection techniques discussed in section 3.3. Potentially, slightly noising the model's responses might increase the number of queries that are needed to accurately extract the model [249]. These techniques increase the difficulty of injecting backdoors yet cannot provide security guarantees against strong attackers. A taxonomy and overview of different attacks on distributed learning processes and possible defenses against them can be found in Lyu et al. [250]. As was mentioned above, many of such integrity problems can be circumvented by limiting the number of training process participants to a relatively small, known group of institutions, such as a set of hospitals.

One way to protect the *query* data, i.e., sensitive patient data serving as inputs to the model, is to deploy the trained model in the "edge", i.e., on a patient or hospital device, thereby alleviating the need to communicate the query data with any external entity. (This corresponds to the setting in the bottom right panel of fig. 4.) However, as discussed above, if the data was sensitive, so is the trained ML model. If solely authorized personnel is involved, i.e., in a hospital setting, this is not a problem and may be no need for special protection. If, however, the ML model is accessible to unauthorized query issuers — as is common in large scale applications —, classical (not privacy-preserving) machine learning techniques can lead to undesired information leakage about the sensitive training data, as discussed above. Again, differentially private machine learning can be used to avoid such leakage. Another challenge in this setting

[14] See https://www.nist.gov/itl/voting/security-recommendations.

[15] See https://www.bsi.bund.de/EN/Topics/ITGrundschutz/itgrundschutz_node.html.

[16] See https://art360.mybluemix.net/.

concerns the validity of the model: if a remote server distributes the model, how can its validity be verified on an end-user device? This problem can be addressed using standard techniques for verifying remotely distributed software, as are used by, e.g., various software package managers and update mechanisms.

In the alternative scenario in which patients *remotely* access the model (which runs on a centralized server), recent work [251] has shown that trusted computing platforms (such as Intel SGX [252]) can be used to ensure that the model owner cannot access the sensitive patient query while computing the response to the query. Additionally, cryptographic solutions exist where both parties either compute on encrypted data via homomorphic encryption [253], using secure multi-party computation [254], or using outsourced computation techniques [255].

Regarding the preservation of the confidentiality of a trained model against *model extraction* attacks, the research literature is also very young. As the current attacks require a large number of queries to the model, the best countermeasure currently seems to be to limit the query rate to a model [256].

Safeguarding the integrity of an ML model against adversarial attacks, a property called *adversarial robustness*, is another young and challenging research area. This is, again, most relevant in Edge AI settings, in which model inputs can be influenced not only by certified personnel. Many so-called *best-effort* defense mechanisms, i.e., mechanisms which do not provide security guarantees, have been proposed in the literature to defend ML models against adversarial examples. Several of them, however, turned out to be unsafe against adaptive attackers who are aware of the deployed defense mechanism and its implementation [257, 258, 259]. See [242, 260, 258, 259] for an overview of defense mechanisms that are nullified by adaptive attacks. A best-effort defense mechanism that is still considered practically valuable is adversarial training [240, 261, 262]. On the other hand, there are adversarial defenses that offer certifiable robustness against a subset of adversarial attacks, such as the mechanisms described by Lécuyer et al. [211] and Cohen et al. [212]. For a given input x, these defenses can certify for a certain distance r (which depends on x) that no adversarial example exists within a ball of radius r centered at x. Alas, for some inputs r might be equal to zero, providing no guarantees for these inputs. An attacker restricted to such a threat model, even if adaptive, will not be able to create adversarial examples within those boundaries. Recently, Zhang et al. [136] have provided an interpretation of adversarial attacks as intentional exploits of spurious correlations learned by the model. The authors also provide a novel defense mechanism that leverages causal modeling for improving adversarial robustness. In general, it is crucial to carefully evaluate any implemented defense mechanism in order to avoid a false sense of security [263]. To perform such an evaluation in practice, pen-testing frameworks like the FoolBox [264, 265], ART [247][16], or Cleverhans [266] toolboxes may be employed. Finally, notice that methods to increase robustness against adversarial attacks do *not* necessarily increase robustness against natural distribution shifts [129].

4.3 Summary

Medical machine learning applications necessitate exceptionally high standards of privacy and security. Distributed learning models such as federated learning and split learning appear to be promising frameworks for solving the problem of medical data silos while maintaining patient data privacy [91, 92, 93, 153, 154]. For MML models solely accessible by authorized personnel, security and privacy concerns appear to be relatively limited. However, in scenarios with more widely distributed access, models should be protected from direct access due to privacy concerns. For MML models that are remotely accessible to a broader audience or for Edge AI applications, a limited query rate, robustness-enhancing training methods, and privacy-preserving learning methods should be considered. In particular, it is advisable to evaluate the security and privacy properties of such a model using popular penetration testing frameworks [264, 265, 247, 266]. Finally, in scenarios where the data collection or the training process is not entirely trusted (e.g., in some federated learning settings), countermeasures against backdoor injections may be necessary [250], such as deploying outlier detection methods [248].

5 Transparency & explainability

By *transparency*, following Dignum [39], we will refer to “the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and the provenance and dynamics of the data that is used and created by the system. [...] transparency is also about being explicit and open about choices and decisions concerning data sources and development processes and stakeholders.” [39]. Thus, transparency encompasses the open communication of some or all factors that influence the decision-making process, including, among others, the training data and their characteristics, the selected model structure, the training procedure, performance on the training and test datasets, real-world performance across different patient groups, and more. Notice that transparency alone does not necessarily enable *comprehension* on the part of the information recipient: making the full structure and weights of a neural network openly available can be seen as fully transparent, but due to the model's opacity likely does not significantly increase recipient comprehension. Thus, for many purposes, transparency alone is not sufficient, and *explainability* must be a second goal. By explainability, we refer to "the ability to explain or to present *in understandable terms* to a human" (emphasis ours) [267].[17] In all except for the very simplest cases, there is a trade-off between the understandability and the completeness of an explanation [269, 270]: facilitating human comprehension necessitates meaningful abstraction and reduction of details, just like a medical doctor would not provide her whole scholarly knowledge as an explanation but rather present the few most important factors influencing her decision. Thus, useful explanations of complex models can never be complete, i.e., can never correctly reflect all influencing factors; a conundrum that is sometimes called the *approximation dilemma* [271].

In consulting current regulatory documents for medical devices, a clear demand for transparency or explainability of solutions for medical use is not to be found (although this appears likely to change in the near future, cf. section 2). In Europe, the GDPR, granting a right to an explanation and a right to contest any automated decision, does require some (application-dependent and currently unclear) degree of explainability and transparency in automated decision-making systems.[18] Similarly, the EU commission's proposal for the regulation of AI systems [29] emphasizes an (albeit vague) need for transparency. On the other hand, transparency and explainability serve many practical purposes: they aid developers to create safe, robust, reliable, and fair systems [214], auditors to evaluate system performance [53], and doctors and patients to understand and potentially challenge algorithmic decisions [11]. From a medical device manufacturer's perspective, transparency and explainability support risk assessment, risk mitigation during development and deployment, and clinical validation of machine learning solutions, besides fulfilling a stakeholder need. Both the FDA's plan to advance the agency's oversight of AI/ML-based medical software [23] and the WHO's guidance [11] emphasize the role of transparency of AI/ML-based devices as an essential factor to achieve patient, clinician, and societal *trust* in AI/ML technologies. From an ethical perspective, transparency to the user is often understood as a prerequisite for human agency in the face of algorithmic decision-making [272, 273, 11]: if people are affected by the consequences of an ML system's decision, they must receive sufficient information to be able to exercise their rights appropriately and, if necessary, challenge the decision (as is also required by the GDPR [82]).

5.1 Goals of transparency and explainability: transparent to *whom*?

There are many stakeholders involved in the lifecycle of an MML system (cf. fig. 1), with different levels of expertise regarding machine learning in general and a given system in particular, and with different reasons to be interested in transparency and explainability of the system [274, 275, 271, 276]. Weller [274] lists a number of potential goals of transparency and explainability, including

- helping a *developer* and, more generally, the *manufacturer* understand the system and the way particular decisions are made, thus aiding the development of a safe, robust, and reliable system,
- helping a particular *user* (which may be a patient or a healthcare professional) and the *society at large* broadly understand the system as a whole and thus aiding the formation of trust or distrust in the fitness of the system for a particular application,
- helping a particular *user* understand the most important factors influencing a particular prediction, thus enabling him to trust or challenge it,
- helping an *auditor* or regulator assess the trustworthiness of the system as a whole or the way a particular decision was made, and
- helping any of the above stakeholders assess the development of system performance over time after market deployment.

Each of these goals requires different kinds of information, different types of explanations, and different means of communicating the explanation. In other words, the transparency and explainability–enhancing features of an MML system must be tailored towards a specific purpose [275, 271]: whom should they serve?

Crucially, transparency does not necessarily require openly communicating all aspects of the system to the public — this may be undesirable for various reasons, including privacy concerns regarding patient data and intellectual property concerns regarding the exact employed models [277, 99, 11]. For instance, it may be preferable to communicate critical implementation details such as model structure, training routine, and the full training, validation, and test datasets to auditors only [277, 99]. A developer, on the other hand, who is struggling with the severe task to create a system that is robust and reliable in the face of all the challenges discussed in sections 3, 4 and 6, will benefit from transparent access and well-tailored explanations of all aspects of the system [271, 214]. In all likelihood, the optimal level of detail differs between explanations for said developer, for a healthcare professional working with the system, and for a patient affected by its decision.

[17] Like Miller [268], Tjoa and Guan [222], we do not differentiate between explainability and *interpretability* and use the two terms interchangeably.

[18] Cf. the discussion in section 2.

Finally, one particular goal of transparency may be to monitor the temporal development of a continuously learning MML system. In this case, transparency and explainability measures may serve to assess whether an iteratively improving system stays safe, robust, reliable, and fair at each time. This connection between transparency requirements and continuously learning systems is emphasized, e.g., by the FDA [52, 23].

5.2 Technical measures to achieve transparency

Transparency should, wherever feasible, be pursued regarding all aspects of a model, its origin, and its performance [278, 140, 11]. To begin with, transparency regarding the (composition of the) training and test datasets [279] as well as the *process* by which these are gathered and curated [139] has been recognized as crucially important. This is hardly surprising, seeing that the robustness, reliability, and fairness of the resulting model are strongly dependent on (albeit not entirely determined by) the properties of these datasets. Frequently, samples are not representative of the target population and, e.g., tend to be disproportional sick or biased towards the local population [119]. Various exploratory data analysis techniques may be used to characterize a dataset in a meaningful way, including, e.g., data prototypes and criticisms [187, 280, 281], which characterize, respectively, dominant and rare (potentially under-represented) clusters of data. One may aim to be transparent regarding the full, raw training and test datasets, transparent regarding the way in which these data were collected, or transparent regarding the prevalence and characteristics of certain (age, diagnosis, sex, ethnic origin, recording type, and location) groups. What is the model's performance overall and within these different (training or test) subgroups? Are there significant discrepancies to be observed? This information may drive decisions concerning, e.g., the proper operating conditions under which the model is believed to operate reliably, the necessity to collect further data, or the necessity to modify the model class or training algorithm to improve the model's performance. Due to the constantly lurking threat of dataset shift between the training data and real-world data, post-market real-world performance monitoring is believed to play a key role in building trust in an MML system [52, 23, 29]. (Post-market surveillance is, of course, already a key requirement of presently existing medical device regulations.) To the clinical user, transparency regarding the data that were used to train the algorithm and their characteristics, the relevance of the different input data provided to the system, the logic employed by the system, the role intended to be served by its output, and data quantifying the device's performance are crucial indicators for assessing the credibility of the system in a certain setting [23]. Moreover, the conditions under which a system is safely applicable must be clearly communicated.[19] Popular transparency initiatives include the TRIPOD statement for reporting multi-variable prediction models for individual prognosis or diagnosis [278], the "Datasheets for Datasets" proposal [279], the Google model cards initiative [140],[20] the dataset nutrition label [282][21] and AI FactSheets [283].[22]

Achieving *contestability* requires a particular kind of transparency, namely, *traceability* [82]. If users (e.g., patients or healthcare professionals) are to be given the power to contest an automated decision post-hoc, i.e., *after* it has been made, the system must implement an infrastructure that enables a) retracing the way that particular decision was made and b) assessing whether this decision-making process adhered to the relevant regulations and requirements [82]. Aler Tubella et al. [82] propose to address this technical challenge by implementing a traceability software infrastructure and enabling post-hoc verification of formalized requirements that must be fulfilled by a lawful (or otherwise desirable) decision-making process. (Also cf. the discussion of formal verification methods in section 3.)

Finally, transparency about prediction or classification *uncertainty* is another essential requirement: how certain is the system regarding a particular prediction? Proper uncertainty quantification is a challenging task and requires a careful analysis and characterization of general model reliability. Uncertainty quantification may be achieved using Bayesian techniques [284, 285, 286], case-based reasoning [287] or proper calibration of the estimated predictor [288, 289, 290]. Unfortunately, proper calibration of deep neural networks has been found to be particularly challenging in the healthcare context due to the limited amount of available data [291]. For a comprehensive review and introduction to different techniques for uncertainty quantification in deep neural networks, refer to Gawlikowski et al. [292].

The following section will discuss measures to aid developers, auditors, and users in understanding the functioning of the trained model.

5.3 Technical measures to achieve explainability

Popular interest in explainable machine learning models has exploded in recent years, with government agencies across the globe demanding that machine learning systems be explainable, multiple major companies providing toolboxes to implement explainability, and whole MOOCs being launched to cover it [293, 294, 295, 271]. In the medical field, calls for the explainability of MML systems abound [86, 200, 10, 119, 1, 121]. To begin the discussion, a distinction has to be made between mod-

[19] If possible, any deviation from these conditions should also be detected automatically, cf section 3.3.

[20] See https://modelcards.withgoogle.com.

[21] See https://datanutrition.org/labels/.

[22] See https://aifs360.mybluemix.net/.

els that are *intrinsically* interpretable and black-box models that are made explainable through post-hoc explanation methods [271, 281].[23] Whereas explaining a non-interpretable, black-box model necessarily always requires omitting (supposedly less important) details, intrinsically interpretable models have the advantage that their functioning can be explained to users without potentially misleading simplification [296].

The class of intrinsically interpretable machine learning models is larger than one might expect. Firstly, it includes all the classical, simple (rather low-dimensional) parametric models, decision trees, generalized linear models, and many more [281]. Various types of probabilistic graphical models such as Bayesian networks — including, in particular, causal models [196, 297] — are interpretable as well, at least up to a certain degree of model complexity. Bayesian rule lists were introduced to achieve interpretable and highly accurate models for stroke prediction [298]. *Causal* models (mainly applicable for use with structured or time-series data) are inherently explainable and provide explanations in a particularly natural form [299, 200, 297], besides bringing many other benefits to MML [133]. (Also cf. the discussion of causal models in section 3.2.) Many more examples of interpretable model classes can be found in Marcinkevičs and Vogt [300]. Bridging the gap towards black-box neural networks, various interpretable yet highly flexible model classes have been proposed. As one example, *case-based reasoning* architectures leverage deep learning for visual perception yet are inherently interpretable [301, 296] and have also been employed successfully in medical imaging tasks [287] without losing accuracy in comparison to purely black-box models [194]. In a similar vein, Chen et al. [193] have recently proposed *concept whitening* as a method that can be used with any of the standard deep learning frameworks, and that enforces the learning of interpretable concepts in one layer of a deep neural network. The authors provide many examples illustrating their proposed concept, including a skin lesion diagnosis task. Notably, the authors did not find concept whitening to reduce the predictive capabilities of the network. For time series analysis, Sha and Wang [302] have discussed the use of an intrinsically interpretable recurrent neural network for mortality prediction based on diagnostic codes, also finding their proposed model to outperform baseline models. Similarly, Guo et al. [303] have proposed an interpretable LSTM model. Additional approaches to equip neural networks with intrinsic interpretability (as opposed to post-hoc explanations) include contextual explanation networks [304] and self-explaining neural networks [305].

A model can be a *black box* either because it is too complicated for humans to comprehend (this famously includes most deep learning models) or because it is proprietary (this may be the case in many medical applications). In both cases, an understanding of the model's functionality can only be extracted by examination of the model's response surface, i.e., by means of *post-hoc* methods. *Saliency maps* [306, 307, 308], which visually represent the importance of each pixel for arriving at a particular prediction, are a popular post-hoc method to achieve explainability of black-box computer vision systems, with popular methods for their computation including layer-wise relevance propagation [309], Grad-CAM [310] and spectral relevance analysis [311]. Implementations are readily available and, in many cases, quickly yield information that is useful to the developer or the user [312, 222, 300, 214]. Their prevalent use, especially in the medical domain, has been criticized, however, because they can be misleading or unreliable [296, 313]. For less high-dimensional applications than visual computing, *feature importance* [281], *locally interpretable model-agnostic explanations (LIME)* [270, 281] and *Shapley values* [314, 281] all quantify the importance of different input components for arriving at a particular prediction, analogously to saliency maps for visual computing. Similar to saliency maps, these post-hoc explanations can be misleading due to their simplifying nature [304, 281]. Very differently from the attention-based prediction explanation mechanisms discussed above, *example-based explanations* attempt to explain a prediction by providing examples that the model considers similar or dissimilar in some sense [281]. Wachter et al. [315] propose a general method for generating counterfactual explanations, i.e., examples that are close to the input data at hand (in some distance metric) but classified differently. The authors also discuss the example of a model predicting a patient's risk of diabetes. Counterfactual examples provide very naturally human-interpretable explanations [316, 268]. While some technical challenges are to be solved for a particular application problem [281] and various implementations have been proposed [300], the main challenge when implementing this method is that there are usually many potential counterfactual examples, and selecting the most explanatory is decisive and non-trivial [317].

The post-hoc explanation methods discussed so far were all *local*: they attempt to explain how a black-box model has derived some particular prediction and what would need to change in order for it to reach a different conclusion. Another branch of explainable ML is concerned with finding *global* explanations of the *model as a whole*. Maybe the most straightforward method for doing so is to identify a *surrogate model* from an interpretable model class that approximates the original black-box model [281]. A more complex but similar approach has been pursued by Harradon et al. [318], who identify

[23] Interpretability is *gradual*, i.e., there is no sharp boundary between interpretable and non-interpretable models. A white-box probabilistic graphical model with thousands of variables and connections between them may no longer be classified as "interpretable".

human-comprehensible concepts within a deep neural network, extract a probabilistic causal model and use this extracted (surrogate) model to generated explanations of the network's decisions. Of course, all methods identifying a global surrogate model have the obvious drawback that they may not faithfully represent various properties of the original model due to their approximative nature. As an alternative, one may, again, also consider different example-based model explanations. *Adversarial examples* [240] may serve to examine the brittleness of a trained model with respect to small perturbations. A model's predictions for *prototypes* and *criticisms* [280] may be examined to analyze whether, e.g., the model performs poorly for some under-represented groups. Finally, akin to classical linear model analysis [319], the influence of individual examples in the training dataset can be analyzed, e.g., using *influence functions* [320]. In this way, examples with a particularly strong influence on the resulting model parameters can be identified and checked for potential errors or confounding properties [281].

It is a popular claim that there is a trade-off between model accuracy and model interpretability [321, 296, 271]. This claim has been refuted [296, 193], and many counter-examples — i.e., interpretable models achieving state-of-the-art performance on par with black-box models — have been provided [130, 298, 296, 287, 193, 304, 194]. While simply reducing model complexity is likely to reduce model performance, there are often ways to constrain the model class to be interpretable while retaining flexibility in the dimensions that matter for achieving a high performance [296, 271, 193]. Currently, this may require hand-crafting an interpretable model class that is suitable to the target domain [296, 194], an effort that may be prohibitive in some cases. This might be changing, however. As an example, the recently proposed concept whitening method is directly applicable using standard deep learning frameworks [193]. Moreover, in a field where safety is crucial, data are scarce, and patient recordings differ in a multitude of dimensions, it may be desirable to sacrifice a bit of predictive performance to gain increased explainability of the employed model. An interesting hybrid approach might be to craft an MML system that consists of both interpretable and black-box components, thereby combining domain-specific knowledge with general black-box type adaption layers. So far, however, the authors are unaware of examples demonstrating a successful medical application of latest-generation general-purpose models (like transformers or performers [322]).

An expository review of explainability and interpretability can be found in Marcinkevičs and Vogt [300]. For a comprehensive review and taxonomy of explainability methods and their current challenges, refer to, e.g., Arrieta et al. [271]; for a review of explainability within the MML context, see Tjoa and Guan [222]. A useful, practical guidance document has been developed by the UK information commissioner's office (ICO) and the Alan Turing institute [88]. A number of both freely accessible and commercial tools implementing (mostly post-hoc) explainability methods is already available, such as the AI Explainability 360 toolkit [323],[24] InterpretML [324],[25] Google's "What-if Tool" [325][26] and similar tools in Microsoft's Azure.[27]

The field of explainability raises many fundamental philosophical and psychological questions, such as:

- What constitutes an explanation? What differentiates it from, e.g., a justification?
- How do we humans explain things to each other, and how do we mentally process explanations?
- Which different types of explanations are there?
- What makes an explanation "good" in a given context?

We did not touch upon these matters here for brevity's sake; the interested reader is referred to the DARPA XAI literature review [295] and Miller [268] for broad overviews of the interdisciplinary challenges related to crafting good explanations.

5.4 Non-technical measures to achieve transparency and explainability

Achieving real transparency and explainability necessitates thoughtful interaction with the target information recipient. What kind of information is she looking for? Which type of explanation does she consider useful? For these reasons, manufacturers should closely involve stakeholders in the design of a transparent and explainable MML system [121, 222] and carefully analyze the clinical decision process into which the MML system should be integrated [40]. Appropriate methods must be developed for *communicating* the information to the user in a readily accessible, comprehensible, clear, and, ideally, interactive way [268, 295, 326, 327]. This may include the creation of informative summarizing *labels* for MML systems, which may inform about the data used for training the model, the model type used, its performance on benchmark datasets and in validation studies, its intended use, and its limitations [278, 140, 328, 23]. It will likely also demand the innovation of novel explanatory interfaces [329, 287, 294, 276, 326, 194, 327]. Notice that *evaluating* the transparency and explainability of (different versions of) an MML system quantitatively represents a difficult challenge and is the subject of current research [267, 330, 300, 295, 271, 312]. Finally, medical users and patients, unaccustomed to working with

[24] See `https://aix360.mybluemix.net/`.

[25] See `https://interpret.ml/`.

[26] See `https://pair-code.github.io/what-if-tool/`.

[27] See `https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability`.

and thinking about ML models and systems, will require proper training [222, 223, 23] to empower them to make informed decisions when interacting with an MML system.

5.5 Summary

Regardless of the exact degree to which they may or may not be legally required, there is a broad consensus that transparency and explainability are crucially important enabling properties to achieve safe, robust, reliable, and fair MML systems. This includes transparency regarding the data used for training a model and its characteristics and limitations, the model's real-world performance across different patient groups, the uncertainty of model predictions, explanations of the model as a whole, and explanations of the most important factors influencing a particular decision of the model. Data-based explanations of model behavior, such as model predictions for minority group prototypes, adversarial examples, and influential examples, may help assess the model's robustness and reliability and inform continuous model improvement. In a similar vein, the generation of counterfactual examples represents an attractive explanatory mechanism that is applicable in many domains. Post-hoc explanation methods, such as saliency maps and LIME, can be used to explain the decisions of otherwise opaque models. Due to their approximate nature, however, they can be inaccurate and misleading and hence should only be used with great care in the medical domain. It has been demonstrated in multiple studies across various domains that intrinsically interpretable model classes can achieve the same performance as black-box models; such models should thus be preferred where possible. Finally, explanations must always be tailored to a particular stakeholder group and use case. Explanation design is a highly interdisciplinary endeavor that should involve the relevant stakeholders from the beginning.

6 Algorithmic fairness & nondiscrimination

Due to the proliferation of machine learning systems making decisions with a significant impact on human lives, recent years have seen a surge of interest in *algorithmic fairness* research [331]. Examples of biased machine learning systems eliciting serious consequences have been widely reported in the media, including

- publicly used face recognition algorithms performing best for White men [332],
- an algorithm used to assign U.S. healthcare resources to patients discriminating against Black patients [17], and
- disparities in true positive rates of chest X-ray diagnosis systems between patients of different sex, age, race, or insurance types [16].

In a recent survey of industry ML practitioners [333], 49% "reported that their team had previously found potential fairness issues in their products. Of the 51% who indicated that their team had not found any issues, most (80%) suspected that there might be issues they had not yet detected, with a majority (55%) reporting they believe undetected issues 'Probably' or 'Definitely' exist." Ethics, nondiscrimination law, and technical accuracy and reliability requirements all demand that medical machine learning (MML) systems treat patients fairly and do not create or reproduce biases [29, 11]. But what exactly constitutes a "fair" MML system, and why are "unfair" biases so prevalent?

6.1 Obstacles to achieving fairness

Fairness is a context-dependent social construct, and thus, there is no universally accepted and applicable definition of fairness in medical ethics [334, 335, 336]. Many different, conflicting definitions of algorithmic fairness have been proposed for general [337] as well as specifically medical [338] applications of machine learning. It is readily apparent that different medical applications demand different definitions of fairness: while a healthcare resource allocation algorithm should not allocate resources based on sensitive attributes such as gender, ethnicity, or social status, a disease or treatment prediction algorithm probably should take at least some sensitive attributes into account [339]. Moreover, predictive performance for one patient group must not be sacrificed in order to achieve fairness by means of equal(ly bad) performance across groups: instead, the aim must be to achieve optimal performance for each individual group [339]. A related challenge concerns the difficulty of identifying the groups that are relevant for fairness considerations in the first place [333]. These groups depend on the application at hand and may not be obvious from the beginning; e.g., native speakers and non-native speakers may form fairness-relevant groups in applications related to text generation and understanding, but not in many others [333]. There is a particular risk to overlook potential sources of unfair discrimination due to blind spots in the development team [333]. For these reasons, the decision for a particular definition of fairness to pursue in a given application should generally not be made by the ML developer alone but by a broad and inclusive group of stakeholders — including, in particular, medical professionals, medical ethicists, and patients [335, 340, 341]. Given a fairness definition to aim for, there are, however, still considerable challenges involved in creating a model that conforms with this definition.

From a technical point of view, none of the negative examples mentioned above are particularly surprising: machine learning models, to a large degree, reflect the data they are trained on — and if these data are biased, then the resulting machine learning model is also likely to be biased. There are many different potential sources of bias in a dataset. As an example of *historical bias*, the health risk score algorithm studied by Obermeyer et al. [17] used past health care cost data as a proxy for health care needs — and since Black patients historically received inferior treatment, the algorithm assigned a lower risk score to them compared to White patients with equally severe symptoms. This example also illustrates the *measurement bias* that may arise from the incautious use of proxy variables for training ML models, such as healthcare costs, hospital visits, or medication usage as proxies for patient health. Measurement bias may also arise due to discrepancies in diagnostic and treatment recommendation accuracy across, e.g., different gender, ethnic, or weight groups [342, 343, 344, 345]. Face recognition algorithms may serve as an example of *representation bias*: the databases these algorithms are trained on often contain many more samples of North American and Western European persons than from the rest of the world [342], leading to better performance of the trained models on "western" subjects [346, 332]. Similarly, most biomedical databases significantly under-represent large parts of the population [146]. The biased data problem is very hard to solve because learning algorithms often amplify even subtle correlations in superficially balanced datasets, an effect known as bias amplification [347, 135, 348].

Further complicating the problem, even an ideal training dataset devoid of any biases does *not* guarantee that the trained ML model will be bias-free [348]: model underspecification [137], model misspecification [349], and biological differences that render the prediction task easier or harder in different groups [350] may still introduce arbitrary biases. Similarly, Hooker [348] discusses the impact of seemingly innocent technical measures and design choices on the fairness of the resulting model. These include privacy-enhancing techniques, quantization, compression, as well as the choice of the learning rate and training length. To summarize, one can broadly distinguish two important technical causes for algorithms to be unfairly biased: biases present in the training data and biases introduced by the learning algorithm or the model [331, 348].

6.2 Data collection & curation measures to achieve fairness

Bias is pervasive in medical datasets [113]. Electronic health records are known to be biased due to their dependence on the patients' interactions with the healthcare system. As an example, Agniel et al. [114] found that the timing of laboratory tests had higher accuracy in predicting three-year survival than the actual test result. Similarly, biases due to national healthcare policies [115, 116], funding systems [118], and patient distance to treatment centers [117] have been demonstrated, while studies on the effect of the implicit racial bias of healthcare workers on clinical decision-making have yielded mixed results to date [351, 352, 351, 352]. There is also a particularly subtle form of bias to be expected once machine learning systems are used for clinical decision making: the algorithms used to make a decision may be biased, which then introduces these biases into the dataset used to train the next machine learning model [353, 354].

Given that observational medical databases are almost surely biased in one way or another, how can one hope to obtain a representative, ideally unbiased dataset? There are two primary solution approaches: performing targeted data collection so that the likelihood of bias occurrence is minimized and using synthetic data to either augment or completely replace human data. Suitably targeted data collection has been demonstrated in many applications to improve not only the fairness of the resulting classifier (e.g., by improving performance on minority groups) but also its overall accuracy [355, 356]. Aiming for a balanced dataset will surely involve targeting similar ratios of different genders, ethnicities, age groups, sensor types, and device manufacturers, but there are many more factors to consider depending on the particular application. Importantly, purely observational studies are often ill-suited for gathering a balanced dataset because of the biases introduced by the existing healthcare system, cf. above. Best practice guidelines on how best to gather balanced, representative datasets are currently sorely lacking [333], although there are example projects where this has been or is being attempted [355, 135, 146]. Unfortunately, it is impractical or even impossible to gather a sufficiently large and balanced dataset in many medical applications. In these cases, synthetic data generation methods, such as data reweighting schemes, artificial data modifications like cropping or rotation, and pathophysiological simulation models, represent an attractive alternative. These methods have been discussed in more detail in section 3.2.

Practice shows that dataset curation is very likely to be an iterative process, where some form of bias is detected either during testing or in the field, and as a remedy, the training or test datasets are adjusted to be unbiased in the observed regard [333]. However, while careful curation of the training and test datasets has been recognized as a key tool for fighting bias [333, 357], it is unlikely to fully remove all biases, and thus a successful bias mitigation strategy is likely to combine data curation techniques with algorithmic measures.

6.3 Algorithmic measures to achieve fairness

Proceeding to the algorithmic level, how can the likelihood of learning an unbiased model be increased? The first class of solution approaches is called *fair representation learning*, popularized by Zemel et al. [358]. The main idea is to find an intermediate representation of the data from which protected attributes can not be recovered but which retains the remaining information. Note that this is a much stronger requirement than merely removing protected features from a dataset because these can often still be recovered utilizing their correlations with other features [359]. The resulting sanitized dataset is then used as the input data for the actual prediction model. Recent work on fair representation learning is usually adversarial in nature, i.e., the correct representation of the original data samples is maximized while minimizing distributional differences between protected groups in the learned representation [360]. Fair representation learning is particularly amenable to the situation where a dataset is to be provided to third parties, and one wants to ensure that the learned models will likely be fair [361, 362]. If there are no third parties involved, fair representation learning may not be an appropriate solution [362].

Further advancing along the learning pipeline, the choice of an appropriate model structure clearly influences the likelihood of identifying a biased model. If a model correctly captures the potential sources of bias in a dataset, their relationships to different attributes, and the factors that can and cannot reasonably influence the prediction, it is much less likely to be biased. *Causal models* allow for an accurate description of these relationships and have therefore been proposed as a promising means to learn fair models from biased datasets [341, 363, 364]. This approach is not limited to simple white-box models based on prior knowledge, as various complex and highly flexible model structures that nevertheless incorporate assumptions about causality have been proposed in the literature [365, 363]. Besides enabling the learning of fair models from biased data, causal models have various other advantages for reliability and explainability, cf. sections 3.2 and 5.3.

Agnostic to the choice of a particular model structure, there is a large body of research on *algorithmic fairness* [331]. In this field, the general approach is to optimize a selected definition of fairness during model training. Popular fairness metrics include *statistical parity* (or *anti-classification*), where classification must be independent of protected attributes, *equalized odds* (or *classification parity*), which requires that common measures of estimator performance are equal across groups, and *calibration*, which requires that conditional on risk estimates, outcomes are group-independent [366, 367]. Many of these fairness metrics, as well as methods for enforcing them, are implemented in the open-source AI Fairness 360 (AIF) python package [368][28]; some of these metrics are also implemented in, e.g., the TensorFlow Fairness Indicators[29] and Google's "What-if Tool" [325].[26] It has been demonstrated that there are inherent problems with this approach, however. Corbett-Davies and Goel [366], Chouldechova [369] demonstrate that most of these fairness definitions are incompatible with one another, i.e., satisfying one fairness constraint implies violating another. Moreover, Pleiss et al. [370] demonstrate that these fairness definitions are in many ways not sufficient, and optimizing for one of these fairness metrics will often even lead to worse outcomes for members of minority classes. A more promising solution, especially in the medical context, has recently been proposed by Dwork et al. [367]. The authors propose the use of *decoupled classifiers*, i.e., using different classifiers for different groups, with the classifiers being designed to maximize performance in that group only. The use of transfer learning is proposed for leveraging the whole dataset even when training the decoupled classifier for a small minority group. Ustun et al. [339] build on this approach and also use decoupled classifiers. Additionally, however, they demonstrate that their learning method satisfies preference guarantees, i.e., the majority of each group will prefer the classifier assigned to that group to a) a pooled classifier that does not differentiate between groups and b) the classifiers assigned to all other groups. This appears to be a very promising approach. Recently, mirroring the rise of the causal inference field [198], various *causal* fairness definitions have been proposed [326, 371], attempting to solve the problems with the previously proposed statistical definitions mentioned above. Again, there are many different causal fairness notions, the most general of which is path-specific counterfactual fairness (PC fairness) [371]. The key advantage of these causal notions of fairness is that they can correctly adjust for confounding factors in the data due to, e.g., historical bias. Recently, Yan et al. [326] have developed an interactive tool that identifies the causal relationships learned by an ML model and enables users to interactively explore potential sources of unfairness in a dataset or model. In a user study, the tool was favorably evaluated compared to the AIF tool, with users mentioning the increased transparency of the identified causal graph and the interactivity of the tool as key reasons [326].

Finally, a number of *post-hoc methods* have been proposed to recalibrate or fine-tune an existing (biased) model in order to satisfy some fairness metric [372, 373, 374]. These methods represent an attractive solution in domains in which retraining a whole model is very costly, and many of them are available in the AIF tool [368]. They do, however, suffer from the same problems as the other, previously mentioned, methods that optimize a global model for some fairness definition [370].

6.4 Summary

Unfortunately, creating a fair ML model requires more work than picking an algorithmic fairness definition from the literature and choosing an appropriate learn-

[28] See https://aif360.mybluemix.net/.

[29] See https://www.tensorflow.org/tfx/guide/fairness_indicators.

ing procedure. Most practical problems are characterized by particular biases and complexities that require special care and are often not foreseeable in the early stages of a project [333]. There is no simple technical fix for the fairness problem [375]; the only real solution is to be aware of the problem in all stages of the development process, to be on the lookout for potential biases, and to gather feedback from diverse stakeholders. In a recent field experiment, dataset representativeness was found to be the strongest factor influencing the fairness of the resulting model [357]. On the algorithmic level, the causal lens currently appears particularly promising for learning fair models from biased data [341, 363, 131, 364] and evaluating the fairness of arbitrary models [371, 326]. Overall, a (reasonably) fair machine learning model will likely be a result of targeted data collection, resampling methods or synthetic data generation, a fairness-aware learning method, and monitoring for potential biases throughout the ML lifecycle. In practice, achieving algorithmic fairness will often be an iterative process [333, 326]: data are collected and curated, a model is selected and trained, and the model's fairness is evaluated using one or (ideally) multiple of the metrics discussed above. The model is found to be unfair in some way, adjustments to the dataset, the model, or the training method are made, and the procedure is repeated. There is an evident close connection to transparency and explainability here: fairness metrics and interactive tools for exploring the relationships learned by a model represent transparency and explainability tools tailored towards investigating the fairness of the learned model.

Algorithmic (un)fairness is a topic that will almost certainly increase in importance as machine learning systems further permeate all areas of society. Preventing biased algorithms requires a significant investment of resources and procedural support, both of which are currently not widely experienced by industry ML practitioners [333]. Development teams' blind spots may make it harder to detect (and mitigate) fairness risks early in the development process [333]; a problem which can be partially remedied by increasing diversity and inclusiveness. Unfair algorithms are already perceived by the public as one of the major problems with AI systems [376], and this perception is likely to be reinforced by future high-profile examples of unfair algorithms. It is thus in the interest of MML developers and manufacturers to be extremely cautious not to introduce unwanted biases in their systems. This need has recently also been noted by the FDA, which "recognizes the crucial importance for medical devices to be well suited for a racially and ethnically diverse intended patient population and the need for improved methodologies for the identification and improvement of machine learning algorithms. This includes methods for the identification and elimination of bias, and on the robustness and resilience of these algorithms to withstand changing clinical inputs and conditions." [23].

7 Discussion & summary

Medical machine learning (MML) applications must adhere to the same standards as classical health or medical device software. Medical device regulations demand following best practices regarding quality and risk management, performing comprehensive verification, validation, and post-market surveillance, as well as the use of state-of-the-art risk control and mitigation techniques. In particular, potentially unreliable, unsafe, insecure, not privacy-preserving, nontransparent, unexplainable, or discriminatory MML systems pose *risks* (to the patient, the hospital, the society at large) that must be mitigated by the use of appropriate technical (state-of-the-art) counter-measures. An ecosystem of technical standards on medical device software development details best practice development procedures that apply to MML systems as well, although none of these standards are currently specifically tailored towards ML applications. In addition, nondiscrimination law and data privacy regulations such as the GDPR impose further regulatory requirements. To summarize, existing regulations already demand some degree of safety, robustness, reliability, security, privacy, explainability, and fairness of MML systems. In the future, foreseeable regulatory changes and standards under development will further substantiate these requirements [29, 25, 26].

Much of what is often called "responsible machine learning" [377, 340, 43] is thus — to some degree, at least — already required by existing regulations. A notable exception from this pattern concerns *algorithmic impact assessments*; especially ones with a scope surpassing the impact on the immediate *data subject* (often: the patient) and including, among others, the impact on the hospital and its staff, the healthcare system, and society as a whole. Such broad algorithmic impact assessments appear highly desirable from a responsibility perspective [46, 47, 48, 49, 87, 40, 11] but are currently not required by regulations.

Unfortunately, there is currently a lack of precise regulatory guidance and technical best practice documents.[30] This lack is not coincidental, however. Practically all of the technical challenges discussed in this document are extremely young, and the theory and research fields devoted to their solution are still emerging. Compared to classical, hand-written health software, creating an ML model that satisfies various desirable requirements (and verifying that it really does) can be much more challenging due to the inscrutability of the training process, its complex dependency on the training data, and the complexity and opaqueness of the resulting model. Moreover, model underspecification [137], data scarcity [119, 6, 1, 120, 121], distribution shift [126, 127, 128, 129], and

[30] The recent guidance by the German Fraunhofer IAIS institute [53] is, to the authors' knowledge, the first comprehensive practical guidance for developers and auditors concerning trustworthy AI.

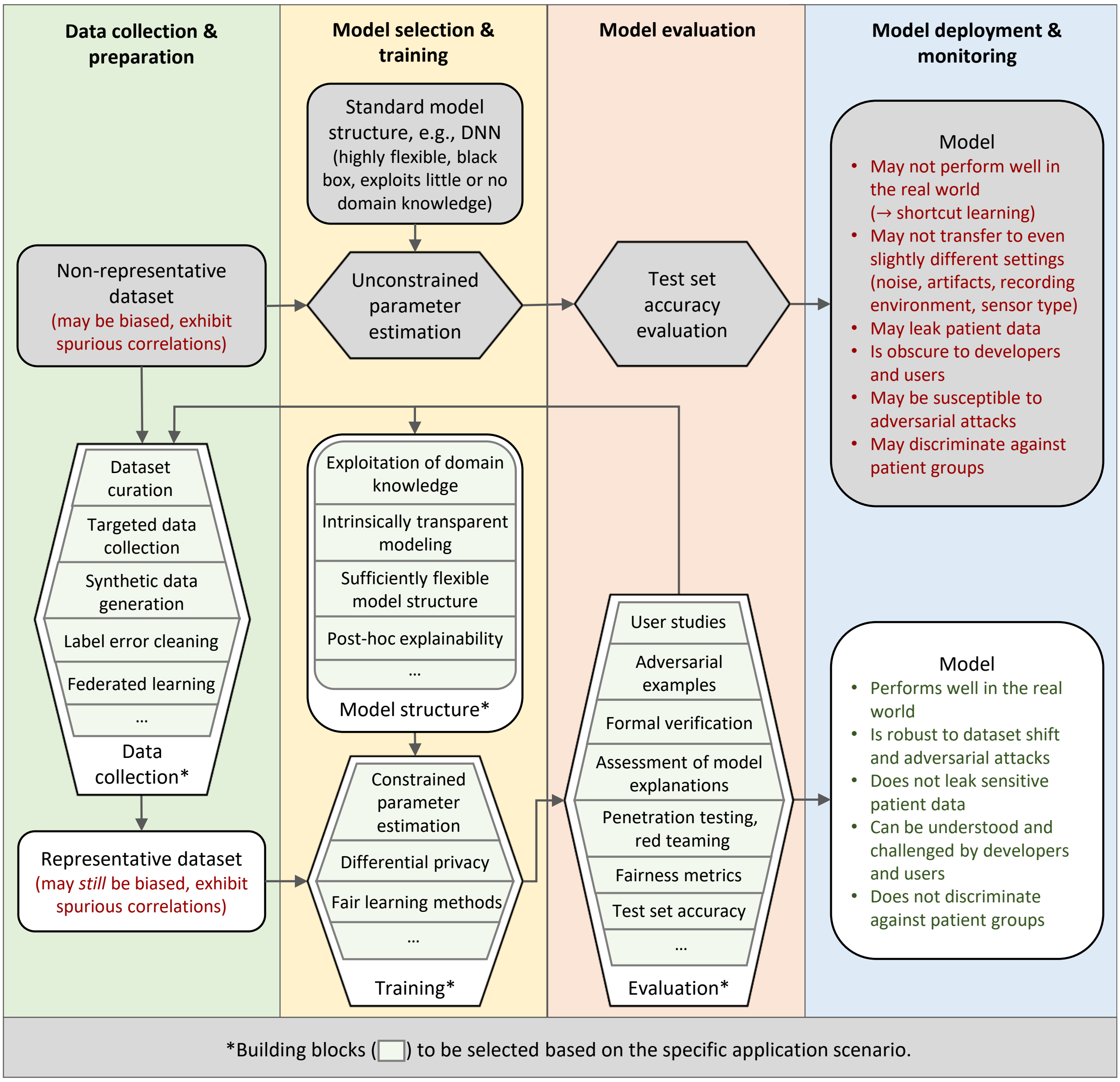
Data collection & preparation
Model selection & training
Model evaluation
Model deployment & monitoring
Standard model structure, e.g., DNN (highly flexible, black box, exploits little or no domain knowledge)
Non-representative dataset (may be biased, exhibit spurious correlations)
Unconstrained parameter estimation
Test set accuracy evaluation
Model
May not perform well in the real world (→ shortcut learning)
May not transfer to even slightly different settings (noise, artifacts, recording environment, sensor type)
May leak patient data
Is obscure to developers and users
May be susceptible to adversarial attacks
May discriminate against patient groups
Dataset curation
Targeted data collection
Synthetic data generation
Label error cleaning
Federated learning
...
Data collection*
Exploitation of domain knowledge
Intrinsically transparent modeling
Sufficiently flexible model structure
Post-hoc explainability
...
Model structure*
User studies
Adversarial examples
Formal verification
Assessment of model explanations
Penetration testing, red teaming
Fairness metrics
Test set accuracy
...
Evaluation*
Representative dataset (may *still* be biased, exhibit spurious correlations)
Constrained parameter estimation
Differential privacy
Fair learning methods
...
Training*
Model
Performs well in the real world
Is robust to dataset shift and adversarial attacks
Does not leak sensitive patient data
Can be understood and challenged by developers and users
Does not discriminate against patient groups
*Building blocks (□) to be selected based on the specific application scenario.

Figure 5: A visual summary of known important problems with the classical ML workflow as well as potential solutions to these challenges. The lower path is not meant to be pursued in its entirety; instead, the requirements in each specific application scenario must be analyzed carefully and the appropriate building blocks selected. Neither the list of problems nor the list of solution building blocks is considered comprehensive.

spurious correlations [15, 134, 13, 14, 135] pose serious challenges to overcome. They are very difficult to solve with black-box models that do not employ any domain knowledge, as evidenced by the increasingly large number of problematic incidents with ML systems [12]. While there is at the moment considerable ambiguity regarding the exact regulatory requirements as well as the auditing of said requirements by notified bodies [58], it is evident that MML developers and manufacturers must not neglect these crucial challenges. Several key strategies towards their solution emerge.

Firstly, the careful curation of large, realistic, and representative *datasets* is indispensable to achieve robust, reliable, and fair MML. In many instances, data curation will be an iterative process: once a problem with a model is noticed, appropriate example data are added to the training data or faulty and misleading examples are removed, and the model is retrained using this extended dataset. While various initiatives are attempting to gather large, openly available, and representative datasets, data scarcity remains a key challenge for MML due to technical, organizational, and legal reasons. Policy initiatives such as the proposed European health data space are an essential step to boost the availability of high-quality medical datasets. Similarly, inclusive consensus processes for the standardized validation and benchmarking of MML systems may play a vital role [378]. On the other hand, distributed learning schemes such as federated learning and split learning appear very promising for enabling the training of models on large and diverse datasets without requiring these data to ever leave their respective source institutions. Due to the increased number of participants in the training process, these methods require careful consideration of privacy and security risks, however. Furthermore, transfer learning, data augmentation, automated labeling, and synthetic data generation represent additional ways to boost dataset size despite a lack of real-world (labeled) data. However, the extent to which these techniques can help (and under which circumstances) is currently unclear.

Secondly, the careful exploitation of *domain knowledge* increases robustness and reliability without decreasing real-world accuracy. Domain knowledge can be exploited in various ways, e.g., by selecting an appropriate model structure or imposing hard constraints or soft priors on the model parameters. Modeling assumptions such as monotonicity, convexity, symmetry, and smoothness may already yield significantly improved robustness and reliability while retaining flexibility in all other regards [189, 190]. The introduction of *causal* assumptions into the modeling and estimation process appears particularly promising for a number of reasons. It is purported to enable learning dataset shift-stable models [131, 197, 198], identifying fair models from biased data [363], and increasing general robustness (also to adversarial examples) [197, 198, 136] and explainability to human users [316, 268]. The causal ML field is still in its infancy, and many technical challenges remain to be solved [197, 198], but this appears to be a very promising direction for future research.[31] Domain knowledge can also be exploited after the fact by formally verifying whether a trained model satisfies application-specific constraints [204, 205] or automatically generating safety-critical corner cases for testing the model [206].

Thirdly, *transparency and explainability* are not only (to some application-specific degree) legally required, but also instrumental properties that help maintain human agency in the face of algorithmic decision-making [11] and achieve robustness, reliability, and fairness of the MML system. Tools to achieve transparency and explainability include transparent communication of the dataset, its properties, influential instances and outliers, the choice of an interpretable model class, post-hoc explainability methods for generating explanations of black-box models, and algorithmic fairness metrics. If used well, they enable a developer to better judge and debug the system, an auditor to better assess the trustworthiness of a system, and a user to better understand the reasoning behind a particular decision of the system. To achieve these goals, however, transparency and explainability tools must always be tailored towards a particular user group and use case [275, 271], requiring early involvement of that user group in the development process: which information do they need? A particular kind of transparency — namely, *traceability* of individual decisions — may be required to enable *contestability* of the system, as is required by, e.g., the GDPR. Finally, care must be taken not to cause more harm than good with an explanation, since especially approximative post-hoc explanation methods such as LIME and saliency maps can be misleading [304, 281]. Where feasible, intrinsically interpretable models thus appear preferable, and such models have been proposed for many domains and have been shown to perform equally well as black-box models [130, 298, 296, 287, 193, 304, 194].

Fourth, and maybe most importantly, developing a responsible and regulatory conform MML system is not only a technical challenge; it is also an organizational and procedural challenge [11]. Truly responsible AI will never be the result of simply following some “ethics checklist” [39]. It will always require virtuous designers, engineers, and managers who are truly motivated to create MML systems that benefit patients, healthcare professionals, and the society at large. Technical guidelines and requirements are no replacement for an intense occupation of all involved with the underlying ethical and technical challenges [380, 39, 40, 11]. The focus must be both on a responsible *process* as well as on a responsible end result [39, 11]. To this end, the inclusion of all rele-

[31] The rise of causal inference has already had a significant impact on, e.g., epidemiological research in recent years [379].

vant stakeholders is indispensable [340, 11]: ML experts, clinicians, regulators, clinical managers, patients, and relatives should be involved in the development process. Moreover, some parts of this responsible innovation process can be formalized, by, e.g., invoking a red team responsible for breaking the MML system [221], standardizing algorithm audits [78], or aiming for an inclusive and transparent development process [11].

To summarize this discussion, fig. 5 illustrates several key problems that may arise using a classical ML workflow, as well as potential solutions to these challenges. Many of the aforementioned challenges and solution approaches become significantly more complex in the context of *dynamic* MML systems. By dynamic MML systems, we refer to systems that interact with a dynamic environment in a closed loop — such as a system controlling a patient's mechanical ventilation — and systems that learn continuously over time. How can we ensure, e.g., the maintained robustness, reliability, and fairness of a closed-loop or continuously learning system? Moreover, such systems raise additional critical challenges regarding the long-term impact of deploying them. These problems are closely related to the field of classical closed-loop control theory, which has investigated the properties of dynamic systems for decades. We will further discuss such dynamic MML systems in a future publication.

Lastly, the aim of this document is, of course, *not* to discourage developers and manufacturers from developing ML-based medical solutions, but to assist in doing so *responsibly* and while conforming with (current and future) regulations.

Acknowledgements

The authors would like to thank Jannik Prüßmann, B.Sc., for assisting with the formatting of the references.

Funding

All authors are affiliated with the KI-SIGS (Artificial Intelligence Space for Smart Health Systems) project,[32] which is funded by the German federal ministry for economic affairs and energy (BMWi) and involves several universities, research institutions and private companies in northern Germany. Researchers EP, YP, SB and LP are employed through funding from the KI-SIGS project. Neither the funding agency nor the involved private companies have had any influence on the planning and design of this survey, the writing of the manuscript, or the publication process.

Conflict of interest statement

EP and PR hold multiple patents with Dräger Medical. EM has received research funding by Volkswagen-Stiftung. All remaining authors declare no potential conflicts of interest.

References

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, J. T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Briefings in Bioinformatics 19 (2017) 1236–1246. doi:10.1093/bib/bbx044.

[2] J. Wiens, E. S. Shenoy, Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology, Clinical Infectious Diseases 66 (2017) 149–153. doi:10.1093/cid/cix731.

[3] V. H. Buch, I. Ahmed, M. Maruthappu, Artificial intelligence in medicine: current trends and future possibilities, British Journal of General Practice 68 (2018) 143–144. doi:10.3399/bjgp18x695213.

[4] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature Medicine 25 (2019) 24–29. doi:10.1038/s41591-018-0316-z.

[5] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. doi:10.1038/s41591-018-0300-7.

[6] A. S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Zeitschrift für Medizinische Physik 29 (2019) 102–127. doi:10.1016/j.zemedi.2018.11.002.

[7] W. Tao, A. N. Concepcion, M. Vianen, A. C. A. Marijnissen, F. P. G. J. Lafeber, T. R. D. J. Radstake, A. Pandit, Multiomics and machine learning accurately predict clinical response to adalimumab and etanercept therapy in patients with rheumatoid arthritis, Arthritis & Rheumatology 73 (2020) 212–222. doi:10.1002/art.41516.

[8] E. Mlodzinski, D. J. Stone, L. A. Celi, Machine learning for pulmonary and critical care medicine: A narrative review, Pulmonary Therapy 6 (2020) 67–77. doi:10.1007/s41030-020-00110-z.

[9] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, npj Digital Medicine 4 (2021) 4. doi:10.1038/s41746-020-00372-6.

[10] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine 17 (2019). doi:10.1186/s12916-019-1426-2.

[11] Ethics and governance of artificial intelligence for health: WHO guidance, World Health Organization (WHO), Geneva, Switzerland, 2021.

[12] S. McGregor, Preventing repeated real world AI failures by cataloging incidents: The AI incident database (2020). arXiv:2011.08512.

[13] M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha,

[32] See https://ki-sigs.de/.

T. M. Snyder, J. T. Dudley, Deep learning predicts hip fracture using confounding patient and healthcare variables, npj Digital Medicine 2 (2019) 31. doi:10.1038/s41746-019-0105-1.

[14] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, E. K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, PLOS Medicine 15 (2018) e1002683. doi:10.1371/journal.pmed.1002683.

[15] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, H. A. Haenssle, Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition, JAMA Dermatology 155 (2019) 1135–1141. doi:10.1001/jamadermatol.2019.1735.

[16] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, M. Ghassemi, CheXclusion: Fairness gaps in deep chest X-ray classifiers, in: Biocomputing 2021, World Scientific, 2020, pp. 232–243. doi:10.1142/9789811232701_0022.

[17] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453. doi:10.1126/science.aax2342.

[18] D. B. Larson, H. Harvey, D. L. Rubin, N. Irani, J. R. Tse, C. P. Langlotz, Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: Summary and recommendations, Journal of the American College of Radiology 18 (2020) 413–424. doi:10.1016/j.jacr.2020.09.060.

[19] M. D. McCradden, S. Joshi, J. A. Anderson, M. Mazwi, A. Goldenberg, R. Z. Shaul, Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning, Journal of the American Medical Informatics Association 27 (2020) 2024–2027. doi:10.1093/jamia/ocaa085.

[20] T. Minssen, S. Gerke, M. Aboy, N. Price, G. Cohen, Regulatory responses to medical machine learning, Journal of Law and the Biosciences 7 (2020) lsaa002. doi:10.1093/jlb/lsaa002.

[21] S. Benjamens, P. Dhunnoo, B. Meskó, The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database, npj Digital Medicine 3 (2020) 118. doi:10.1038/s41746-020-00324-0.

[22] M. Ryan, B. C. Stahl, Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, Journal of Information, Communication and Ethics in Society 19 (2021) 61–68. doi:10.1108/jices-12-2019-0138.

[23] Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan, Food and Druck administration (FDA), US, 2021. URL: https://www.fda.gov/media/145022/download.

[24] P. Cihon, Standards for AI governance: international standards to enable global coordination in AI research & development, Future of Humanity Institute, University of Oxford (2019).

[25] IEEE P7000 family of standards for ethically aligned autonomous and intelligent systems, Institute of Electrical and Electronics Engineers (IEEE), under development. URL: https://ethicsinaction.ieee.org/p7000/.

[26] ISO/IEC JTC 1/SC 42 Artificial intelligence, Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence, International Organization for Standardization (ISO), 2020.

[27] ISO/IEC JTC 1/SC 42 Artificial intelligence, ISO/IEC DIS 23053 – Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), International Organization for Standardization (ISO), under development.

[28] Normungsroadmap KI, German Institute for Standardization (DIN), Germany, 2020. URL: https://www.dke.de/resource/blob/2008010/0c29125fa99ac4c897e2809c8ab343ff/nr-ki-deutsch---download-data.pdf.

[29] European Commission, Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[30] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, European Commission, 2019. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[31] L. Floridi, Establishing the rules for building trustworthy AI, Nature Machine Intelligence 1 (2019) 261–262. doi:10.1038/s42256-019-0055-y.

[32] OECD/LEGAL/0449 – Recommendation of the Council on Artificial Intelligence, Organisation for Economic Cooperation and Development (OECD), 2019.

[33] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, Minds and Machines 28 (2018) 689–707. doi:10.1007/s11023-018-9482-5.

[34] L. Floridi, J. Cowls, A unified framework of five principles for AI in society, Harvard Data Science Review 1 (2019). doi:10.1162/99608f92.8cd550d1.

[35] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. doi:10.1038/s42256-019-0088-2.

[36] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices, Science and Engineering Ethics 26 (2020) 2141–2168. doi:10.1007/s11948-019-00165-5.

[37] S. Bowers, D. Cohen, How lobbying blocked European safety checks for dangerous medical implants, BMJ (2018) k4999. doi:10.1136/bmj.k4999.

[38] G. Rieder, J. Simon, P.-H. Wong, Mapping the stony road toward trustworthy AI: Expectations, problems, conundrums, in: M. Pelillo, T. Scantamburlo (Eds.), Machines

We Trust: Perspectives on Dependable AI, MIT Press, Cambridge, MA, US, 2020. doi:10.2139/ssrn.3717451.

[39] V. Dignum, Responsible artificial intelligence: How to develop and use AI in a responsible way, Springer Nature, Cham, Switzerland, 2019.

[40] M. C. Elish, E. A. Watkins, Repairing Innovation. A Study of Integrating AI in Clinical Care, Data & Society, New York, NY, US, 2020. URL: https://datasociety.net/library/repairing-innovation/.

[41] D. Acemoglu, P. Restrepo, The wrong kind of AI? artificial intelligence and the future of labour demand, Cambridge Journal of Regions, Economy and Society 13 (2019) 25–35. doi:10.1093/cjres/rsz022.

[42] M. Coeckelbergh, AI Ethics, MIT Press, Cambridge, MA, US, 2020.

[43] D. S. Char, M. D. Abràmoff, C. Feudtner, Identifying ethical considerations for machine learning healthcare applications, The American Journal of Bioethics 20 (2020) 7–17. doi:10.1080/15265161.2020.1819469.

[44] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, L. Floridi, The ethics of AI in health care: A mapping review, Social Science & Medicine 260 (2020) 113172. doi:10.1016/j.socscimed.2020.113172.

[45] Redesigning AI for Shared Prosperity: an Agenda, Partnership on AI, 2021. URL: https://www.partnershiponai.org/shared-prosperity-initiative/.

[46] D. Reisman, J. Schultz, K. Crawford, M. Whittaker, Algorithmic impact assessments: A practical framework for public agency accountability, AI Now Institute, New York, NY, US, 2018. URL: https://ainowinstitute.org/aiareport2018.pdf.

[47] Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, first edition ed., The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019. URL: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

[48] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, M. C. Elish, Algorithmic impact assessments and accountability, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (ACM), New York, NY, US, 2021, pp. 735–746. doi:10.1145/3442188.3445935.

[49] R. A. Calvo, D. Peters, S. Cave, Advancing impact assessment for intelligent systems, Nature Machine Intelligence 2 (2020) 89–91. doi:10.1038/s42256-020-0151-z.

[50] M. E. Kaminski, G. Malgieri, Multi-layered explanations from algorithmic impact assessments in the GDPR, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (ACM), New York, NY, US, 2020, pp. 68–79. doi:10.1145/3351095.3372875.

[51] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, M. Lennon, Principles to practices for responsible AI: Closing the gap (2020). arXiv:2006.04707.

[52] Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), Food and Druck administration (FDA), US, 2019. URL: https://www.fda.gov/media/122535/download.

[53] M. Poretschkin, S. Houben, A. Schmitz, M. Mock, M. Akila, J. Rosenzweig, L. Adilova, J. Sicking, D. Becker, E. Schulz, A. B. Cremers, A. Voss, D. Hecker, S. Wrobel, Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS), 2021. URL: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf.

[54] P.-Y. Benhamou, S. Franc, Y. Reznik, C. Thivolet, P. Schaepelynck, E. Renard, B. Guerci, L. Chaillous, C. Lukas-Croisier, N. Jeandidier, H. Hanaire, S. Borot, M. Doron, P. Jallon, I. Xhaard, V. Melki, L. Meyer, B. Delemer, M. Guillouche, L. Schoumacker-Ley, A. Farret, D. Raccah, S. Lablanche, M. Joubert, A. Penfornis, G. Charpentier, Closed-loop insulin delivery in adults with type 1 diabetes in real-life conditions: a 12-week multicentre, open-label randomised controlled crossover trial, The Lancet Digital Health 1 (2019) e17–e25. doi:10.1016/s2589-7500(19)30003-2.

[55] I. Fox, J. Lee, R. Pop-Busui, J. Wiens, Deep reinforcement learning for closed-loop blood glucose control, in: F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), Proceedings of the 5th Machine Learning for Healthcare Conference, volume 126 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 508–536. URL: http://proceedings.mlr.press/v126/fox20a.html.

[56] B. Zhu, U. Shin, M. Shoaran, Closed-loop neural interfaces with embedded machine learning (2020) 1–4. doi:10.1109/ICECS49266.2020.9294844.

[57] Regulation of Artificial Intelligence in Selected Jurisdictions, The Law Library of Congress, Global Legal Research Directorate, 2019. URL: https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf.

[58] U. J. Muehlematter, P. Daniore, K. N. Vokinger, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis, The Lancet Digital Health 3 (2021) e195–e203. doi:10.1016/s2589-7500(20)30292-2.

[59] N. Aisu, M. Miyake, K. Takeshita, M. Akiyama, R. Kawasaki, K. Kashiwagi, T. Sakamoto, T. Oshika, A. Tsujikawa, Regulatory-approved deep learning/machine learning-based medical devices in japan as of 2020: A systematic review (2021). doi:10.1101/2021.02.19.21252031.

[60] R. Beckers, Z. Kwade, F. Zanca, The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics, Physica Medica 83 (2021) 1–8. doi:10.1016/j.ejmp.2021.02.011.

[61] J. Meszaros, M. Corrales Compagnucci, T. Minssen, The interaction of the medical device regulation and the GDPR: Do European rules on privacy and scientific research impair the safety & performance of AI medical devices?, in: I. G. Cohen, T. Minssen, W. N. Price II, C. Robertson, C. Shachar (Eds.), The Future of Medi-

cal Device Regulation: Innovation and Protection, Cambridge University Press, Cambridge, MA, US, 2021. URL: https://ssrn.com/abstract=3808384.

[62] C. Johner, Validation of machine learning libraries, 2020. URL: https://www.johner-institute.com/articles/software-iec-62304/and-more/validation-of-machine-learning-libraries/, last accessed: July 5, 2021.

[63] C. Johner, Regulatory requirements for medical devices with machine learning, 2020. URL: https://www.johner-institute.com/articles/regulatory-affairs/and-more/regulatory-requirements-for-medical-devices-with-machine-learning/, last accessed: July 5, 2021.

[64] IEC 62304:2006. Medical device software – Software life cycle processes, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2006.

[65] IEC 82304-1:2016. Health software – Part 1: General requirements for product safety, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2016.

[66] IEC 60601-1:2005. Medical electrical equipment – Part 1: General requirements for basic safety and essential performance, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2005.

[67] IMDRF Software as a Medical Device (SaMD) Working Group, Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations, International Medical Device Regulators Forum (IMDRF), 2014. URL: http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf.

[68] ISO 13485:2016. Medical devices – Quality management systems – Requirements for regulatory purposes, International Organization for Standardization (ISO), Geneva, Switzerland, 2016.

[69] Code of federal regulations (CFR) title 21 part 820. Medical devices – quality system regulation, Food and Drug Administration (FDA), US, 2020.

[70] G. Bos, ISO 13485:2003/2016—medical devices—quality management systems—requirements for regulatory purposes, in: J. Wong, R. K. Y. Tong (Eds.), Handbook of Medical Device Regulatory Affairs in Asia, second edition ed., Pan Stanford Publishing, 2018, pp. 153–174. doi:10.1201/9780429504396.

[71] ISO 14971:2019. Medical devices – Application of risk management to medical devices, International Organization for Standardization (ISO), Geneva, Switzerland, 2019.

[72] T. Chan, R. K. Y. Tong, ISO 14971: Application of risk management to medical devices, in: J. Wong, R. K. Y. Tong (Eds.), Handbook of Medical Device Regulatory Affairs in Asia, second edition ed., Pan Stanford Publishing, 2018, pp. 175–191. doi:10.1201/9780429504396.

[73] IEC 62366-1:2015. Medical devices – Part 1: Application of usability engineering to medical devices, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2015.

[74] Regulation (EU) 2017/745 (EU MDR) on medical devices, The European parliament and the council of the European union (EU), 2017.

[75] IMDRF Software as a Medical Device (SaMD) Working Group, Software as a Medical Device (SaMD): Clinical Evaluation, International Medical Device Regulators Forum (IMDRF), 2016. URL: http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-samd-ce.pdf.

[76] Postmarket Surveillance Under Section 522 of the Federal Food, Drug, and Cosmetic Act. Guidance for Industry and Food and Drug Administration Staff, Food and Drug Administration (FDA), US, 2016. URL: https://www.fda.gov/media/81015/download.

[77] ISO/TR 20416:2020. Medical devices – Post-market surveillance for manufacturers, International Organization for Standardization (ISO), Geneva, Switzerland, 2020.

[78] UK Information Commissioner's Office, Guidance on the AI auditing framework, 2020. URL: https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

[79] MEDDEV 2.7/1 revision 4. Guidelines on Medical Devices – Clinical Evaluation: A Guide for Manufacturers and Notified Bodies under Directives 93/42/EEC and 90/385/EEC, European Commission, 2016.

[80] C. Johner, Transfer learning in medical devices: Regulatory recklessness or an ethical necessity?, 2021. URL: https://johner-institute.com/articles/software-iec-62304/and-more/transfer-learning-in-medical-devices/, last accessed: July 5, 2021.

[81] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Magazine 38 (2017) 50–57. doi:10.1609/aimag.v38i3.2741.

[82] A. Aler Tubella, A. Theodorou, V. Dignum, L. Michael, Contestable black boxes, in: Rules and Reasoning, Springer International Publishing, Cham, Switzerland, 2020, pp. 159–167. doi:10.1007/978-3-030-57977-7_12.

[83] C. Ryngaert, M. Taylor, The GDPR as global data protection regulation?, AJIL Unbound 114 (2020) 5–9. doi:10.1017/aju.2019.80.

[84] A. D. Selbst, J. Powles, Meaningful information and the right to explanation, International Data Privacy Law 7 (2017) 233–242. doi:10.1093/idpl/ipx022.

[85] L. Edwards, M. Veale, Slave to the algorithm? Why a right to explanation is probably not the remedy you are looking for, Duke Law & Technology Review 16 (2017) 18–84. doi:10.2139/ssrn.2972855.

[86] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable AI systems for the medical domain? (2017). arXiv:1712.09923.

[87] G. Malgieri, Automated decision-making in the EU member states: The right to explanation and other "suitable safeguards" in the national legislations, Computer Law & Security Review 35 (2019) 105327. doi:10.1016/j.clsr.2019.05.002.

[88] UK Information Commissioner's Office, The Alan Turing Institute, Explaining decisions made with AI, 2020. URL:

https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/.

[89] D. Baldini, Article 22 GDPR and prohibition of discrimination. An outdated provision?, 2019. URL: https://www.cyberlaws.it/2019/article-22-gdpr-and-prohibition-of-discrimination-an-outdated-provision/, last accessed: July 5, 2021.

[90] L. Rocher, J. M. Hendrickx, Y.-A. de Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications 10 (2019). doi:10.1038/s41467-019-10933-3.

[91] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, M. J. Cardoso, The future of digital health with federated learning, npj Digital Medicine 3 (2020) 119. doi:10.1038/s41746-020-00323-1.

[92] G. A. Kaissis, M. R. Makowski, D. Rückert, R. F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, Nature Machine Intelligence 2 (2020) 305–311. doi:10.1038/s42256-020-0186-1.

[93] F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R. T. H. Leijenaar, A. Jochems, B. Miraglio, D. Townend, P. Lambin, Systematic review of privacy-preserving distributed machine learning from federated databases in health care, JCO Clinical Cancer Informatics 4 (2020) 184–200. doi:10.1200/cci.19.00047.

[94] European Commission, White paper. On artificial intelligence — a European approach to excellence and trust, 2020. URL: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[95] I. G. Cohen, T. Evgeniou, S. Gerke, T. Minssen, The European artificial intelligence strategy: implications and challenges for digital health, The Lancet Digital Health 2 (2020) e376–e379. doi:10.1016/s2589-7500(20)30112-6.

[96] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, L. Floridi, The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation, AI & SOCIETY 36 (2020) 59–77. doi:10.1007/s00146-020-00992-2.

[97] IEEE P2801 – Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence, Institute of Electrical and Electronics Engineers (IEEE), under development. URL: https://standards.ieee.org/project/2801.html.

[98] IEEE P2802 – Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology, Institute of Electrical and Electronics Engineers (IEEE), under development. URL: https://standards.ieee.org/project/2802.html.

[99] Enquete-Kommission Künstliche Intelligenz, Bericht der Enquete-Kommission Künstliche Intelligenz — Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, Deutscher Bundestag, 2020. URL: https://dserver.bundestag.de/btd/19/237/1923700.pdf.

[100] IEC 61508:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2010.

[101] A. Morikawa, Y. Matsubara, Safety design concepts for statistical machine learning components toward accordance with functional safety standards (2020). arXiv:2008.01263.

[102] M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, J. Törnqvist, Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry, Journal of Automotive Software Engineering 1 (2019) 1. doi:10.2991/jase.d.190131.001.

[103] IEEE 610.12-1990 - IEEE Standard Glossary of Software Engineering Terminology, Institute of Electrical and Electronics Engineers (IEEE), 1990.

[104] R. Billinton, R. N. Allan, Reliability Evaluation of Engineering Systems, Springer US, 1992. doi:10.1007/978-1-4899-0685-4.

[105] M. Daszykowski, K. Kaczmarek, Y. V. Heyden, B. Walczak, Robust statistics in data analysis — a review, Chemometrics and Intelligent Laboratory Systems 85 (2007) 203–219. doi:10.1016/j.chemolab.2006.06.016.

[106] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, X. Yi, A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Computer Science Review 37 (2020) 100270. doi:10.1016/j.cosrev.2020.100270.

[107] I. Goodfellow, P. McDaniel, N. Papernot, Making machine learning robust against adversarial inputs, Communications of the ACM 61 (2018) 56–66. doi:10.1145/3134599.

[108] S. Saria, A. Subbaswamy, Tutorial: Safe and reliable machine learning, 2019. URL: https://arxiv.org/abs/1904.07204.

[109] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, M. P. Lungren, Preparing medical imaging data for machine learning, Radiology 295 (2020) 4–15. doi:10.1148/radiol.2020192224.

[110] K. T. Aakre, C. D. Johnson, Plain-radiographic image labeling: A process to improve clinical outcomes, Journal of the American College of Radiology 3 (2006) 949–953. doi:10.1016/j.jacr.2006.07.005.

[111] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Medical Image Analysis 65 (2020-10) 101759. doi:10.1016/j.media.2020.101759.

[112] L. Oakden-Rayner, Exploring large-scale public medical image datasets, Academic Radiology 27 (2020) 106–112. doi:10.1016/j.acra.2019.10.006.

[113] D. R. Williams, S. A. Mohammed, J. Leavell, C. Collins, Race, socioeconomic status, and health: Complexities, ongoing challenges, and research opportunities, Annals

of the New York Academy of Sciences 1186 (2010) 69–101. doi:10.1111/j.1749-6632.2009.05339.x.

[114] D. Agniel, I. S. Kohane, G. M. Weber, Biases in electronic health record data due to processes within the healthcare system: retrospective observational study, BMJ (2018) k1479. doi:10.1136/bmj.k1479.

[115] Y. Kitagawa, M. Namiki, Prostate-specific antigen based population screening for prostate cancer: current status in japan and future perspective in asia, Asian Journal of Andrology 17 (20142015) 475–480. doi:10.4103/1008-682x.143756.

[116] W. Levinson, M. Kallewaard, R. S. Bhatia, D. Wolfson, S. Shortt, E. A. Kerr, 'Choosing Wisely': a growing international campaign, BMJ Quality & Safety 24 (2014) 167–174. doi:10.1136/bmjqs-2014-003821.

[117] C. Kelly, C. Hulme, T. Farragher, G. Clarke, Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? a systematic review, BMJ Open 6 (2016) e013059. doi:10.1136/bmjopen-2016-013059.

[118] K. McLintock, A. M. Russell, S. L. Alderson, R. West, A. House, K. Westerman, R. Foy, The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis, BMJ Open 4 (2014) e005178. doi:10.1136/bmjopen-2014-005178.

[119] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, C. S. Greene, Opportunities and obstacles for deep learning in biology and medicine, Journal of The Royal Society Interface 15 (2018) 20170387. doi:10.1098/rsif.2017.0387.

[120] Z. Obermeyer, E. J. Emanuel, Predicting the future — big data, machine learning, and clinical medicine, New England Journal of Medicine 375 (2016) 1216–1219. doi:10.1056/NEJMp1606181.

[121] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, K. D. Mandl, Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency, npj Digital Medicine 3 (2020) 47. doi:10.1038/s41746-020-0254-2.

[122] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, A. Schwaighofer, Dataset shift in machine learning, MIT Press, Cambridge, MA, US, 2009.

[123] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern Recognition 45 (2012) 521–530. doi:10.1016/j.patcog.2011.06.019.

[124] A. Qayyum, J. Qadir, M. Bilal, A. I. Al-Fuqaha, Secure and robust machine learning for healthcare: A survey, IEEE Reviews in Biomedical Engineering 14 (2021) 156–180. doi:10.1109/RBME.2020.3013489.

[125] A. van Opbroek, M. A. Ikram, M. W. Vernooij, M. de Bruijne, Transfer learning improves supervised image segmentation across imaging protocols, IEEE Transactions on Medical Imaging 34 (2015) 1018–1030. doi:10.1109/tmi.2014.2366792.

[126] B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do ImageNet classifiers generalize to ImageNet?, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 5389–5400. URL: http://proceedings.mlr.press/v97/recht19a.html.

[127] A. Azulay, Y. Weiss, Why do deep convolutional networks generalize so poorly to small image transformations?, Journal of Machine Learning Research 20 (2019) 1–25. URL: http://jmlr.org/papers/v20/19-519.html.

[128] M. A. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W.-S. Ku, A. Nguyen, Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4845–4854. doi:10.1109/CVPR.2019.00498.

[129] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, L. Schmidt, Measuring robustness to natural distribution shifts in image classification, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 18583–18599. URL: https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.

[130] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery (ACM), New York, NY, US, 2015, pp. 1721–1730. doi:10.1145/2783258.2788613.

[131] A. Subbaswamy, S. Saria, From development to deployment: dataset shift, causality, and shift-stable models in health AI, Biostatistics (2019). doi:10.1093/biostatistics/kxz041.

[132] S. Sathitratanacheewin, P. Sunanta, K. Pongpirul, Deep learning for automated classification of tuberculosis-related chest X-ray: dataset distribution shift limits diagnostic performance generalizability, Heliyon 6 (2020) e04614. doi:10.1016/j.heliyon.2020.e04614.

[133] D. C. Castro, I. Walker, B. Glocker, Causality matters in medical imaging, Nature Communications 11 (2020). doi:10.1038/s41467-020-17478-w.

[134] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673. doi:10.1038/s42256-020-00257-z.

[135] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, V. Ordonez, Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5310–5319. doi:10.1109/

ICCV.2019.00541.

[136] Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, K. Zhang, Adversarial robustness through the lens of causality (2021). arXiv:2106.06196.

[137] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., Underspecification presents challenges for credibility in modern machine learning (2020). arXiv:2011.03395.

[138] A. G. Wilson, P. Izmailov, Bayesian deep learning and a probabilistic perspective of generalization, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 4697–4708. URL: https://proceedings.neurips.cc/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf.

[139] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, M. Mitchell, Towards accountability for machine learning datasets, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (ACM), New York, NY, US, 2021, pp. 560–575. doi:10.1145/3442188.3445918.

[140] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (ACM), New York, NY, US, 2019, pp. 220–229. doi:10.1145/3287560.3287596.

[141] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Systems with Applications 73 (2017) 220–239. doi:10.1016/j.eswa.2016.12.035.

[142] S. Shilaskar, A. Ghatol, P. Chatur, Medical decision support system for extremely imbalanced datasets, Information Sciences 384 (2017) 205–219. doi:10.1016/j.ins.2016.08.077.

[143] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, S. Sakr, Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford exercIse Testing (FIT) project, PloS one 12 (2017) e0179805. doi:10.1371/journal.pone.0179805.

[144] K. Hirano, G. W. Imbens, Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization, Health Services and Outcomes research methodology 2 (2001) 259–278. doi:10.1023/A:1020371312283.

[145] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, Journal of Statistical Planning and Inference 90 (2000) 227–244. doi:10.1016/s0378-3758(00)00115-4.

[146] T. A. R. P. Investigators, The "all of us" research program, New England Journal of Medicine 381 (2019) 668–676. doi:10.1056/nejmsr1809937.

[147] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology 10 (2019) 1–19. doi:10.1145/3298981.

[148] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Processing Magazine 37 (2020) 50–60. doi:10.1109/msp.2020.2975749.

[149] O. Gupta, R. Raskar, Distributed learning of deep neural network over multiple agents, Journal of Network and Computer Applications 116 (2018) 1–8. doi:10.1016/j.jnca.2018.05.003.

[150] Y. Gao, M. Kim, S. Abuadbba, Y. Kim, C. Thapa, K. Kim, S. A. Camtep, H. Kim, S. Nepal, End-to-end evaluation of federated learning and split learning for internet of things, in: Proceedings of the 39th International Symposium on Reliable Distributed Systems, IEEE, 2020, pp. 91–100. doi:10.1109/srds51746.2020.00017.

[151] V. Turina, Z. Zhang, F. Esposito, I. Matta, Combining split and federated architectures for efficiency and privacy in deep learning, in: Proceedings of the 16th International Conference on emerging Networking EXperiments and Technologies, Association for Computing Machinery (ACM), New York, NY, US, 2020, pp. 562–563. doi:10.1145/3386367.3431678.

[152] C. Thapa, M. A. P. Chamikara, S. Camtepe, SplitFed: When federated learning meets split learning (2020). arXiv:2004.12088.

[153] M. G. Poirot, P. Vepakomma, K. Chang, J. Kalpathy-Cramer, R. Gupta, R. Raskar, Split learning for collaborative deep learning in healthcare (2019). arXiv:1912.12115.

[154] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, F. Wang, Federated learning for healthcare informatics, Journal of Healthcare Informatics Research 5 (2020) 1–19. doi:10.1007/s41666-020-00082-4.

[155] A. Moncada-Torres, F. Martin, M. Sieswerda, J. van Soest, G. Geleijnse, VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange, in: AMIA Annual Symposium Proceedings, 2020, pp. 870–877.

[156] G. A. Reina, A. Gruzdev, P. Foley, O. Perepelkina, M. Sharma, I. Davidyuk, I. Trushkin, M. Radionov, A. Mokrov, D. Agapov, J. Martin, B. Edwards, M. J. Sheller, S. Pati, P. N. Moorthy, H. S. Wang, P. Shah, S. Bakas, OpenFL: An open-source framework for federated learning (2021). arXiv:2105.06413.

[157] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, P. P. B. de Gusmão, N. D. Lane, Flower: A friendly federated learning research framework (2021). arXiv:2007.14390.

[158] IEEE 3652.1-2020 – IEEE Guide for Architectural Framework and Application of Federated Machine Learning, Institute of Electrical and Electronics Engineers (IEEE), 2021. URL: https://standards.ieee.org/standard/3652_1-2020.html.

[159] J. E. van Engelen, H. H. Hoos, A survey on semi-supervised learning, Machine Learning 109 (2019) 373–440. doi:10.1007/s10994-019-05855-6.

[160] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, D. Rueckert, Semi-supervised learning for network-based cardiac MR image segmentation, in: Lecture Notes in Computer Sci-

ence, Springer International Publishing, 2017, pp. 253–260. doi:10.1007/978-3-319-66185-8_29.

[161] H. M. Trivedi, M. Panahiazar, A. Liang, D. Lituiev, P. Chang, J. H. Sohn, Y.-Y. Chen, B. L. Franc, B. Joe, D. Hadley, Large scale semi-automated labeling of routine free-text clinical records for deep learning, Journal of Digital Imaging 32 (2018) 30–37. doi:10.1007/s10278-018-0105-8.

[162] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. Chapman, T. Amrhein, D. Mong, D. L. Rubin, O. Farri, M. P. Lungren, Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification, Artificial Intelligence in Medicine 97 (2019) 79–88. doi:10.1016/j.artmed.2018.11.004.

[163] P. H. Yi, A. Lin, J. Wei, A. C. Yu, H. I. Sair, F. K. Hui, G. D. Hager, S. C. Harvey, Deep-learning-based semantic labeling for 2D mammography and comparison of complexity for machine learning tasks, Journal of Digital Imaging 32 (2019) 565–570. doi:10.1007/s10278-019-00244-w.

[164] C.-K. Shie, C.-H. Chuang, C.-N. Chou, M.-H. Wu, E. Y. Chang, Transfer representation learning for medical image analysis, in: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2015, pp. 711–714. doi:10.1109/embc.2015.7318461.

[165] P. Alirezazadeh, B. Hejrati, A. Monsef-Esfahani, A. Fathi, Representation learning-based unsupervised domain adaptation for classification of breast cancer histopathology images, Biocybernetics and Biomedical Engineering 38 (2018) 671–683. doi:10.1016/j.bbe.2018.04.008.

[166] T. Bai, S. Zhang, B. L. Egleston, S. Vucetic, Interpretable representation learning for healthcare via capturing disease progression through time, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery (ACM), New York, NY, US, 2018, pp. 43–51. doi:10.1145/3219819.3219904.

[167] S. Latif, M. Asim, M. Usman, J. Qadir, R. Rana, Automating motion correction in multishot MRI using generative adversarial networks, in: NeurIPS 2018 Workshop 'Medical Imaging meets NeurIPS', 2018. URL: https://arxiv.org/abs/1811.09750.

[168] C. S. Perone, P. Ballester, R. C. Barros, J. Cohen-Adad, Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, NeuroImage 194 (2019) 1–1. doi:10.1016/j.neuroimage.2019.03.026.

[169] W. M. Kouw, M. Loog, A review of domain adaptation without target labels, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2019) 766–785. doi:10.1109/TPAMI.2019.2945942.

[170] A. Choudhary, L. Tong, Y. Zhu, M. D. Wang, Advancing medical imaging informatics by deep learning-based domain adaptation, Yearbook of Medical Informatics 29 (2020) 129–138. doi:10.1055/s-0040-1702009.

[171] S. I. Nikolenko, Synthetic Data for Deep Learning, Springer, Cham, Switzerland, 2021. doi:https://doi.org/10.1007/978-3-030-75178-4.

[172] A. Fawzi, H. Samulowitz, D. S. Turaga, P. Frossard, Adaptive data augmentation for image classification, in: Proceedings of the 23rd IEEE International Conference on Image Processing, 2016, pp. 3688–3692. doi:10.1109/ICIP.2016.7533048.

[173] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953.

[174] A. Fernandez, S. Garcia, F. Herrera, N. V. Chawla, SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, Journal of Artificial Intelligence Research 61 (2018) 863–905. doi:10.1613/jair.1.11192.

[175] E. E. Berkson, J. D. VanCor, S. Esposito, G. Chern, M. Pritt, Synthetic data generation to mitigate the low/no-shot problem in machine learning, in: Proceedings of the 4th IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2019, pp. 1–7. doi:10.1109/aipr47015.2019.9174596.

[176] T. M. Deist, A. Patti, Z. Wang, D. Krane, T. Sorenson, D. Craft, Simulation-assisted machine learning, Bioinformatics 35 (2019) 4072–4080. doi:10.1093/bioinformatics/btz199.

[177] C. Esteban, S. L. Hyland, G. Rätsch, Real-valued (medical) time series generation with recurrent conditional GANs (2017). arXiv:1706.02633.

[178] S. Dash, R. Dutta, I. Guyon, A. Pavao, A. Yale, K. P. Bennett, Synthetic event time series health data generation, in: NeurIPS Workshop on Machine Learning for Health, 2019. URL: https://arxiv.org/abs/1911.06411.

[179] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness., in: International Conference on Learning Representations (ICLR), 2019. URL: https://openreview.net/forum?id=Bygh9j09KX.

[180] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, pp. 23–30. doi:10.1109/iros.2017.8202133.

[181] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, S. Birchfield, Training deep networks with synthetic data: Bridging the reality gap by domain randomization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 1082–1088. doi:10.1109/cvprw.2018.00143.

[182] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, W. Brendel, Benchmarking robustness in object detection: Autonomous driving when winter is coming (2019). arXiv:1907.07484.

[183] L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423. doi:10.1109/cvpr.2016.265.

[184] H. Li, H. Han, Z. Li, L. Wang, Z. Wu, J. Lu, S. K. Zhou, High-resolution chest X-ray bone suppression using un-

paired CT structural priors, IEEE Transactions on Medical Imaging 39 (2020) 3053–3063. doi:10.1109/tmi.2020.2986242.

[185] D. Rolnick, A. Veit, S. Belongie, N. Shavit, Deep learning is robust to massive label noise (2017). arXiv:1705.10694.

[186] C. G. Northcutt, A. Athalye, J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, in: ICLR 2021 RobustML and Weakly Supervised Learning Workshops; NeurIPS 2020 Workshop on Dataset Curation and Security, 2021. URL: https://arxiv.org/abs/2103.14749.

[187] C. O. Schmidt, S. Struckmann, C. Enzenbach, A. Reineke, J. Stausberg, S. Damerow, M. Huebner, B. Schmidt, W. Sauerbrei, A. Richter, Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R, BMC Medical Research Methodology 21 (2021) 63. doi:10.1186/s12874-021-01252-7.

[188] ISO/IEC JTC 1/SC 42 Artificial intelligence, ISO/IEC AWI 52591 series – Data quality for analytics and ML, under development. URL: https://www.iso.org/standard/81088.html.

[189] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, A. Van Esbroeck, Monotonic calibrated interpolated look-up tables, The Journal of Machine Learning Research 17 (2016) 3790–3836. URL: https://jmlr.org/papers/volume17/15-243/15-243.pdf.

[190] A. Gupta, N. Shukla, L. Marla, A. Kolbeinsson, K. Yellepeddi, How to incorporate monotonicity in deep networks while preserving flexibility?, in: NeurIPS 2019 Workshop on Machine Learning with Guarantees, 2019. URL: https://arxiv.org/abs/1909.10662.

[191] T. Cohen, M. Welling, Group equivariant convolutional networks, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, PMLR, 2016, pp. 2990–2999. URL: http://proceedings.mlr.press/v48/cohenc16.html.

[192] R. Evans, E. Grefenstette, Learning explanatory rules from noisy data, Journal of Artificial Intelligence Research 61 (2018) 1–64. doi:10.1613/jair.5714.

[193] Z. Chen, Y. Bei, C. Rudin, Concept whitening for interpretable image recognition, Nature Machine Intelligence 2 (2020) 772–782. doi:10.1038/s42256-020-00265-z.

[194] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, C. Rudin, IAIA-BL: A case-based interpretable deep learning model for classification of mass lesions in digital mammography (2021). arXiv:2103.12308.

[195] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms, MIT Press, Cambridge, MA, US, 2017.

[196] J. Pearl, The seven tools of causal inference, with reflections on machine learning, Communications of the ACM 62 (2019) 54–60. doi:10.1145/3241036.

[197] B. Schölkopf, Causality for machine learning (2019). arXiv:1911.10500.

[198] B. Scholkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proceedings of the IEEE 109 (2021) 612–634. doi:10.1109/jproc.2021.3058954.

[199] J. G. Richens, C. M. Lee, S. Johri, Improving the accuracy of medical diagnosis with causal machine learning, Nature Communications 11 (2020) 3069. doi:10.1038/s41467-020-17419-7.

[200] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019). doi:10.1002/widm.1312.

[201] J. Kukačka, V. Golkov, D. Cremers, Regularization for deep learning: A taxonomy (2017). arXiv:1710.10686.

[202] A. Lavin, Machine learning is no place to "move fast and break things", Forbes (2020). URL: https://www.forbes.com/sites/alexanderlavin/2020/02/17/machine-learning-is-no-place-to-move-fast-and-break-things/, last accessed: July 10, 2021.

[203] R. Ehlers, Formal verification of piece-wise linear feed-forward neural networks, in: D. D'Souza, K. Narayan Kumar (Eds.), Automated Technology for Verification and Analysis. ATVA 2017. Lecture Notes in Computer Science, volume 10482, Springer, Cham, 2017. doi:10.1007/978-3-319-68167-2_19.

[204] G. Katz, C. W. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer, Reluplex: An efficient SMT solver for verifying deep neural networks, in: R. Majumdar, V. Kuncak (Eds.), Proceedings of the 29th International Conference on Computer Aided Verification, volume 10426 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 97–117. doi:10.1007/978-3-319-63387-9_5.

[205] D. Guidotti, F. Leofante, A. Tacchella, C. Castellini, Improving reliability of myocontrol using formal verification, IEEE Transactions on Neural Systems and Rehabilitation Engineering 27 (2019) 564–571. doi:10.1109/tnsre.2019.2893152.

[206] K. Pei, Y. Cao, J. Yang, S. Jana, Towards practical verification of machine learning: The case of computer vision systems (2017). arXiv:1712.01785.

[207] J. Törnblom, S. Nadjm-Tehrani, Formal verification of random forests in safety-critical applications, in: Formal Techniques for Safety-Critical Systems, Springer International Publishing, 2019, pp. 55–71. doi:10.1007/978-3-030-12988-0_4.

[208] J. Törnblom, S. Nadjm-Tehrani, Formal verification of input-output mappings of tree ensembles, Science of Computer Programming 194 (2020) 102450. doi:10.1016/j.scico.2020.102450.

[209] N. Sato, H. Kuruma, Y. Nakagawa, H. Ogawa, Formal verification of decision-tree ensemble model and detection of its violating-input-value ranges, IEICE Transaction D, Feb, 2020 (2019). doi:10.1587/transinf.2019EDP7120. arXiv:1904.11753.

[210] X. Huang, M. Kwiatkowska, S. Wang, M. Wu, Safety verification of deep neural networks, in: R. Majumdar, V. Kuncak (Eds.), Proceedings of the 29th International Conference on Computer Aided Verification, volume 10426 of *Lecture Notes in Computer Science*, Springer,

2017, pp. 3–29. doi:10.1007/978-3-319-63387-9_1.

[211] M. Lécuyer, V. Atlidakis, R. Geambasu, D. Hsu, S. Jana, Certified robustness to adversarial examples with differential privacy, in: Proceedings of the 40th IEEE Symposium on Security and Privacy, 2019, pp. 656–672. doi:10.1109/SP.2019.00044.

[212] J. M. Cohen, E. Rosenfeld, J. Z. Kolter, Certified adversarial robustness via randomized smoothing, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 1310–1320. URL: http://proceedings.mlr.press/v97/cohen19c.html.

[213] S. A. Seshia, S. Jha, T. Dreossi, Semantic adversarial deep learning, IEEE Design & Test 37 (2020) 8–18. doi:10.1109/MDAT.2020.2968274.

[214] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, P. Eckersley, Explainable machine learning in deployment, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery (ACM), New York, NY, USA, 2020, pp. 648–657. doi:10.1145/3351095.3375624.

[215] K. Aslansefat, I. Sorokos, D. Whiting, R. Tavakoli Kolagari, Y. Papadopoulos, SafeML: Safety monitoring of machine learning classifiers through statistical difference measures, in: M. Zeller, K. Höfig (Eds.), Model-Based Safety and Assessment. IMBSA 2020. Lecture Notes in Computer Science, volume 12297, Springer, Cham, 2020. doi:10.1007/978-3-030-58920-2_13.

[216] H. Jiang, B. Kim, M. Guan, M. Gupta, To trust or not to trust a classifier, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/7798-to-trust-or-not-to-trust-a-classifier.

[217] S. Liu, R. Garrepalli, T. Dietterich, A. Fern, D. Hendrycks, Open category detection with PAC guarantees, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3169–3178. URL: http://proceedings.mlr.press/v80/liu18e.html.

[218] P. Schulam, S. Saria, Can you trust this prediction? Auditing pointwise reliability after learning, in: K. Chaudhuri, M. Sugiyama (Eds.), Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, volume 89 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 1022–1031. URL: http://proceedings.mlr.press/v89/schulam19a.html.

[219] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, B. Lakshminarayanan, Likelihood ratios for out-of-distribution detection, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/9611-likelihood-ratios-for-out-of-distribution-detection.

[220] K. R. Varshney, Engineering safety in machine learning, in: Proceedings of the Information Theory and Applications Workshop (ITA), IEEE, 2016, pp. 1–5. doi:10.1109/ita.2016.7888195.

[221] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. . hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, M. Anderljung, Toward trustworthy AI development: Mechanisms for supporting verifiable claims (2020). arXiv:2004.07213v2.

[222] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Transactions on Neural Networks and Learning Systems (2020) 1–21. doi:10.1109/tnnls.2020.3027314.

[223] M. Bukowski, R. Farkas, O. Beyan, L. Moll, H. Hahn, F. Kiessling, T. Schmitz-Rode, Implementation of eHealth and AI integrated diagnostics with multidisciplinary digitized data: are we ready from an international perspective?, European Radiology 30 (2020) 5510–5524. doi:10.1007/s00330-020-06874-x.

[224] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, K. Tsaneva-Atanasova, Artificial intelligence, bias and clinical safety, BMJ Quality & Safety 28 (2019) 231–237. doi:10.1136/bmjqs-2018-008370.

[225] K. Goddard, A. Roudsari, J. C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, Journal of the American Medical Informatics Association 19 (2012) 121–127. doi:10.1136/amiajnl-2011-000089.

[226] R. Parasuraman, D. H. Manzey, Complacency and bias in human use of automation: An attentional integration, Human Factors: The Journal of the Human Factors and Ergonomics Society 52 (2010) 381–410. doi:10.1177/0018720810376055.

[227] X-Force Threat Intelligence Index, IBM, 2020. URL: https://www.ibm.com/security/data-breach/threat-intelligence.

[228] Verizon, Data breach investigations report, 2020. URL: https://enterprise.verizon.com/resources/reports/dbir/.

[229] D. Truxius, E. Müller, N. Krupp, J. Suleder, O. Matula, D. Kniel, Cyber Security Review of Network-Connected Medical Devices, BSI Federal Office for Information Security, 2020. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/DigitaleGesellschaft/ManiMed_Abschlussbericht_EN.html.

[230] Postmarket Management of Cybersecurity in Medical Devices, U.S. Food and Druck administration (FDA), 2016. URL: https://www.fda.gov/media/95862/download.

[231] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: Proceedings of the 38th IEEE Symposium on Security and Privacy, 2017, pp. 3–18. doi:10.1109/SP.2017.41.

[232] D. Bernau, P.-W. Grassal, J. Robl, F. Kerschbaum, Assessing differentially private deep learning with membership inference (2019). arXiv:1912.11328.

[233] K. Packhäuser, S. Gündel, N. Münster, C. Syben, V. Christlein, A. Maier, Is medical chest X-ray data anonymous? (2021). arXiv:2103.08562.

[234] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[235] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, N. Papernot, High accuracy and high fidelity extraction of neural networks, in: 29th USENIX Security Symposium (USENIX Security 20), 2020, pp. 1345–1362. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/jagielski.

[236] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, J. S. Rellermeyer, A survey on distributed machine learning, ACM Computing Surveys 53 (2020) 1–33. doi:10.1145/3377454.

[237] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, A. Das, Differential privacy-enabled federated learning for sensitive health data, in: Machine Learning for Health (ML4H) at NeurIPS, 2019. URL: https://arxiv.org/abs/1910.02578.

[238] B. Hitaj, G. Ateniese, F. Perez-Cruz, Deep models under the GAN, in: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery (ACM), New York, NY, US, 2017, pp. 603–618. doi:10.1145/3133956.3134012.

[239] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 2938–2948. URL: http://proceedings.mlr.press/v108/bagdasaryan20a.html.

[240] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Y. Bengio, Y. LeCun (Eds.), Proceedings of the 3rd International Conference on Learning Representations, 2015. URL: http://arxiv.org/pdf/1412.6572.

[241] K. Papangelou, K. Sechidis, J. Weatherall, G. Brown, Toward an understanding of adversarial examples in clinical trials, in: M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, G. Ifrim (Eds.), Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part I, volume 11051 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 35–51. doi:10.1007/978-3-030-10925-7_3.

[242] N. Carlini, D. A. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, in: B. M. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, A. Sinha (Eds.), Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, US, November 3, 2017, Association for Computing Machinery (ACM), New York, NY, US, 2017, pp. 3–14. doi:10.1145/3128572.3140444.

[243] N. Chattopadhyay, A. Chattopadhyay, S. S. Gupta, M. Kasper, Curse of dimensionality in adversarial examples, in: Proceedings of the International Joint Conference on Neural Networks, 2019, pp. 1–8. doi:10.1109/IJCNN.2019.8851795.

[244] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery (ACM), New York, NY, US, 2016, p. 308–318. doi:10.1145/2976749.2978318.

[245] K. Chaudhuri, C. Monteleoni, A. D. Sarwate, Differentially private empirical risk minimization, Journal of Machine Learning Research 12 (2011) 1069–1109. URL: https://jmlr.org/papers/volume12/chaudhuri11a.html.

[246] H. B. McMahan, D. Ramage, K. Talwar, L. Zhang, Learning differentially private recurrent language models, in: Proceedings of the 6th International Conference on Learning Representations, 2018. URL: https://arxiv.org/abs/1710.06963v3.

[247] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, et al., Adversarial robustness toolbox v1.0.0 (2018). URL: https://adversarial-robustness-toolbox.org.

[248] Z. Sun, P. Kairouz, A. T. Suresh, H. B. McMahan, Can you really backdoor federated learning? (2019). arXiv:1911.07963.

[249] M. Du, R. Jia, D. Song, Robust anomaly detection and backdoor attack detection via differential privacy, in: International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=SJx0q1rtvS.

[250] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, P. S. Yu, Privacy and robustness in federated learning: Attacks and defenses (2020). arXiv:2012.06337.

[251] F. Tramer, D. Boneh, Slalom: Fast, verifiable and private execution of neural networks in trusted hardware, in: Proceedings of the 7th International Conference on Learning Representations, 2019. URL: https://arxiv.org/abs/1806.03287.

[252] Intel Software Guard Extensions (Intel SGX), Intel, 2021. URL: https://software.intel.com/en-us/isa-extensions/intel-sgx.

[253] F. Armknecht, C. Boyd, C. Carr, K. Gjøsteen, A. Jäschke,

C. A. Reuter, M. Strand, A guide to fully homomorphic encryption, IACR Cryptology ePrint Archive 2015 (2015) 1192. URL: https://eprint.iacr.org/2015/1192.pdf.

[254] P. Mohassel, Y. Zhang, SecureML: A system for scalable privacy-preserving machine learning, in: Proceedings of the 38th IEEE Symposium on Security and Privacy, 2017, pp. 19–38. doi:10.1109/SP.2017.12.

[255] B. Parno, J. Howell, C. Gentry, M. Raykova, Pinocchio: Nearly practical verifiable computation, in: Proceedings of the 34th IEEE Symposium on Security and Privacy, 2013, pp. 238–252. doi:10.1109/SP.2013.47.

[256] N. Carlini, M. Jagielski, I. Mironov, Cryptanalytic extraction of neural network models, in: D. Micciancio, T. Ristenpart (Eds.), Advances in Cryptology – CRYPTO 2020, Springer International Publishing, 2020, pp. 189–218. doi:10.1007/978-3-030-56877-1_7.

[257] A. Athalye, N. Carlini, On the robustness of the CVPR 2018 white-box adversarial example defenses, CoRR abs/1804.03286 (2018). arXiv:1804.03286.

[258] A. Athalye, N. Carlini, D. A. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, in: J. G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 274–283. URL: http://proceedings.mlr.press/v80/athalye18a.html.

[259] F. Tramer, N. Carlini, W. Brendel, A. Madry, On adaptive attacks to adversarial example defenses, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1633–1645. URL: https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf.

[260] N. Carlini, D. A. Wagner, Towards evaluating the robustness of neural networks, in: Proceedings of the 38th IEEE Symposium on Security and Privacy, 2017, pp. 39–57. doi:10.1109/SP.2017.49.

[261] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: Proceedings of the 6th International Conference on Learning Representations, 2018. URL: https://arxiv.org/abs/1706.06083.

[262] X. Yin, S. Kolouri, G. K. Rohde, Adversarial example detection and classification with asymmetrical adversarial training (2019). arXiv:1905.11475.

[263] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, A. Kurakin, On evaluating adversarial robustness, CoRR abs/1902.06705 (2019). arXiv:1902.06705.

[264] J. Rauber, W. Brendel, M. Bethge, Foolbox: A python toolbox to benchmark the robustness of machine learning models, in: Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning, 2017. URL: https://arxiv.org/abs/1707.04131.

[265] J. Rauber, R. Zimmermann, M. Bethge, W. Brendel, Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX, Journal of Open Source Software 5 (2020) 2607. doi:10.21105/joss.02607.

[266] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, R. Long, P. McDaniel, Technical report on the CleverHans v2.1.0 adversarial examples library, 2018. URL: https://github.com/cleverhans-lab/cleverhans. arXiv:1610.00768.

[267] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning (2017). arXiv:1702.08608.

[268] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[269] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80–89. doi:10.1109/DSAA.2018.00018.

[270] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery (ACM), New York, NY, US, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[271] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[272] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, the Precise4Q consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Medical Informatics and Decision Making 20 (2020) 310. doi:10.1186/s12911-020-01332-6.

[273] N. Bostrom, E. Yudkowsky, The ethics of artificial intelligence, in: K. Frankish, W. M. Ramsey (Eds.), The Cambridge handbook of artificial intelligence, volume 1, Cambridge university press, Cambridge, UK, 2014, pp. 316–334.

[274] A. Weller, Transparency: Motivations and challenges, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, volume 11700, Springer, Cham, 2019, pp. 23–40. doi:10.1007/978-3-030-28954-6_2.

[275] M. Ribera, A. Lapedriza, Can we do better explanations? A proposal of user-centered explainable AI, in: Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, US, 2019.

[276] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: Informing design practices for explainable AI user experi-

ences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), New York, NY, US, 2020, pp. 1–15. doi:10.1145/3313831.3376590.

[277] P. B. de Laat, Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability?, Philosophy & Technology 31 (2017) 525–541. doi:10.1007/s13347-017-0293-z.

[278] G. S. Collins, J. B. Reitsma, D. G. Altman, K. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BMC Medicine 13 (2015) 1. doi:10.1186/s12916-014-0241-z.

[279] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford, Datasheets for datasets, in: Workshop on Fairness, Accountability, and-Transparency in Machine Learning, 2018. URL: https://www.fatml.org/media/documents/datasheets_for_datasets.pdf.

[280] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! Criticism for interpretability, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf.

[281] C. Molnar, Interpretable Machine Learning, 2020. URL: https://christophm.github.io/interpretable-ml-book/.

[282] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The Dataset Nutrition Label: A framework to drive higher data quality standards, in: D. Hallinan, R. Leenes, S. Gutwirth, P. De Hert (Eds.), Data Protection and Privacy, Hart, Oxford, UK, 2020, pp. 1–25. URL: https://datanutrition.org/.

[283] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, FactSheets: Increasing trust in AI services through supplier's declarations of conformity, IBM Journal of Research and Development 63 (2019) 6:1–6:13. doi:10.1147/JRD.2019.2942288.

[284] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, PMLR, 2016, pp. 1050–1059. URL: http://proceedings.mlr.press/v48/gal16.html.

[285] A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision?, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[286] B. Patro, M. Lunayach, S. Patel, V. Namboodiri, U-CAM: Visual explanation using uncertainty based class activation maps, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7443–7452. doi:10.1109/ICCV.2019.00754.

[287] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, Artificial Intelligence in Medicine 94 (2019) 42–53. doi:10.1016/j.artmed.2019.01.001.

[288] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330. doi:10.5555/3305381.3305518.

[289] Z. Shao, J. Yang, S. Ren, Calibrating deep neural network classifiers on out-of-distribution datasets (2020). arXiv:2006.08914.

[290] J. J. Thiagarajan, P. Sattigeri, D. Rajan, B. Venkatesh, Calibrating healthcare AI: Towards reliable and interpretable deep predictive models (2020). arXiv:2004.14480.

[291] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, S. Michalak, On mixup training: Improved calibration and predictive uncertainty for deep neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://papers.nips.cc/paper/9540-on-mixup-training-improved-calibration-and-predictive-uncertainty-for-deep-neural-networks.

[292] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X. X. Zhu, A survey of uncertainty in deep neural networks (2021). arXiv:2107.03342.

[293] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/access.2018.2870052.

[294] D. Gunning, D. Aha, DARPA's explainable artificial intelligence (XAI) program, AI Magazine 40 (2019) 44–58. doi:10.1609/aimag.v40i2.2850.

[295] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications and bibliography for explainable AI, DARPA XAI Program (2019).

[296] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[297] R. Moraffah, M. Karami, R. Guo, A. Raglin, H. Liu, Causal interpretability for machine learning – problems, methods and evaluation, SIGKDD Explorations Newsletter 22 (2020) 18–33. doi:10.1145/3400051.3400058.

[298] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model, The Annals of Applied Statistics 9 (2015). doi:10.1214/15-aoas848.

[299] T. Lombrozo, N. Vasilyeva, Causal explanation, in: M. R.

Waldmann (Ed.), Oxford handbook of causal reasoning, Oxford University Press Oxford, UK, 2017, pp. 415–432. doi:10.1093/oxfordhb/9780199399550.013.22.

[300] R. Marcinkevičs, J. E. Vogt, Interpretability and explainability: A machine learning zoo mini-tour (2020). arXiv:2012.01805.

[301] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11771.

[302] Y. Sha, M. D. Wang, Interpretable predictions of clinical outcomes with an attention-based recurrent neural network, in: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Association for Computing Machinery (ACM), New York, NY, US, 2017, pp. 230–240. doi:10.1145/3107411.3107445.

[303] T. Guo, T. Lin, N. Antulov-Fantulin, Exploring interpretable LSTM neural networks over multi-variable data, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2494–2504. URL: http://proceedings.mlr.press/v97/guo19b.html.

[304] M. Al-Shedivat, A. Dubey, E. Xing, Contextual explanation networks, Journal of Machine Learning Research 21 (2020) 1–44. URL: http://jmlr.org/papers/v21/18-856.html.

[305] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018, pp. 7786–7795. URL: https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

[306] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps (2013). arXiv:1312.6034.

[307] J. Yosinski, J. Clune, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, in: Deep Learning Workshop, 31st International Conference on Machine Learning, 2015. URL: https://arxiv.org/abs/1506.06579.

[308] P. Dabkowski, Y. Gal, Real time image saliency for black box classifiers, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf.

[309] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, in: A. E. Villa, P. Masulli, A. J. Pons Rivero (Eds.), Proceedings of the International Conference on Artificial Neural Networks, Springer, Cham, 2016, pp. 63–71. doi:10.1007/978-3-319-44781-0_8.

[310] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[311] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications 10 (2019). doi:10.1038/s41467-019-08987-4.

[312] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, N. Berthouze, Evaluating saliency map explanations for convolutional neural networks: a user study, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 275–285. doi:10.1145/3377325.3377519.

[313] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M. D. Li, J. Kalpathy-Cramer, Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging (2020). doi:10.1101/2020.07.28.20163899.

[314] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and Information Systems 41 (2013) 647–665. doi:10.1007/s10115-013-0679-x.

[315] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law and Technology 31 (2017) 841–887. doi:10.2139/ssrn.3063289.

[316] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplements 27 (1990) 247–266. doi:10.1017/S1358246100005130.

[317] R. Byrne, Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 6276–6282. doi:https://doi.org/10.24963/ijcai.2019/876.

[318] M. Harradon, J. Druce, B. Ruttenberg, Causal learning and explanation of deep neural networks via autoencoded activations (2018). arXiv:1802.00541.

[319] R. D. Cook, Detection of influential observation in linear regression, Technometrics 19 (1977) 15–18. doi:10.1080/00401706.1977.10489493.

[320] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1885–1894. URL: http://proceedings.mlr.press/v70/koh17a/koh17a.pdf.

[321] B. Liu, M. Udell, Impact of accuracy on model interpretations (2020). arXiv:2011.09903.

[322] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, A. Weller, Rethink-

ing attention with performers, in: International Conference on Learning Representations (ICLR), 2021. URL: https://openreview.net/pdf?id=Ua6zuk0WRH.

[323] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques (2019). arXiv:1909.03012.

[324] H. Nori, S. Jenkins, P. Koch, R. Caruana, InterpretML: A unified framework for machine learning interpretability (2019). URL: https://interpret.ml. arXiv:1909.09223.

[325] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, J. Wilson, The What-If Tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics (2019) 1–1. doi:10.1109/tvcg.2019.2934619.

[326] J. N. Yan, Z. Gu, H. Lin, J. M. Rzeszotarski, Silva: Interactively assessing machine learning fairness using causality, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), New York, NY, US, 2020, pp. 1–13. doi:10.1145/3313831.3376447.

[327] A. Aler Tubella, A. Theodorou, J. C. Nieves, Interrogating the black box: Transparency through information-seeking dialogues, in: Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '21, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2021, p. 106–114.

[328] J. A. Kroll, The fallacy of inscrutability, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2018) 20180084. doi:10.1098/rsta.2018.0084.

[329] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Visual Informatics 1 (2017) 48–56. doi:10.1016/j.visinf.2017.01.006.

[330] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects (2018). arXiv:1812.04608.

[331] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning (2019). arXiv:1908.09635.

[332] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

[333] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudik, H. Wallach, Improving fairness in machine learning systems, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), New York, NY, US, 2019, pp. 1–16. doi:10.1145/3290605.3300830.

[334] R. Gillon, Medical ethics: four principles plus attention to scope, BMJ 309 (1994) 184. doi:10.1136/bmj.309.6948.184.

[335] B. Green, L. Hu, The myth in the methodology: Towards a recontextualization of fairness in machine learning, in: Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 2018.

[336] R. Gillon, Raising the profile of fairness and justice in medical practice and policy, Journal of Medical Ethics 46 (2020) 789–790. doi:10.1136/medethics-2020-107039.

[337] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness - FairWare '18, Association for Computing Machinery (ACM), New York, NY, US, 2018, pp. 1–7. doi:10.1145/3194770.3194776.

[338] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, M. H. Chin, Ensuring fairness in machine learning to advance health equity, Annals of Internal Medicine 169 (2018) 866. doi:10.7326/m18-1990.

[339] B. Ustun, Y. Liu, D. Parkes, Fairness without harm: Decoupled classifiers with preference guarantees, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6373–6382. URL: http://proceedings.mlr.press/v97/ustun19a.html.

[340] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaney-Israni, A. Goldenberg, Do no harm: a roadmap for responsible machine learning for health care, Nature Medicine 25 (2019) 1337–1340. doi:10.1038/s41591-019-0548-6.

[341] M. J. Kusner, J. R. Loftus, The long road to fairer algorithms, Nature 578 (2020) 34–36. doi:10.1038/d41586-020-00274-3.

[342] H. Suresh, J. V. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle (2021). arXiv:1901.10002.

[343] K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites, Proceedings of the National Academy of Sciences 113 (2016) 4296–4301. doi:10.1073/pnas.1516047113.

[344] K. Hamberg, Gender bias in medicine 4 (2008) 237–243. doi:10.2217/17455057.4.3.237.

[345] S. M. Phelan, D. J. Burgess, M. W. Yeazel, W. L. Hellerstedt, J. M. Griffin, M. Ryn, Impact of weight bias and stigma on quality of care and outcomes for patients with obesity, Obesity Reviews 16 (2015) 319–326. doi:10.1111/obr.12266.

[346] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, D. Sculley, No classification without representation: Assessing geodiversity issues in open data sets for the developing world, in: NeurIPS Workshop on Machine Learning for the Developing World, 2017. URL: https://arxiv.org/abs/1711.08536.

[347] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Men also like shopping: Reducing gender bias amplifi-

cation using corpus-level constraints, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 2979–2989. doi:10.18653/v1/d17-1323.

[348] S. Hooker, Moving beyond "algorithmic bias is a data problem", Patterns 2 (2021) 100241. doi:10.1016/j.patter.2021.100241.

[349] R. Berk, L. Brown, A. Buja, E. George, L. Zhao, Working with misspecified regression models, Journal of Quantitative Criminology 34 (2017) 633–655. doi:10.1007/s10940-017-9348-7.

[350] A. Krishnan, A. Almadan, A. Rattani, Understanding fairness of gender classification algorithms across gender-race groups, in: IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 1028–1035. doi:10.1109/ICMLA51294.2020.00167.

[351] E. Dehon, N. Weiss, J. Jones, W. Faulconer, E. Hinton, S. Sterling, A systematic review of the impact of physician implicit racial bias on clinical decision making, Academic Emergency Medicine 24 (2017) 895–904. doi:10.1111/acem.13214.

[352] I. W. Maina, T. D. Belton, S. Ginzberg, A. Singh, T. J. Johnson, A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test, Social Science & Medicine 199 (2018) 219–229. doi:10.1016/j.socscimed.2017.05.009.

[353] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed impact of fair machine learning, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3150–3158. URL: http://proceedings.mlr.press/v80/liu18c.html.

[354] M. DeCamp, C. Lindvall, Latent bias and the implementation of artificial intelligence in medicine, Journal of the American Medical Informatics Association 27 (2020) 2020–2023. doi:10.1093/jamia/ocaa094.

[355] S. Escalera, M. A. Bagheri, M. Valstar, M. T. Torres, B. Martinez, X. Baro, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, ChaLearn looking at people and faces of the world: Face analysis workshop and challenge 2016, in: Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 706–713. doi:10.1109/cvprw.2016.93.

[356] H. J. Ryu, H. Adam, M. Mitchell, InclusiveFaceNet: Improving face attribute detection with race and gender diversity, in: Fairness, Accountability, and Transparency in Machine Learning Workshop, 34th International Conference on Machine Learning, 2018. URL: https://arxiv.org/abs/1712.00193.

[357] B. Cowgill, F. Dell'Acqua, S. Deng, D. Hsu, N. Verma, A. Chaintreau, Biased programmers? Or biased data? A field experiment in operationalizing AI ethics, in: Proceedings of the 21st ACM Conference on Economics and Computation, 2020, pp. 679–681. doi:10.1145/3391403.3399545.

[358] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, volume 28 of *Proceedings of Machine Learning Research*, PMLR, 2013, pp. 325–333. URL: http://proceedings.mlr.press/v28/zemel13.html.

[359] D. Pedreshi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, 2008, pp. 560–568. doi:10.1145/1401890.1401959.

[360] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, C. Wang, Learning fair representations via an adversarial framework (2019). arXiv:1904.13341.

[361] D. Madras, E. Creager, T. Pitassi, R. Zemel, Learning adversarially fair and transferable representations, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3384–3393. URL: http://proceedings.mlr.press/v80/madras18a.html.

[362] D. McNamara, C. S. Ong, R. C. Williamson, Costs and benefits of fair representation learning, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 263–270. doi:10.1145/3306618.3317964.

[363] D. Madras, E. Creager, T. Pitassi, R. Zemel, Fairness through causal awareness, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (ACM), New York, NY, US, 2019, pp. 349–358. doi:10.1145/3287560.3287564.

[364] E. Creager, D. Madras, T. Pitassi, R. Zemel, Causal modeling for fairness in dynamical systems, in: H. Daumé III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 2185–2195. URL: https://arxiv.org/abs/1909.09141.

[365] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, M. Welling, Causal effect inference with deep latent-variable models, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf.

[366] S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning (2018). arXiv:1808.00023.

[367] C. Dwork, N. Immorlica, A. T. Kalai, M. Leiserson, Decoupled classifiers for group-fair and efficient machine learning, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 119–133. URL: http://proceedings.mlr.press/v81/dwork18a.html.

[368] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI Fairness 360: An extensible toolkit for detecting, un-

derstanding, and mitigating unwanted algorithmic bias, 2018. URL: https://arxiv.org/abs/1810.01943.

[369] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data 5 (2017) 153–163. doi:10.1089/big.2016.0047.

[370] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration, in: Advances in Neural Information Processing Systems, 2017, pp. 5680–5689. URL: https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html.

[371] K. Makhlouf, S. Zhioua, C. Palamidessi, Survey on causal-based machine learning fairness notions (2020). arXiv:2010.09553.

[372] K. D. Johnson, D. P. Foster, R. A. Stine, Impartial predictive modeling: Ensuring group fairness in arbitrary models (2016). arXiv:1608.00528.

[373] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in neural information processing systems, 2016, pp. 3315–3323. URL: https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

[374] Y. Savani, C. White, N. S. Govindarajulu, Intra-processing methods for debiasing neural networks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2798–2810. URL: https://proceedings.neurips.cc/paper/2020/file/1d8d70dddf147d2d92a634817f01b239-Paper.pdf.

[375] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (ACM), New York, NY, US, 2019, pp. 59–68. doi:10.1145/3287560.3287598.

[376] A. Smith, Public attitudes toward computer algorithms, Pew Research Center, 2018. URL: https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/.

[377] E. Vayena, A. Blasimme, I. G. Cohen, Machine learning in medicine: Addressing ethical challenges, PLOS Medicine 15 (2018) e1002689. doi:10.1371/journal.pmed.1002689.

[378] T. Wiegand, R. Krishnamurthy, M. Kuglitsch, N. Lee, S. Pujari, M. Salathé, M. Wenzel, S. Xu, WHO and ITU establish benchmarking process for artificial intelligence in health, The Lancet 394 (2019) 9–11. doi:10.1016/S0140-6736(19)30762-7.

[379] M. L. Petersen, M. J. van der Laan, Causal models and learning from data, Epidemiology 25 (2014) 418–426. doi:10.1097/ede.0000000000000078.

[380] B. Mittelstadt, Principles alone cannot guarantee ethical AI, Nature Machine Intelligence 1 (2019) 501–507. doi:10.1038/s42256-019-0114-4.