Buy the Book >>

CHAPTER 11 Artificial Intelligence

By Koen Cobbaert, MSc

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Introduction

Although different people may understand artificial intelligence (AI) differently, it has been a reality in healthcare for decades. Healthcare has adopted AI technology in medical devices, workflows, and decision-making processes. Rather than replacing the human component of healthcare delivery, artificial intelligence has become a vital tool or a companion to improve patient outcomes.

Artificial intelligence refers to a wide variety of techniques.[1] While neural networks are in the spotlight today, this chapter covers all forms of AI, including classical AI (e.g., search and optimization techniques), expert systems,[2] Hidden Markov Models,[3] and older forms of computer vision), symbolic AI[4] (e.g., logical reasoning[5] and decision making), Abstract Syntax Tree[6] (AST) modifying code, probabilistic reasoning,[7] machine learning,[8] knowledge representation,[9] planning and navigation, natural language processing,[10] and perception.[11] Hybrid approaches also exist, which use a combination of techniques, e.g., neural networks and symbolic AI. The connectionist AI takes care of the messiness and correlations of the real world, for example, to detect patient position and anatomy, and help convert those into symbols that a symbolic AI can use to interact with the patient during physiotherapy. The influence of the latter will likely increase in the future.[12]

See **Figure 11-1** for a vastly simplified diagram of the different types of AI.[13]

Artificial General Intelligence (AGI) can learn incrementally, reason abstractly, and act effectively over a wide range of domains. As the author does not anticipate AGI to appear on the market in the near- or medium-term future, this chapter focuses on so-called 'narrow AI,' i.e., artificial intelligence with a specific, rather than a general, purpose.

Current medical device legislation applies to machine learning devices. It appears fit to assure safety and reliable performance, including AI that continues to change after it is placed on the market.[14] What is lacking is guidance on how to comply with regulations practically. Despite a flurry of AI standardization activity, practical guidance for medical devices is scarce. This chapter provides an introduction to AI, its characteristics, and how these impact regulatory compliance. The chapter concludes with a horizon scan of legislative initiatives that may affect AI in medical devices.

Definition

Having one common internationally accepted definition for AI would be helpful when comparing AI investments and regulations across the world. Several definitions of AI exist,[15] each with different flavors. At the time of writing, the International Medical Device Regulators Forum (IMDRF) is drafting a definition for machine learning medical devices.[16]

In trying to capture AI's essence, definitions tend to focus on AI's learning ability or its intelligent nature, two aspects that pose significant interpretation issues in a legal context.[17] Regarding AI's intelligent nature, no scientific consensus on the meaning of 'intelligence' exists.[18] As society tends to strip older AI techniques of their 'intelligent' status, the term poses a large moving edge. In terms of learning ability, learning is generally understood as 'acquiring knowledge or new skills.' If learning knowledge is sufficient, many digital products qualify as AI as soon as they acquire input data. If learning a new skill is necessary, the meaning of the term narrows significantly.

Today, it is uncertain what is and what is not covered by the term AI. Therefore, this chapter focuses on characteristics commonly associated with AI and their regulatory implications.

Machine Learning

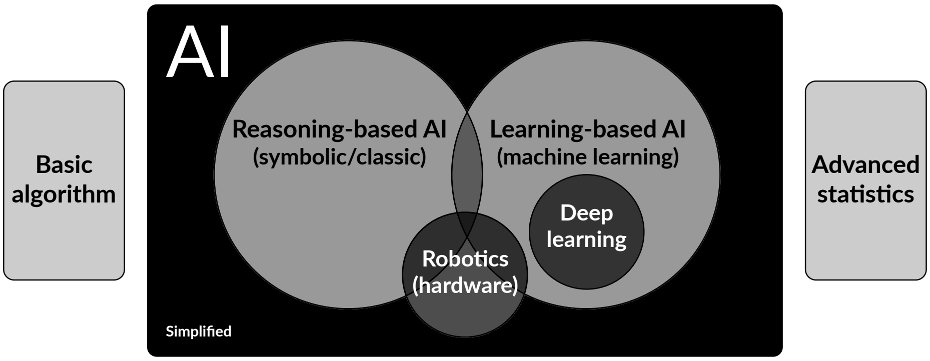
A characteristic of machine learning devices is that they can change based on training data (samples to fit a machine learning model), without being programmed explicitly. In contrast, other AI technologies learn without training data, such as through genetic programming[19] or reasoning. For example, semantic computing learns through semantic networks (a knowledge base that represents semantic relations between concepts in a network). Through reason, deduction, and inference, the AI may evolve or adapt during use. There are different perspectives to the aspect of change.

Global Versus Local Change

During global change, the manufacturer or health institution trains a machine learning model that is part of a device, i.e., 'the global model.' After the validation and conformity assessment, if applicable, the device is deployed

Figure 11-1. A Simplified Diagram of Different AI Techniques

There are roughly two different techniques: reason-based AI, which reasons based on symbols or rules programmed in the system. Instinctively, the reader can sense that the human has a bit more control of the rules. A second technique is learning-based AI. These are systems that use large amounts of data to construct a mathematical model to recognize correlations. This auto-construction can be done in a supervised or unsupervised manner. Supervised means that humans determine the categories of data themselves, e.g., these are cats; these are dogs, now learn to recognize them. Unsupervised, and deep learning, a sub-category of machine learning that belongs to this, is not given any labels. The developer throws data into the system, and by itself, it recognizes patterns and categories in the data. The big advantage is that the system recognizes things that a human might not have thought about. On the other hand, it might also generate categories that are problematic from a moral point of view. When people talk about black box systems, they predominantly talk about deep learning. So, it is incorrect that all AI is not transparent; it is mainly this category where it is difficult to determine how a decision was made. This is not the case for all AI systems. Sometimes, basic algorithms and statistics are also called AI. Robotics, also known as 'embodied AI,' is shown to be half in the box, half out of the box. Robots can work using AI but also can operate using simpler algorithms. The drawing overlaps all these types of AI because there are also hybrid forms of AI.



into clinical use through one or more copies. Each copy has a model identical to that of the original device. The manufacturer or health institution may continue to train the global model. Through periodic or real-time updates, the manufacturer or health institution synchronizes the local and global models. Updates can comprise changes to the settings or the design of the device. The local models are identical to a version of the global model.

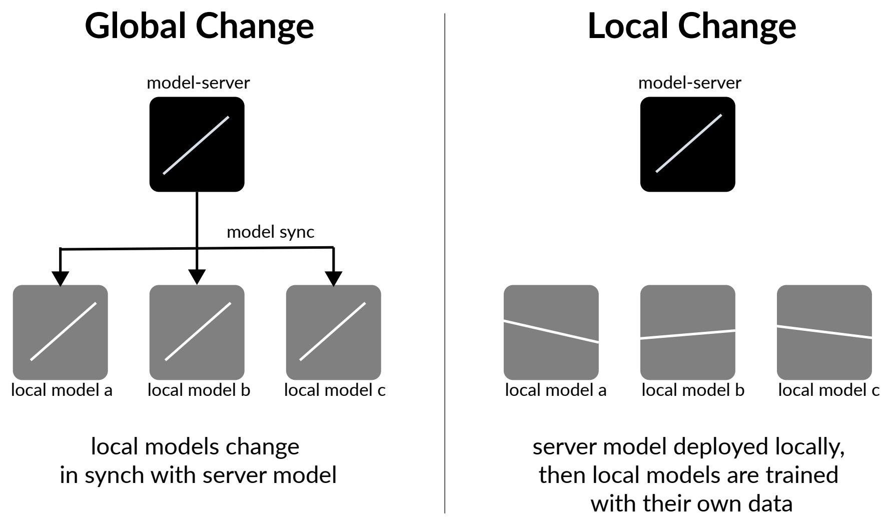
The synchronization can occur in real-time or with a delay, depending on the need for validation and conformity assessment. Conversely, during local change, the local models learn and change independently of the global model (see **Figure 11-2**).

Federated Learning

Global change is irrespective of whether the global model learns based on local data or on data collected separately, e.g., through the use of data repositories (e.g., registries or decentralized personal data stores) or data generated through clinical investigations and postmarket clinical follow-up (PMCF).

Federated learning (see **Figure 11-3**) is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples without exchanging them with the developer, e.g., in health institutions. This approach contrasts with traditional centralized machine learning techniques, where all the local datasets are uploaded to one server, and more classic decentralized methods are used that often assume that local data samples are identically distributed. The main advantage of using federated approaches is to ensure data privacy or data secrecy. Indeed, no local data is uploaded externally, concatenated, or exchanged. Since the entire database is segmented into local bits, it is more difficult to hack into it. With federated learning, only machine learning parameters are exchanged. Also, such parameters can be encrypted before sharing between learning rounds to extend privacy. Homomorphic encryption schemes can be used to directly make computations on the encrypted data without decrypting them beforehand.

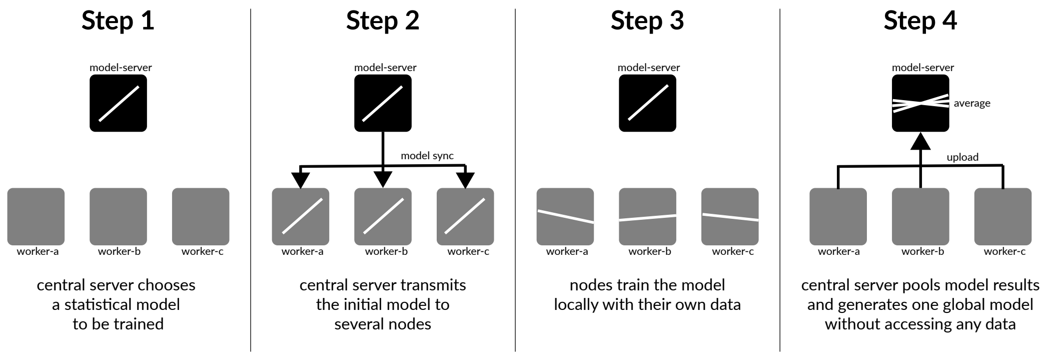
Figure 11-2. Global and Local Model Change



©Koen Cobbaert 2021

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 11-3. Visualization of Federated Learning


Step 1
model-server
worker-a
worker-b
worker-c
central server chooses a statistical model to be trained
Step 2
model-server
model sync
worker-a
worker-b
worker-c
central server transmits the initial model to several nodes
Step 3
model-server
worker-a
worker-b
worker-c
nodes train the model locally with their own data
Step 4
model-server
average
upload
worker-a
worker-b
worker-c
central server pools model results and generates one global model without accessing any data

Pre-Determined Change

Local change occurs in machine learning devices that adapt, refine their performance, or calibrate themselves to a specific patient or healthcare setting's characteristics. An AI that learns and changes itself during clinical use requires manufacturers to determine the limits for such change to occur safely while assuring performance for the intended purpose. Manufacturers should establish those boundaries of change before placing the product on the market. Set boundaries determine the framework in which regulatory approval allows for changes.

Pre-determined change is not unique to machine learning devices. Many medical devices adapt to the patient or their environment or compensate for wear and tear by reconfiguring or self-healing their design. For example, a CT machine undergoes regular calibration cycles to adjust for wear and tear. The calibration reconfigures the CT software to compensate for, and adapt to, hardware changes.

Pre-determined changes through machine learning personalized healthcare include a modification of the AI's 'working point' based on the local/patient environment; it allows AI to maximize performance for a given patient or situation and enables the device to adapt to the patient rather than force them to adapt to the device. For example, AI used in the joints of a bionic foot allows the kinetics to be adapted to the patient, rather than letting the patient adapt their gait to the prosthesis's kinetics.

Change Dynamics

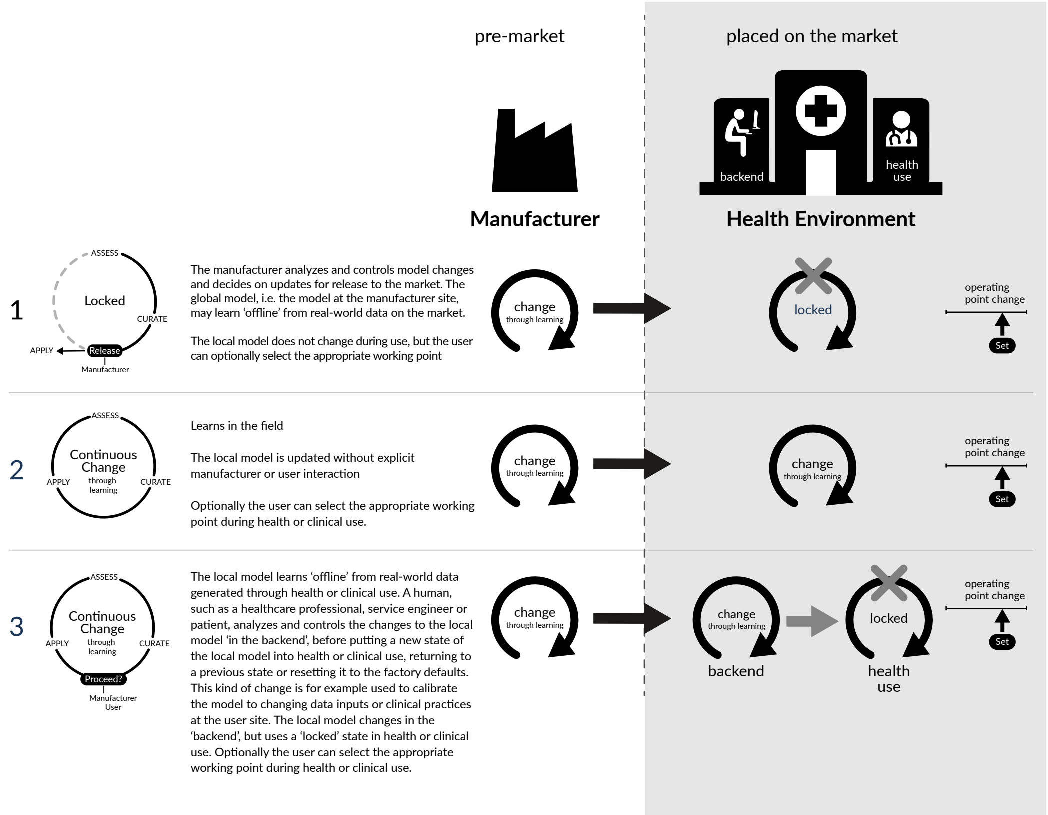
Depending on the device, the manufacturer, the health institution, the caregiver, the patient, or a combination of these, can control the change. The actor responsible for the change and whether the change occurs before or after the device was placed on the market brings forth different regulatory implications (see **Figure 11-4**).

In the first scenario, the AI is locked, and the manufacturer controls the learning. 'Locked AI' does not change its design during runtime. Examples of locked AI are static lookup tables, decision trees, and complex classifiers. Locked AI generally provides the same result each time the same input is applied. However, there are some exemptions to this, e.g., if the AI contains non-deterministic[20] algorithms, or the user can change the working point on the operating curve. The performance of a 'locked' AI can therefore still change. Consider an algorithm to screen for tuberculosis at airports. While the algorithm's design is locked, the user may still choose a point on the operating curve that is different from the factory default, e.g., to trade-off increased sensitivity with reduced specificity.

AI can change during runtime in various other ways. In the second scenario (continuous change), the model continues to learn within the manufacturer's boundaries. Consider, for example, an AI for precision medicine. The AI calculates the optimal medication dose to reduce the tremor of a patient with Parkinson's disease while limiting the drug's side effects. As the disease progresses over months or years, symptoms change. The algorithm continues to evolve together with the patient's disease state. In this scenario, the user may still be able to change the working point on the operating curve. A patient may, for example, on a particular day, decide to pick a point on the operating curve that lowers the dose, allowing for more tremor, but improving their cognitive performance. A patient may prefer a different operating point because the medication causes mild cognitive impairments, such as distraction, disorganization, difficulty planning, and accomplishing tasks.

In the third scenario (discrete changes), the learning initially occurs under the manufacturer's control. The model is then handed over to the user (or another party) for further calibration or adjustment to the local context or to a specific patient. The change occurs within the intended

Figure 11-4. Three Scenarios for Change of Machine Learning Devices in Relation to Their Placement on the Market



The manufacturer analyzes and controls model changes and decides on updates for release to the market. The global model, i.e. the model at the manufacturer site, may learn 'offline' from real-world data on the market.

The local model does not change during use, but the user can optionally select the appropriate working point

Learns in the field

The local model is updated without explicit manufacturer or user interaction

Optionally the user can select the appropriate working point during health or clinical use.

The local model learns 'offline' from real-world data generated through health or clinical use. A human, such as a healthcare professional, service engineer or patient, analyzes and controls the changes to the local model 'in the backend', before putting a new state of the local model into health or clinical use, returning to a previous state or resetting it to the factory defaults. This kind of change is for example used to calibrate the model to changing data inputs or clinical practices at the user site. The local model changes in the 'backend', but uses a 'locked' state in health or clinical use. Optionally the user can select the appropriate working point during health or clinical use.

©Koen Cobbaert 2021

use, within the change boundaries, and according to the manufacturer's algorithm change protocol (ACP) (see next section for a description of change boundaries and ACP). The manufacturer remains responsible for the device. Consider, for example, an AI for the prediction of sepsis. The hospital makes a change in its local practices, now also encoding blood parameters procalcitonin and IL-6 and making these available for the AI. The manufacturer has proved these blood parameters work as valid inputs, described in the ACP. The user can now further improve the local model's performance by training the AI on these extra blood parameters, following the manufacturer's ACP.

Also, hybrid scenarios may exist, whereby the model continues to evolve, but the user has the ability to revert back to an earlier state of the model or to the factory default.

Note: Having a human in the loop during learning to control what model state is put into clinical use is different from having a human in the loop to control the device during clinical use. A continuously changing AI does not have a human in the loop to control the learning but can nevertheless have a human in the loop to control device functions during clinical use, for example, through a device override or emergency stop.

Postmarket Significant Change

Medical device legislation requires a manufacturer to determine whether a new device release changes significantly ('substantial' in the words of the IMDRF). If a new software release changes significantly, the manufacturer must perform a new conformity assessment before placing the device on the market. Under the *EU MDR* and *EU IVDR*, a health institution developing a machine learning device for in-house use also must perform such significant change determination and perform a new conformity assessment before putting the device into clinical use. In some cases, the health institution may use a manufacturer's machine learning component to build a new device, or the health institution may change an existing device in a significant way. A significant change that occurs postmarket provides a fourth scenario (see **Figure 11-5**).

In the fourth scenario, the user intentionally changes the local model in a way not allowed by the manufacturer's change control plan, either by making a change:

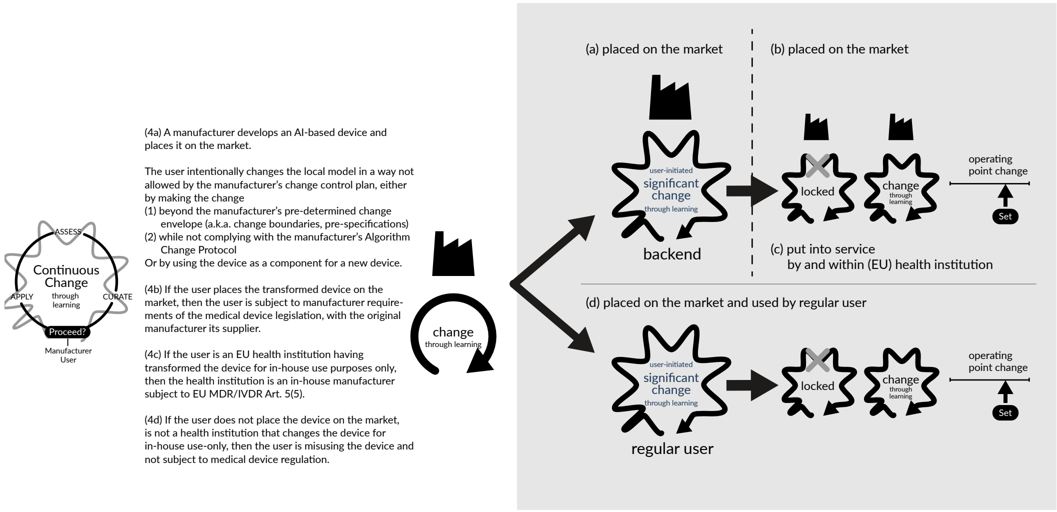
- Beyond the manufacturer's pre-determined change envelope (also known as change boundaries or pre-specifications), e.g., for a purpose not covered by the intended use or for more specific purposes than claimed by the manufacturer
- While not complying with the manufacturer's ACP

Alternatively, the user incorporates the device as a component in a new device.

In the fourth scenario, the device is intentionally changed by the user in a significant way. The user is misusing the device. Consider, for example, an AI to read quantitative Polymerase Chain Reaction (qPCR) assays. The AI can handle qPCR curves generated for assays for plant, animal, or human specimens. Assume a second manufacturer or a health institution produces an assay for the detection of SARS-COV-2. By continuing the AI training, it becomes especially good at reading qPCR curves associated with a SARS-COV-2 assay. 'Reading qPCR curves for SARS-COV-2' is a more specific claim and requires a higher level of performance before it is considered state-of-the-art than 'reading qPCR curves for assays for plant, animal or human specimen.'

Consequently, the more specific claim is considered a significant change. A manufacturer or EU-based health institution (following *EU MDR* Art. 5(5) on in-house manufacturing) performing such change carries the manufacturer responsibilities under medical device regulations. If the second manufacturer can prove safety and performance without having the technical documentation of the original device, the original manufacturer's intellectual property remains protected if the notified body or competent author-

Figure 11-5. Visualization of Postmarket Significant Change



(4a) A manufacturer develops an AI-based device and places it on the market.

The user intentionally changes the local model in a way not allowed by the manufacturer's change control plan, either by making the change
(1) beyond the manufacturer's pre-determined change envelope (a.k.a. change boundaries, pre-specifications)
(2) while not complying with the manufacturer's Algorithm Change Protocol
Or by using the device as a component for a new device.

(4b) If the user places the transformed device on the market, then the user is subject to manufacturer requirements of the medical device legislation, with the original manufacturer its supplier.

(4c) If the user is an EU health institution having transformed the device for in-house use purposes only, then the health institution is an in-house manufacturer subject to EU MDR/IVDR Art. 5(5).

(4d) If the user does not place the device on the market, is not a health institution that changes the device for in-house use-only, then the user is misusing the device and not subject to medical device regulation.

©Koen Cobbaert 2021

ity accepts the 'new' device's technical documentation containing just a reference to the master file of the original device. That master file then has to be made available by the original manufacturer to the notified body. **Note:** FDA provides the possibility of using master files. Depending on the technology used, the second manufacturer or health institution operating under this scenario may need to implement technical measures, a quality agreement, and monitoring of original device updates to ensure the new device is safe and performant in light of state-of-the-art.

Change Boundaries and Algorithm Change Protocol

The manufacturer must specify a change envelope or change boundaries and an ACP. As long as the device operates within the related change boundaries, its safety and performance are guaranteed (e.g., minimum specificity and sensitivity). The manufacturer can ensure this through procedural (e.g., acceptance testing) or technical measures. How change envelopes can be defined depends on the technology used. The manufacturer should demonstrate why the chosen parameters and threshold values are valid in consideration of the clinical state of the art and the intended purpose of the device. Currently, no standard is available that guides the drafting of an ACP.[21]

Manufacturers can increase the trust necessary for implementation by informing regulators and users about the following:

- How the AI learns over time
- What the allowed change envelope/boundaries are of the AI
- What caused a significant change in the AI behavior
- How the performance and safety is being assured as it adapts
- How quality control of new training data is assured
- What triggers algorithm change
- What performance and confidence levels[22] are during a given timeframe
- Whether and how a user can reject an algorithm change or roll-back to a previous state

For troubleshooting purposes, a manufacturer also may want to monitor actual performance and change/drift of an evolving algorithm to detect performance deficits, e.g., by implementing traceability through an audit trail.[23]

Change According to Medical Device Legislation

Most medical device regulations require a conformity assessment before a manufacturer can place a device on the market. The conformity assessment must demonstrate the regulatory requirements are met.[24] Devices can change in terms of design and characteristics after conformity assessment, but only if the manufacturer has a quality management system to address these changes in a timely manner with regard to regulatory compliance, including compliance with conformity assessment procedures.[25] Manufacturers are prohibited from suggesting uses for the device other than those stated in the intended use for which the conformity assessment was carried out.[26] So, what does this mean for devices that change during use?

In the EU, notified bodies must have documented procedures and contractual arrangements in place with the manufacturer relating to the assessment of changes to the approved design of a device and the intended use or claims made for the device. Manufacturers must submit plans for such changes for prior approval. Such changes may affect conformity with the general safety and performance requirements or with the conditions prescribed for the use of the product, including changes related to limitations of the intended purpose or conditions of use. The notified body must assess the proposed changes and verify whether, after these changes, the design of a device or type of a device still meets the requirements of the regulation, must notify the manufacturer of its decision, and must provide a report or, as applicable, a supplementary report to the EU technical documentation assessment certificate containing the justified conclusions of its assessment.[27]

This implies that a manufacturer can place a device on the market that can change within a pre-defined change envelope or tolerances for which a conformity assessment was carried out, provided the manufacturer respects the contractual agreements with the notified body. This approach appears possible under most medical device legislation.

In contrast, no medical device legislation currently exists that allows manufacturers to place machine learning devices on the market that are intended to change outside of the change envelope or to suggest claims, intended uses, or use conditions to the device for which no conformity assessment was carried out (see **Figure 11-6**). For example, the functionality of a machine learning device placed on the market intended to detect frontotemporal dementia evolves during runtime to detect dementia with Lewy Bodies or Creutzfeldt Jakob disease; this is considered a new intended purpose and requires a new conformity assessment.

Changes outside of the change envelope or assigning new claims or use conditions require an update of the technical documentation, including the clinical evaluation and a new conformity assessment to be carried out.[28] Significant changes to the pre-determined ACP require an update of the technical documentation, including the clinical evaluation and a new conformity assessment.

Assume a natural or legal person wants to place an existing device on the market in the EU (1) by changing its intended purpose or (2) by modifying it in a way that compliance with the applicable requirements may be affected. In that case, that person shall assume the obligations incumbent on manufacturers, except if they change the device, without changing its intended purpose, to adapt it for an individual patient. Then manufacturer obligations do not ensue, but a Member State may still require the person to register as a manufacturer of custom-made devices. This implies that a person can adapt a machine learning bionic eye to restore a patient's eyesight if the eye is intended for that purpose, even if this involves a modification in such a way that compliance may be affected.

In the EU, a health institution can (1) change the intended purpose or (2) modify a device in such a way that compliance with the applicable requirements are affected so that it can be used on multiple patients, but only if the health institution meets the conditions for in-house manufacturing, meaning:

- The device is not transferred to another legal entity.
- It has an appropriate quality management system in place.
- It justifies the target patient group's specific needs cannot be met or cannot be met at the appropriate level of performance by an equivalent device available on the market.
- It provides information on the manufacturing, modification, and use to its competent authority upon request and draws up a declaration that it shall make publicly available.[29]

The exception to this rule is when a health institution adapts a device for a purpose not in the scope of the medical device definition, e.g., a machine learning bionic limb is adapted for superhuman purposes rather than to alleviate or compensate for a disability. The health institution then must not meet the conditions for in-house manufacturing. The author is not aware of any restrictions outside the EU that apply to health institutions performing manufacturing.

Controllability and Human Oversight

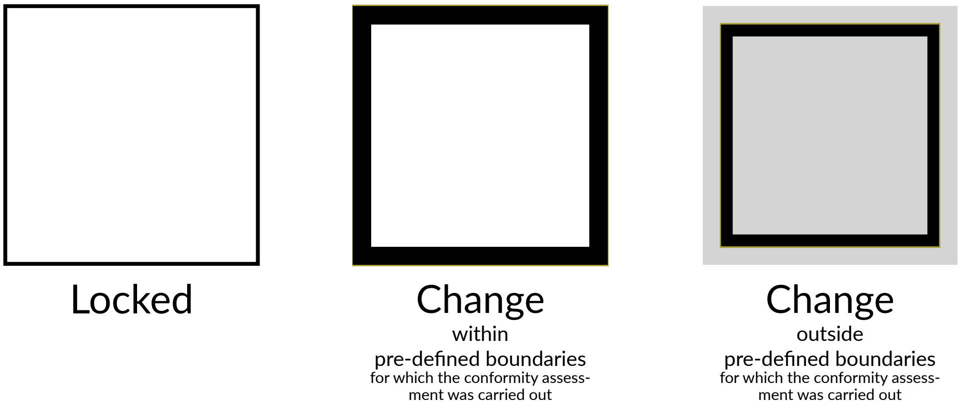
Controllability refers to the ability of an external operator to intervene or deactivate a machine learning device in a timely manner.[30] **Figure 11-7** illustrates the different types of controllability.[31] Having a human in the loop leverages the user's situational awareness and judgement, which may be beneficial for the performance and safety of machine learning devices.

Having a human in the loop requires situational awareness, enough time to intervene, and a mechanism to interfere (a communication link or physical controls) and take control or deactivate the device as required.

From a risk management and performance perspective, it may sometimes be necessary to take the human out of the loop to reduce the risk as far as possible to avoid human-machine

Figure 11-6. Illustration of AI Change Types

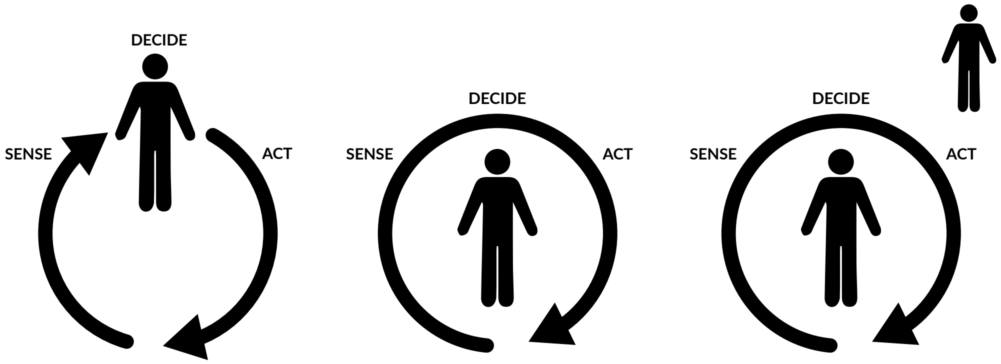
Most medical device regulations allow hardware devices that recalibrate or reconfigure themselves within certain boundaries for which a conformity assessment was carried out to be placed on the market. Similarly, manufacturers can place devices on the market that comprise AI that changes within pre-defined boundaries for which a conformity assessment was carried out. Currently, no medical device legislation is known that allows manufacturers to place devices on the market comprising AI that changes outside of pre-defined boundaries.


Locked
Change
within
pre-defined boundaries
for which the conformity assessment was carried out
Change
outside
pre-defined boundaries
for which the conformity assessment was carried out

©Koen Cobbaert 2021

Figure 11-7. Types of Controllability

Devices that are controlled or supervised by humans are also known as heteronomous devices. Devices that do not provide a means to intervene in a timely fashion are also known as autonomous devices. Hybrid devices provide direct control or supervisory control on certain functions and no control on other functions.



Direct Control

The AI performs a task and then waits for the human user to take an action before continuing

Supervisory Control

The AI can sense, decide, and act on its own. The human user supervises its operation and can intervene when required.

No Control

The AI can sense, decide, and act on its own. The human user cannot intervene in a timely fashion.

©Koen Cobbaert 2021

interaction problems,[32] such as a lack of operator situational awareness (sensing uncertainties/limited situational awareness), i.e., the operator having insufficient knowledge of the state of the device at the time of intervention.

When specific actions require the device to perform at high speed or with safety and accuracy, it may be safer to circumvent the limitations of immediate, real-time human control. For example, a machine learning robot for eye surgery[33] may require the human to be taken out of the loop because of the user's limited decision making-capacity,[34] limited situational awareness, and sensing uncertainties. In this case, the most effective human supervision will be based on continuous oversight and periodic retrospective review of the performance for individual patients or for cohorts of patients, for example, through PMCF.

Automation is, of course, not a panacea. Manufacturers and health institutions must be aware that automation leads to deskilling in some circumstances or may require more user training and higher-skilled individuals than actions performed without automation.[35]

Overtrust

Manufacturers of machine learning devices sometimes require a human in the loop to control when the AI is less confident of its decision. Human behavior, however, comes with its risks. For example, a Level 2 self-driving car[36] requires a human in the loop to drive the car safely. It might take 100.000 km before the car has an accident; the human may by then have put too much confidence (overtrust) and automation bias[37] in the car and no longer pays enough attention to take back control of the car in a timely and safe manner.

In a perfect world, we would like users who trust AI when it works at 100% accuracy but to be hypersensitive and identify when it is not.

In reality, people often tend to sway. If the first impression is positive, humans tend not to see when the AI makes mistakes or may forget or forgive the AI. How we must design our devices allowing users to calibrate their trust appropriately and how we must educate them in their first interactions with the device is an open field of research.

The radiology domain can inspire. For example, a radiology study showed that the human-device combination's accuracy might improve when a computer-aided-detection algorithm identifies more than one choice to the radiologist.[38] Offering multiple options can maintain trust in the system and mitigate the risks of overtrust by putting the human expertise at work.

In other cases, the AI's performance alone might be better than the performance of the human-AI team. Sometimes, it is necessary to take the human out of the loop altogether to get the best performance and reduce or eliminate use error. For example, in a clinical diagnostic application used to read quantitative Polymerase Chain Reaction qPCR) assays, FDA requires manufacturers to deactivate the possibility of the molecular biologist intervening, because the precision and reliability of the AI outperforms that of the human-AI team.[39] Taking the human out of the loop takes away human variability and mitigates the risk of a lab using fewer control samples than required by the assay manufacturer. On the other hand, when AI is trained on rare diseases with fewer datasets, it may require humans to be in the loop to reach maximum performance. Striking the right balance is important and differs on a case-by-case basis. Most medical device legislation requires manufacturers to eliminate or reduce the risk of use error.[40] Inappropriate levels of trust in automation may cause suboptimal performance.[41]

Transparency and Explicability

Only by gaining the user's trust will machine learning devices find their way into the care pathways. One way to gain confidence is to ensure transparency, both in terms of the organization that creates the AI and the AI itself. Transparency is also useful to clarify the liability, i.e., did the doctor or the user make a mistake? Was it the AI, incorrect or unforeseen input data, or a malicious actor, e.g., hackers or disgruntled employees?

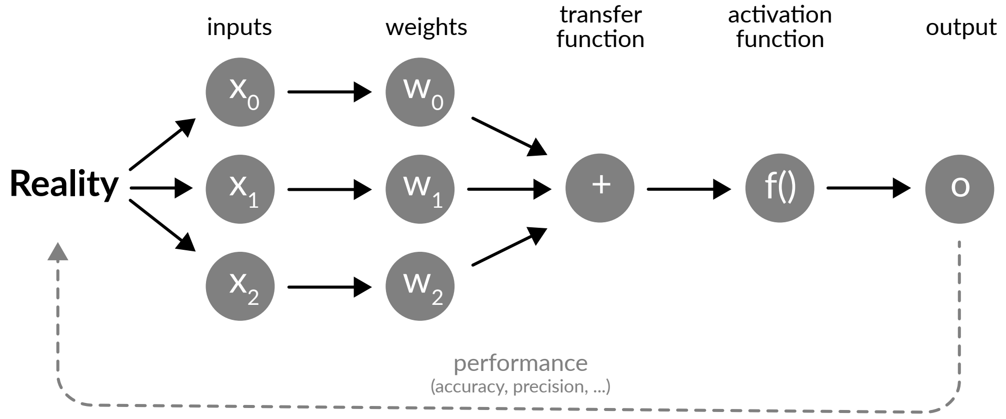
Transparency may also be needed (1) to allow manufacturers to determine operating parameters (when the device works or does not work), limitations to the intended use, contra-indications, inclusion, and exclusion criteria for input data or (2) to enable debugging of the system and detect potential issues of bias.

Transparent AI presents core decision-making elements in an open, comprehensive, accessible, clear, unambiguous, and interpretable way.[42] The first question that comes to mind when considering algorithmic interpretability is: 'Interpretable to whom?' The very word 'interpretable' implies an observer or subjective recipient who will judge whether they can understand the algorithm's model or its behaviors. Another challenge is the question of what we want to be interpretable, i.e., the historical data used to train the model, the model, the performance of the model found by the algorithm on a population cohort, or the model's decisions for a particular case?

Early well-known machine learning models are rather simple (see **Figure 11-8**) and principled (maximizing a natural and clearly stated objective, such as accuracy), and thus are interpretable or understandable to some extent. However, the 'rules' used by such models can be complicated to understand fully. They may capture complex and opaque relationships between the variables in what seemed to be a simple dataset. While radiologists may look for "a curved tube resembling a ram's horn, located in the inner region of the brain's temporal lobe," to identify the hippocampus, AI may use features and patterns that are not articulable in human language. While this makes AI an extremely powerful tool for clinical decision-making, it

Figure 11-8. Transparency and Interpretability of AI

Neural networks comprise relatively simple models, with weights, transfer, and activation function. Still, due to the vast amount of data, a person will not be able to process this information to the point of understanding it. AI being technically interpretable or transparent does not automatically imply that a doctor or a patient can interpret it, i.e., to understand cause and effect. A different level of abstraction is required to make the neural network interpretable to users.



©Koen Cobbaert 2021

also brings the risk of the AI reflecting spurious correlations in the data or overfitting to this particular dataset at the cost of transferability. Thus, simple algorithms applied to simple datasets can nevertheless lead to inscrutable models. However, it is also entirely possible that even a model we cannot understand 'in the large' or that is hidden from the user can make specific decisions that we can understand or rationalize post hoc.[43] For example, a doctor can review the output of an algorithm that reports on the wound healing stage (haemostasis, inflammatory, proliferative, or maturation) by looking at a wound picture to determine whether the algorithm identified the healing phase correctly. Alternatively, we can interrogate the model by having it tell us what it would do on any input. We can explore counterfactuals such as "What would be the smallest change in input data that would change the decision?" This type of explanatory understanding at the level of individual decisions or predictions is the basis for some of the more promising research on interpretability.[44]

Explained variance, i.e., "Given a blank sheet, what would be the minimum input data needed to receive this decision?" is on the opposing end of counterfactuals, i.e., "Given the complete picture, what is the minimum change needed to also change the answer?" Explained variance involves the AI providing "the minimum set of input data needed to come close to its decision." The minimum set of information depends on the desired level of closeness, which may differ for novice versus expert users. For example, an AI predicting the probability of survival from COVID-19 infection may explain to the user that 'age' contributed to 80% of its prediction. For some users, this may be sufficient information. In contrast, other users may want the AI to explain 99% of its prediction, adding that patient history contributed to 15%, specific lab results 3%, symptoms 0.5%, etc.[45]

'Inexplicable' devices are not unusual, as healthcare has long been known for accepting 'inexplicable' devices for certain purposes, as long as the technical file includes a description of the technology and adequate evidence of safety and performance. For example, manufacturers demonstrated via randomized controlled studies that electro-convulsive therapy is highly effective for severe depression, even though the mechanism of action remains unknown. The same holds true for many drugs under the medicines regulations, such as Selective Serotonin Reuptake Inhibitors (SSRI) or anesthetic agents.

Suppose a technology is significantly more effective than traditional methods in terms of diagnostic or therapeutic capabilities, but it cannot be explained. In that case, it poses ethical issues to hold back technology, simply on the basis that we cannot explain or understand it. Explicability is a means (to trust), not a goal. A blanket requirement that machine learning systems in medicine be explicable or interpretable is therefore unfounded and potentially harmful.[46] Of course, the advantage of making the model interpretable is that it helps the user gain confidence in the AI system faster, allowing the company to be successful commercially.

Enclosed AI, i.e., AI with no actionable outcome, may not require transparency toward the healthcare provider or patient but requires sufficient explicability[47] to the manufacturer or service engineer to allow verification, error detection, and troubleshooting. An example would be an AI that controls the cooling of a motor coil.

Manufacturers must also be transparent on the use of automated decision making. The rules in the EU General Data Protection Regulation (EU) 2016/679 (GDPR)[48] imply that when it is not immediately obvious that the user is interacting with an automated decision making process rather than a human (e.g., because there is no meaningful human involvement, for example, to improve a sensor or optics within a device), a software device must inform users of this, in particular patients, and include meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

Ethics

AI ethics is used in the meaning of respecting human values, including safety, transparency, accountability, and morality, but also in the meaning of confidentiality, data protection, and fairness. These aspects are not new, as philosophers have been debating ethics for millennia.

On the other hand, the science behind creating ethical algorithms is relatively new. Computer scientists have the responsibility to think about the ethical aspects of technologies they are involved in and mitigate or resolve any issues. Of these ethical aspects, fairness is probably the most complicated because fairness can mean different things in different contexts to different people.[49]

To consider the ethical aspects of software that changes itself through learning, the European Parliamentary Research Service[50] frames it as a real-world experiment. In doing so, it shifts the question of evaluating the moral acceptability of AI in general to the question of 'under what conditions is it acceptable to experiment with AI in society?' It uses the analogy to healthcare, where medical experimentation requires manufacturers to be explicit about the experimental nature of healthcare technologies by following the rigorous procedures of a clinical investigation, subject to ethics committee approval, patient consent, and careful monitoring to protect the subjects involved or impacted.

Many AI ethics frameworks have appeared in recent years.[51] These differ based on the organization's goals and operating contexts. Such frameworks have limits, for example, because many AI systems comprise a tradeoff between algorithmic fairness and accuracy.[52,53] A fair algorithm should provide the same benefits for the group while protecting it from discrimination. An accurate algorithm, on the other hand, should make a prediction that is as precise as possible for a certain subgroup, e.g., according to age, gender, smoking history, previous illnesses, etc. When an algorithm lies on the tradeoff curve between fairness and accuracy, it is often a matter of public policy, rather than an isolated decision made by the organization. For example, a hypothetical example of an AI algorithm used during the COVID-19 pandemic determines what patients should receive treatment. Certain countries or regions would prefer to allocate their scarce resources to patients that are predicted to have the highest chance of survival (accuracy prevails), whereas others may prefer to apply a fair allocation of resources rather than considering a patient's age and gender (fairness prevails). Such national and regional differences require a handshake between the company's ethics framework and that at the policy level. To accommodate for national and regional differences, algorithms can be designed so that they can be adjusted by the ethics committees at hospitals or at the regional level to meet the level of accuracy versus fairness desired.

Ethics committees emerged in healthcare in the 1970s at the very heart of hospitals. Initially, they focused on research ethics for clinical investigations on human subjects. Such committees exist throughout Europe and are regulated by law. Today, many of these committees have organized themselves at the regional or national level and also focus on clinical ethics. They have evolved into democratic platforms of public debate on medicine and human values. There is no single model. Every country has created its own organizational landscape, according to its moral and ideological preferences, adapted to the political structure of its health system, and its method of financing healthcare.[54] This complex reality and the lack of a unified approach make it challenging for companies to engage with ethics committees and find a single working point on the tradeoff curve between accuracy and fairness that is acceptable worldwide.

Bias

From a scientific point of view, bias is the tendency of a statistic to overestimate or underestimate a parameter.[55] From a legal perspective, bias is any prejudiced or partial personal or social perception of a person or group.[56] It is beyond this chapter's scope to discuss the differences between the scientific and legal definition; suffice it to say that we attribute a broader meaning to the legal definition, mainly because the scientific definition generally is understood to refer to systematic estimation errors.[57] In contrast, the legal definition also can apply to one-off errors in perception.

Aiming for software to be unbiased is desirable. Zero bias is, however, impossible to achieve. Bias may enter the AI development chain at different stages (see **Figure 11-9**). We humans all have our blind spots. Therefore, any data set that relies on humans making decisions will have some form of bias. Every hospital is different, every country is different, and every patient is different. You can never achieve zero bias when extrapolating. In the medical world, clinical investigations of devices for adults have historically underrepresented women, minority racial or ethnic groups, and to some extent, patients over age 65.[58] AI can maintain or even amplify such bias through its decisions. In trying to optimize its function, AI might ignore a minority if the minority looks different from the general population to optimize for the general population.

For manufacturers to establish that bias is minimized, they need to assess the AI for bias and ensure bias has been minimized if considered harmful. For example, they can compare the software output against an independent reference standard (e.g., a ground truth, biopsy, or expert consensus). They can perform specific sanity tests in terms of accuracy on every group to efficiently identify from the data.[59] The challenge is for computer scientists to get an accurate picture of which group(s) could be potentially biased. The difference might not show up in aggregate, but only when focusing on a sub-population within that group, where no test can exhaustively cover the space of all permutations. A big challenge to address bias is that sufficient and complete data must be available, which is rarely possible under

GDPR, causing the tradeoff between fairness and accuracy. Also, testing AI for non-discrimination on an ethical basis is at odds with GDPR and poses risks to users' privacy. Under current GDPR requirements, developers should not be able to access attributes such as ethnicity and, therefore, could not test for ethnic representation in a dataset.

Conversely, software can play an essential role in identifying and minimizing bias. This is not specific to artificial intelligence but to software in general. Historically, it was hard to prove unintended discriminatory bias based on race, for example. Software can enable feedback loops that make it easier to detect and fix bias issues.

Standardization bodies are currently developing standards to characterize data sets. Manufacturers can use these standards to establish data quality for training or evaluation purposes (see **Chapter 5** for a discussion of clinical evaluation of software and machine learning devices). They can use these characteristics to develop bias and determine whether the AI is suitable for a specific target population. However, mandatory certification of training data against these standards is not an effective mechanism to assure the AI is safe and effective for the target population or that bias is minimized. Manufacturers do not always have access to the training data (see machine learning section). There are forms of AI that learn without training data. Also, bias can enter the AI development chain at different points. Training data is only one of those entry points. Instead, manufacturers can make a more comprehensive assessment of bias through the use of qualitative evaluation data.

EU AI Legislation

Legislators across the world are focusing on ethical aspects of AI and the presence of bias. A 2019 heat map published by Anna Jobin[60] shows that the number of published AI ethics guidelines has increased, especially in Europe and the US. As ethical guidelines are not enforceable, the EU is assessing if and how to regulate ethical aspects of AI (see **Figure 11-10**).

People often attribute a broad meaning to ethics. The High-Level Expert Group on AI (HLEG) 'Ethics Guidelines for Trustworthy

Figure 11-9. Bias Entry Points in the AI Development Chain

Bias can creep in at different stages of AI development: 1. data collection, 2. data processing operations such as cleaning, enrichment, and aggregation or through assumptions in the design of data processing pipelines, 3. human judgment about the model construction and evaluation metrics can introduce biases in the processes of algorithm development and testing, and 4. through context changes or user interaction.

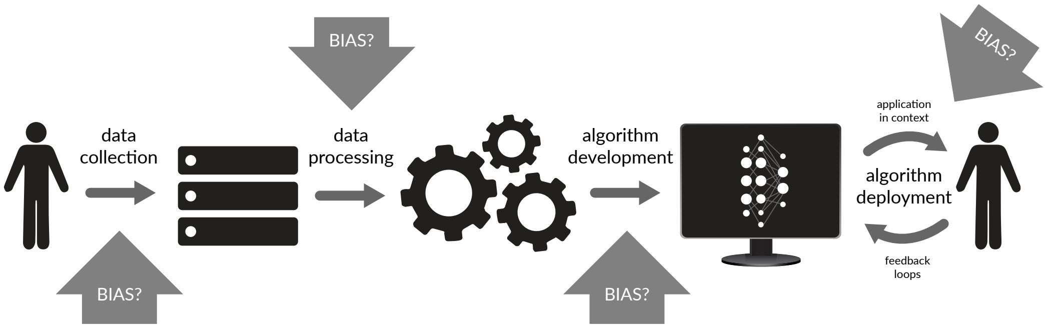
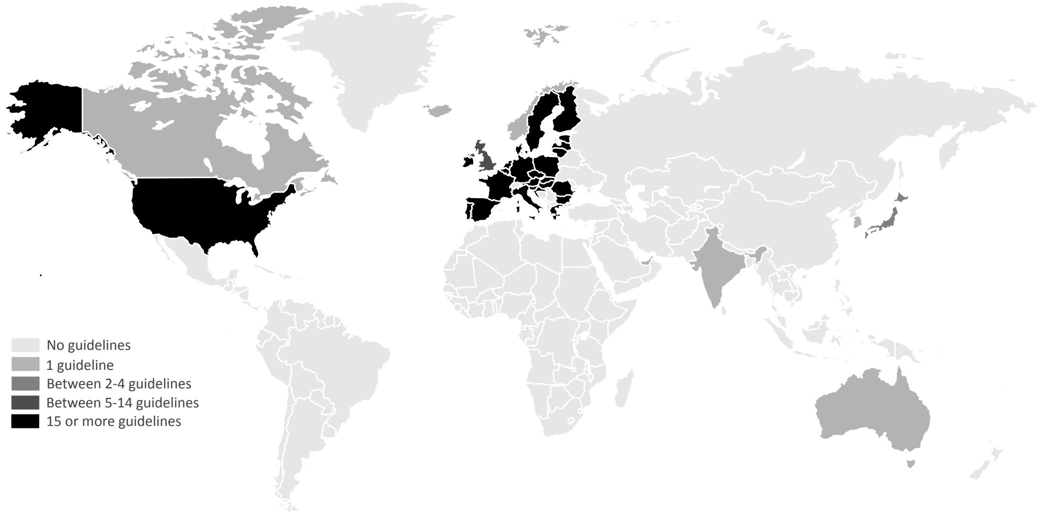


Figure 11-10. Geographic Distribution of Issuers of Ethical AI Guidelines by Number of Documents Released



Source: Jobin A. Artificial Intelligence: The global landscape of ethics guidelines. June 2019. https://www.researchgate.net/publication/334082218_Artificial_Intelligence_the_global_landscape_of_ethics_guidelines. Accessed 16 February 2021.

AI,’[61] for example, includes both algorithmic as well as data aspects: non-discrimination, diversity, bias, privacy and data governance, societal and environmental wellbeing (e.g., energy needed to train AI and its impact on global warming), human agency and oversight, transparency, accountability (related to logging), technical robustness, and safety. Most of these aspects are already covered by existing regulations, such as the medical device, data protection, product liability, and CO2 emission regulations, although they are not as specific to AI as digital rights advocates and the European Parliament would like. The European Commission intends to publish a legislative proposal on 21 April 2021 to address the ethical aspects of AI. It will likely cover several of these ethical dimensions and apply on a horizontal level, i.e., potentially include machine learning medical devices. The EU is also assessing whether and how to address the liability and intellectual property of AI.

Chinese AI Legislation

The Organization Center for Medical Device Evaluation (CMDE), a division of the Chinese regulatory authority, the National Medical Products Administration (NMPA), issued comprehensive requirements[62] encompassing scrutiny of machine learning devices across their entire device lifecycle. The National Institute for Food and Drug control (NIFDC), another division of NMPA, supplements these requirements with a growing body of standards, e.g., to characterize the data sets used for the training or evaluation of AI (see **Figure 11-11**). Manufacturers are well-advised to take these recommended standards into account early in the development project when drafting the clinical evaluation and clinical development plan so that there are no surprises when the regulatory submission for China is prepared.

Through standardization, the Artificial Intelligence Medical Device Innovation and Cooperation Platform (AIMDICP), a subdivision of CMDE, actively encourages the creation of evaluation databases and test platforms, starting with high prevalent diseases, such as lung cancer and diabetic retinopathy. As the world is a long way from having databases covering all 55,000 diseases and conditions listed in the 11th International Classification of Diseases (ICD 11) published by the World Health Organisation (WHO), this Chinese initiative is a welcome start.

US AI Legislation

In the US, FDA published a discussion paper that focuses on machine learning devices that change during runtime, citing the Precertification (Pre-Cert) Program[63] as a possible regulatory pathway for AI. The Pre-Cert Program is intended to be a regulatory model that is more streamlined and efficient, resulting in getting products to market and to patients faster than existing 510(k), *de novo*, and pre-market pathways.

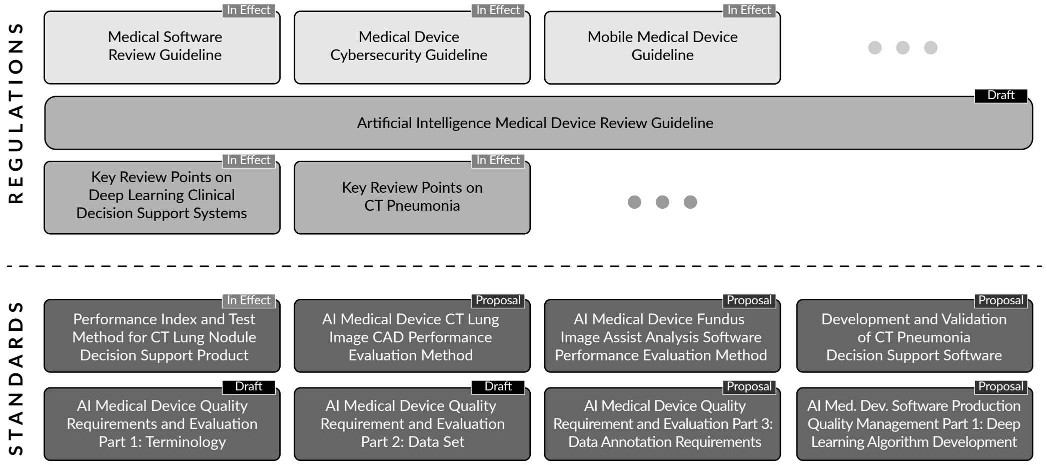
The Pre-Cert Program involves focusing on the product developer instead of focusing primarily on the product itself. If the developer can demonstrate a culture of quality, excellence, and responsiveness, FDA believes that a streamlined approval process could be allowed. The shift from a pure product focus to a product and process viewpoint is a new pathway for FDA and is a step towards convergence with the quality management system approach used within the EU. At the time of writing, FDA is piloting the Pre-Cert Program as a regulatory sandbox.

The Role of AI Standards

Generally, legislation provides high-level requirements to which a product must comply, while a standard provides requirements on how a product must comply. Consequently, standards are more prescriptive than legislation.

As standards are generally voluntary, they are the ideal vehicle to experiment with regulatory sandboxes. As countries develop AI legislation, they generally try to avoid being too prescriptive to avoid killing innovation. On the other hand, they do not want to place the bar too low, which could lead to unsafe uses of AI, and also could create inadequate user trust, resulting in less uptake of AI-products and causing harm to the competitive position of a country. Finding the sweet spot between not being too prescriptive

Figure 11-11. Overview of NMPA Regulation and Standardization Applicable to AI-Enabled Medical Devices


REGULATIONS
In Effect
Medical Software Review Guideline
In Effect
Medical Device Cybersecurity Guideline
In Effect
Mobile Medical Device Guideline
Draft
Artificial Intelligence Medical Device Review Guideline
In Effect
Key Review Points on Deep Learning Clinical Decision Support Systems
In Effect
Key Review Points on CT Pneumonia
STANDARDS
In Effect
Performance Index and Test Method for CT Lung Nodule Decision Support Product
Proposal
AI Medical Device CT Lung Image CAD Performance Evaluation Method
Proposal
AI Medical Device Fundus Image Assist Analysis Software Performance Evaluation Method
Proposal
Development and Validation of CT Pneumonia Decision Support Software
Draft
AI Medical Device Quality Requirements and Evaluation Part 1: Terminology
Draft
AI Medical Device Quality Requirement and Evaluation Part 2: Data Set
Proposal
AI Medical Device Quality Requirement and Evaluation Part 3: Data Annotation Requirements
Proposal
AI Med. Dev. Software Production Quality Management Part 1: Deep Learning Algorithm Development

and not placing the bar too low is key to a country's success. Keeping the legislative requirements high-level while experimenting with standard endorsement allows governments to establish their sweet spot relatively safely.

IEC, ISO, ITU, IEEE, CEN-CENELEC, BSI, and AAMI are examples of standardization bodies working in the AI space. As AI cuts across all sectors, these organizations have feverishly tried to claim their territory, which sometimes results in overlapping and competing standards. A certain level of competition among standardization bodies is desirable from a societal perspective to avoid blind spots from becoming fixated into the system. The downside is that companies active in different countries that use competing standards face higher costs, i.e., competing standards translate to a higher societal cost. Ideally, organizations that draft these competing standards can meet halfway and come together to form one standard. Convergence requires standardization bodies to work together.

Standards are created and voted on democratically. Any expert can join a standardization body and propose an AI standard idea or participate in its drafting. People participating in AI standardization are not necessarily AI experts. Participants such as consultants and certification service providers may see standards as a vehicle to grow their businesses, including in the domain of AI certification. Such hidden agendas bring the risk of more complex standards or standards that do not bring value to society.

As the world's regulatory powers, certification companies, and consultants push for the adoption of their home-grown AI standard to become the world standard, it appears that international standardization bodies have become the new battlegrounds.

Conclusion

More important than a definition for AI are its characteristics and how these affect compliance with existing legislation. Manufacturers should pay attention to its controllability, whether it changes through learning while in clinical use, and whether it is explainable or not. However, there is no room for a blanket requirement for AI's controllability and explainability because of trade-offs against safety and performance.

The European Commission is introducing initiatives to regulate the ethical aspects of AI. Ethics contains many dimensions, several of which are already covered by medical device legislation, including machine learning devices, which are heavily regulated.

Medical device legislation can benefit from guidance on how to apply it to machine learning devices. Legislators have started publishing such guidance, with China taking the lead. Legislators generally do not want to place the bar too low, which could affect the ability to create enough trust for AI to be adopted and hamper the country's competitive position. Nor do legislators want to be too prescriptive, as requirements could kill innovation and harm society in the long run. Finding the sweet spot is essential.

As demonstrated in China, standards play a crucial role in supporting legislation. Standards are a relatively safe tool to experiment with finding the best level of requirements. The push for AI standardization has generated feverish activity. Because of vested interests and how standards can advance a country's competitive position, standardization bodies appear to have become a new battleground for AI.

References

1. European Commission AI High-Level Expert Group. 8 April 2019. "A definition of AI – Main Capabilities and Disciplines." European Commission website. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines. Accessed 16 February 2021.
2. An expert system is a computer system emulating the decision-making ability of a human expert. Expert systems are designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if–then rules rather than through conventional procedural code. Wikipedia website. https://en.wikipedia.org/wiki/Expert_system. Accessed 16 February 2021.

3. Hidden Markov Model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobservable ('hidden') states. Wikipedia website. https://en.wikipedia.org/wiki/Hidden_Markov_model. Accessed 16 February 2021.

4. Symbolic AI: AI system that encodes knowledge using symbols and structures. CD2 ISO-IEC 22989. 3 September 2020. Symbolic AI performs manipulations of abstract objects/concepts for logic deduction. Symbolic AI focuses much less on (mathematical/numerical) calculation and much more on composition and manipulation.

5. Logical reasoning is a form of thinking in which premises and relations between premises are used in a rigorous manner to infer conclusions that are entailed (or implied) by the premises and the relations. Springer website. https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-1428-6_790#:~:text=Logical%20reasoning%20is%20a%20form,of%20science%20and%20artificial%20intelligence. Accessed 16 February 2021.

6. Abstract Syntax Tree is a tree representation of the abstract syntactic structure of source code written in a programming language. Wikipedia website. https://en.wikipedia.org/wiki/Abstract_syntax_tree. Accessed 16 February 2021.

7. Probabilistic reasoning combines probability theory with logic to handle uncertainty of knowledge. Javatpoint website. https://www.javatpoint.com/probabilistic-reasoning-in-artifical-intelligence. Accessed 16 February 2021.

8. Machine learning: process using computational techniques to enable systems to learn from data or experience. Source: IEC 23053, 3.16.

9. Knowledge representation is the field of artificial intelligence dedicated to representing information about the world in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language. Wikipedia website. https://en.wikipedia.org/wiki/Knowledge_representation_and_reasoning. Accessed 16 February 2021.

10. Natural language processing: <system> information processing based upon natural-language understanding and natural-language generation. CD2 ISO-IEC 22989, 3 September 2020.

11. Artificial Intelligence: A Modern Approach (3rd Edition), Stuart Russell and Peter Norvig, 2009, Pearson.

12. Today, in the world of AI there are two schools of thought: (1) that of Yann LeCun, who thinks we can reach Artificial General Intelligence via Deep Learning alone and (2) that of Gary Marcus, who thinks other forms of AI are needed, notably symbolic AI or hybrid forms. See Deep learning: A critical appraisal. arXiv 2018 G Marcus. arXiv preprint arXiv:1801.00631, 2019.

13. Deep learning: approach to creating rich hierarchical representations through the training of neural networks with many hidden layers. CD2 ISO-IEC 22989, 3 September 2020.

14. COCIR (2020), COCIR Analysis on AI in medical Device Legislation. September 2020. COCIR website. https://www.cocir.org/media-centre/publications/article/cocir-analysis-on-ai-in-medical-device-legislation-september-2020.html. Accessed 16 February 2021.

15. Samoili S, López Cobo M, Gómez E, De Prato G, Martínez-Plumed F, and Delipetrev B. AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163. This study assesses a large set of AI literature from three complementary perspectives: policy, research, and market. The paper contains a collection of 46 definitions. Most notable definitions:

    - The High-Level Expert Group (an independent expert group established by the European Commission in June 2018) in the context of the European AI strategy (See AI HLEG "A definition of AI: Main Capabilities and Disciplines," 8 April 2019. Accessed 19 February 2020) provides the following definition: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions. […]"
    - Draft IEC 22989 Information Technology: Artificial Intelligence: Artificial Intelligence Concepts and Terminology, published 31 October 2019, defines artificial intelligence as the capability of an engineered system to acquire, process and apply knowledge and skills. Note: Knowledge are facts, information, and skills acquired through experience or education.

16. IMDRF (2020), Artificial Intelligence Medical Devices (AIMD). IMDRF website. http://www.imdrf.org/workitems/wi-aimd.asp. Accessed 16 February 2021.

17. Scherer MU. 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies.' 2016 Harvard Journal of Law and Technology; and Miriam C. Buiten, 'Towards Intelligent Regulation of Artificial Intelligence.' 2019 European Journal of Risk Regulation.

18. One way of defining intelligence is provided by Steven Pinker. He defines intelligence as the ability to deploy novel means to attain a goal, whereby the goals are extraneous to the intelligence itself. It also should be noted that intelligence is a multi-dimensional variable, i.e., with many facets to it (e.g., visual, motor, mathematics, language); hence, just like humans, an AI can be intelligent on one facet, but not on another.

19. Genetic programming is a technique to create algorithms that can program themselves by simulating biological breeding and Darwinian evolution. Instead of programming a model that can solve a particular problem, genetic programming only provides a general objective and lets the model figure out the details itself.

20. While computer programs generally excel at repetitive behavior, certain AI (e.g., probabilistic AI, soft computing, and fuzzy logic), in order to cope with very noisy input data, will employ an execution flow that is very un-deterministic, but will nevertheless each time reach a deterministic answer, within certain margins of tolerance. For example, you can run the program a hundred times; it will never run the same way twice, but it will (should) consistently produce the same result, within the margins specified by the manufacturer. It has inherent variability to tolerate imprecision, uncertainty, and partial truth in real-world data to solve complex problems while achieving tractability, robustness, and low cost. Also, AI may use differential privacy, a technique that carefully injects random noise to protect individual privacy. The goal of such systems is to protect privacy by making it impossible to reverse-engineer the precise inputs used while still delivering an output that is close enough to the accurate answer.

21. At the time of writing, BSI/AAMI is investigating whether a standard can be created to provide guidance and possibly also requirements on algorithm change protocols.

22. Note: It is important to calibrate the uncertainty. What you do not want is AI that gives the wrong answer and is extremely confident in that wrong answer. See also: The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence. 7 January 2019. Nature website. https://www.nature.com/articles/s42256-018-0004-1. Accessed 16 February 2021.

23. Privacy safeguards inhibit storing of necessary data. For example, AI systems that provide video or audio recommendations are updated over time, changing in response to the availability of content and user reactions. The only way such systems could be precisely replicable over time would be if every interaction of every user was stored indefinitely, which would be unacceptable from a privacy point of view and also questionable from an environmental sustainability point of view.

24. As is for example defined through *EU MDR* Art. 2 (40) and Art. 52(1) and *EU IVDR* Art. 2 (32) and Art. 48(1).

25. As is for example defined through *EU MDR* Art. 10(9) and *EU IVDR* Art. 10(8).

26. As is for example defined through *EU MDR* and EU *IVDR* Art. 7(d).

27. Annex VII Requirements to be met by Notified Bodies Section 4.9, Annex IX Conformity Assessment based on a quality management system and on assessment of technical documentation, Chapter II Assessment of the technical documentation Section 4.10, and Annex X Conformity Assessment based on Type-Examination Section 5 Changes to the type.

28. *EU MDR* Art. 27(3) and *EU IVDR* Art. 24(3).

29. *EU MDR* and *EU IVDR* Art. 5(5).

30. Draft IEC 22989 Information Technology: Artificial Intelligence: Artificial Intelligence Concepts and Terminology. Published 31 October 2019.

31. The European Commission's High-Level Expert Group on AI in its Assessment List for Trustworthy AI (17 July 2020) distinguishes human-in-the-loop (HITL, i.e., direct control), human-on-the-loop (HOTL, i.e., supervisory control) and human-in-command (HIC). Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal, and ethical impact) and the ability to decide when and how to use the AI system in any particular situation. The latter can include the decision not to use an AI system in a particular situation to establish levels of human discretion during the use of the system or to ensure the ability to override a decision made by an AI system. European Commission website. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence. Accessed 16 February 2021.

32. Autonomy, artificial intelligence, and robotics: Technical aspects of human control, ICRC, August 2019.

33. Urias MG, Patel N, He C, et al. Artificial intelligence, robotics, and eye surgery: are we overfitted? *Int J Retin Vitr 5, 52 (2019).*

34. High throughput computing can handle situations much faster than the human brain can let a signal pass from eye to brain to hands. Any device that ultimately relies solely or primarily on human attention and oversight cannot possibly keep up with the volume and velocity of algorithmic decision-making or will necessarily be outmatched by the scale of the problem and hence be insufficient.

35. Ironies of automation, *Bainbridge Automatica,* Vol. 19, No. 6, 1983.

36. Wikipedia. Level 2 ("hands off"): The automated system takes full control of the vehicle: accelerating, braking, and steering. The driver must monitor the driving and be prepared to intervene immediately at any time if the automated system fails to respond properly. The shorthand "hands off" is not meant to be taken literally – contact between hand and wheel is often mandatory during Level 2 driving to confirm that the driver is ready to intervene. Wikipedia website. https://en.wikipedia.org/wiki/Self-driving_car. Accessed 2 February 2020.

37. Overtrust is a form of risk compensation. Goddard K, Roudsari A, and Wyatt J. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*. 19(1)p121–127, 2012. Wikipedia. Accessed 3 March 2021.

38. Jorritsma W, Cnossen F, Van Ooijen PMA. Improving the radiologist-CAD interaction: Designing for appropriate trust. (2014). *Clinical Radiology*. 70.10.1016/j.crad.2014.09.017.

39. The US FDA obliged the manufacturer to remove the human from the loop. K. Cobbaert (personal communication, 17 February 2020).

40. For example, *EU MDR* and *EU IVDR* Annex I GSPR 5: In eliminating or reducing risks related to use error, the manufacturer shall:

    - Reduce as far as possible the risks related to the ergonomic features of the device and the environment in which the device is intended to be used (design for patient safety)
    - Give consideration to the technical knowledge, experience, education, training and use environment, where applicable, and the medical and physical conditions of intended users (design for lay, professional, disabled or other users).

41. Klein, et al. "Ten challenges for making automation a 'team player' in joint human-agent activity." *IEEE Computer*. Nov/Dec 2004.

42. Transparency is defined as open, comprehensive, accessible, clear, and understandable presentation of information (ISO 20294:2018, 3.3.11) or as openness about activities and decisions that affect stakeholders and willingness to communicate about these in an open, comprehensive, accessible, clear, and understandable manner (draft IEC 22989 Information Technology: Artificial Intelligence: Artificial Intelligence Concepts and Terminology. Published 31 October 2019.

43. This phenomenon is called emergence. Emergence occurs when an entity is observed to have properties its parts do not have on their own. Similar to how the function of a biological organism emerges from the interaction of its cells: properties emerge in the organism that are not present and understood when looking at the cellular level. Wikipedia website. https://en.wikipedia.org/wiki/Emergence. Accessed 16 February 2021.

44. Kearns M and Roth A. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. 1 November 2019, Oxford University Press.

45. Other methods are emerging for programmatically interpretable reinforcement learning, in which the black box and other models become not the ultimate output of the learning process, but an intermediate step along the way. As an example, consider DeepMind's OCT AI. It uses optical coherence tomography (OCT) scans. These 3D images provide a detailed map of the back of the eye. DeepMind split the AI in two parts. The first AI identifies all that is abnormal, such as bleeding in retina, leakage of fluid, or water logging of the retina. It highlights all those features. The second AI then categorizes them, for example, diabetic eye disease with a percentage representing the confidence. Even if the AI appeared to have it wrong, some of cases made the scientists realize that the algorithm had noticed something that the healthcare professionals had not spotted. Some of those cases were very ambiguous, challenging cases, and the researchers realized that their gold standard might have to be adapted. DeepMind Podcast, Episode 5 "Out of the lab." 27 August 2019. https://deepmind.com/blog/article/podcast-episode-5-out-of-the-lab. Accessed 16 February 2021.

46. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. London AJ, Hastings Cent Rep. 2019;49(1):15–21. doi:10.1002/hast.973.

47. Explicability: Property of an AI system that important factors influencing the prediction decision can be expressed in a way that humans would understand. Modified from "explainability" as defined by draft IEC 22989 Information Technology: Artificial Intelligence: Artificial Intelligence Concepts and Terminology. Published 31 October 2019.

48. *GDPR* Art. 13(2) (f) controllers must, at the time when the personal data are obtained, provide the data subjects with further information necessary to ensure fair and transparent processing about the existence of automated decision-making and when such is the case, include meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. Art. 22 The data subject shall have the right not to be subject to a decision based solely on automated processing […] which […] significantly affects him or her. EUR-Lex website. https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed 16 February 2021.
    […] data controller shall implement suitable measures to safeguard the data subject's rights, freedoms, and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express

his or her point of view and to contest the decision. For guidance related to automated-decision making and profiling under *GDPR*, see wp251rev.01 published by Art. 29 Working Part (WP29), re-endorsed by European Data Protection Board (EDPB) as *GDPR: Guidelines, Recommendations, Best Practices*. EDPB is the successor of WP29. EDPB is the entity comprising the national information protection authorities. It guards and enforces the consistent implementation of the *GDPR* across the EU. European Commission website. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053. Accessed 16 February 2021.

49. There are many different interpretations of fairness. The European Parliament Panel for the Future Science and Technology (STOA) in its study, Artificial intelligence: From ethics to policy (June 2020), defines fairness as the requirement to equally distribute goods, wealth, harms, and risks. The *General Data Protection Regulation (GDPR*) refers to substantive fairness (Recital 71) as fairness of the content of an automated inference or decision, according to the STOA report. The impact of the *GDPR* on artificial intelligence (25 June 2020) can be summarized as (a) Acceptability, i.e., the input data (the predictors) for the AI decision being normatively acceptable as a basis for the inferences concerning individuals (e.g., the exclusion of ethnicity if this does not impact disease determination), (b) Relevance: the inferred information (the target) should be relevant to the purpose of the decision and normatively acceptable in that connection, (c) Reliability: both input data, including the training set, and the methods to process them should be accurate and statistically reliable (which is an aspect that must be proven in the clinical evaluation or performance evaluation of a medical device). On the other hand, the *GDPR* also refers to informational fairness (Recital (60): that the subject must be informed of the existence of the processing operation and its purposes. The High-Level Expert Group on Artificial Intelligence considers fairness to have both a substantive and a procedural dimension. The substantive dimension implies a commitment to ensuring equal and just distribution of both benefits and costs and ensuring that individuals and groups are free from unfair bias, discrimination, and stigmatization. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. Ethics Guidelines for Trustworthy AI, 8 April 2019. The European Parliament Panel for the Future of Science and Technology (STOA) defines fairness as the requirement to equally distribute goods, wealth, harms, and risks. Artificial Intelligence: From Ethics to policy. 24 June 2020. European Parliament website. https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2020)641507. Accessed 16 February 2021.

50. Artificial intelligence: From ethics to policy. European Parliament Panel for the Future of Science and Technology (STOA). June 2020. European Parliament website. https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2020)641507. Accessed 16 February 2021.

51. Winfield A, et. al. The Ethics of Artificial Intelligence: Issues and initiatives. Panel for the Future of Science and Technology (STOA) Panel. 11 March 2020. Science Communication Unit at the University of the West of England at the request of STOA. European Parliament website. https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2020)634452. Accessed 16 February 2021.

52. Op cit 44.

53. Op cit 50.

54. Lebeer G. Ethical Function in Hospital Ethics Committees. IOS Press, 2002.

55. A statistic and a parameter are very similar. They are both descriptions of groups. The difference between a statistic and a parameter is that statistics describe a sample. A parameter describes an entire population. A statistic is biased if it is calculated in such a way that it is systematically different from the population parameter being estimated.

56. Article 4 (m) of draft Report with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)). 4 April 2020. Committee on Legal Affairs. European Parliament website. https://www.europarl.europa.eu/doceo/document/JURI-PR-650508_EN.html?redirect Accessed 16 February 2021.

57. The scientific definition of bias uses the words "tendency" and "statistic," which alludes to there being a systematic nature.

58. Women, because it is harder to find a woman that is not childbearing to participate in clinical investigations; minority racial or ethnic groups because it is harder to find enough people of that subgroup; elderly, because it sometimes carries more risks to include them. Fox-Rawlings SR, Gottschalk LB, Doamekpor LA, Zuckerman DM, and Milbank Q. Diversity in Medical Device Clinical Trials: Do We Know What Works for Which Patients? 2018;96(3):499–529. doi:10.1111/1468-0009.12344.

59. Hébert-Johnson U, Kim MP, Reingold O, Rothblum GN. Multicalibration: Calibration for the (Computationally-Identifiable) Masses." Cornell University website. https://arxiv.org/abs/1711.08513. Accessed 16 February 2021.

60. Jobin A, et.al. Artificial Intelligence: the global landscape of ethics guidelines. June 2019. ResearchGate website. https://www.researchgate.net/publication/334082218_Artificial_Intelligence_the_global_landscape_of_ethics_guidelines. Accessed 16 February 2021.

61. High-Level Expert Group on AI (2020). Ethics Guidelines for Trustworthy AI.

62. NMPA (September 2019). Key Review Points for Deep Learning Decision Support Medical Devices Software.

63. Digital Health Software Precertification (Pre-Cert) Program. FDA website. https://www.fda.gov/medical-devices/digital-health/digital-health-software-precertification-pre-cert-program. Accessed 16 February 2021.