Artificial intelligence: Characteristics, regulatory compliance, and legislation

This article provides an introduction to artificial intelligence (AI), its characteristics, and how those characteristics affect regulatory compliance. It also examines efforts to regulate the ethical aspects of AI in the EU, future legislative initiatives that may affect AI in medical devices, and the crucial role standards play in supporting legislation.

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

Artificial intelligence refers to a wide variety of techniques. While neural networks are in the spotlight today, this article covers all forms of AI, including classical AI (e.g., search and optimization techniques, expert systems, and older forms of computer vision); symbolic AI (e.g., logical reasoning and decision making); abstract syntax tree modifying code, probabilistic reasoning, machine learning, knowledge representation, planning and navigation, natural language processing, and perception. Hybrid approaches also exist, which use a combination of techniques, such as 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. Figure 1 presents a simplified diagram of the different types of AI.

Artificial general intelligence (AGI) can learn incrementally, reason abstractly, and act effectively over a wide range of domains. The author does not anticipate AGI to appear on the market in the near or medium term, so this article will focus on “narrow AI,” that is, artificial intelligence with a specific, rather than general, purpose.

Current medical device legislation applies to machine learning devices. It appears fit to assure safety and reliable performance, including AI that continues changing after it is placed on the market. 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.

Definition
Having one common, internationally accepted definition for AI would be helpful
when comparing AI investments and regulations across the world. There are several definitions of AI, each with different flavors. At the time of writing, the International Medical Device Regulators Forum (IMDRF) was drafting a definition for machine learning medical devices.

In trying to capture the essence of AI, definitions tend to focus on AI’s learning ability or its intelligent nature, two aspects that pose significant interpretation issues in a legal context. Regarding AI’s intelligent nature, there is no scientific consensus on the meaning of “intelligence.” 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 article 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 program-
Artificial intelligence: Characteristics, regulatory compliance, and legislation

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, that is, “the global model.” After the validation and conformity assessment, if applicable, the device is deployed into clinical use through one or more local models. 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 2).

**Federated learning**

Global change is irrespective of whether the global model learns based on local data or on data collected separately, such as 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 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 – for example, in health institutions. This approach contrasts with traditional centralized machine learning.
techniques, in which all the local datasets are uploaded to one server, and more classic decentralized methods that often assume local data samples are identically distributed are used. 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 make computations directly on the encrypted data without decrypting them beforehand.

Predetermined change

Local change occurs in machine learning devices that adapt, refine their performance, or calibrate themselves to a specific patient or health care 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.

Predetermined 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.

Predetermined changes through machine learning-enable personalized health care 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 forcing 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 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 look-up 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, for example, if the AI contains nondeterministic 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, for example, 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, involving 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 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 interleukin 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, there may be hybrid scenarios, whereby the model continues to evolve, but the user has the ability to revert to an earlier state of the model or to the factory default. It is worth noting that 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 substantially (the IMDRF refers to “substantial” changes). 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 Medical Device Regulation (EU MDR) and EU In Vitro Diagnostic (EU IVDR), a health institution developing a machine learning device for in-house use must also 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 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 predetermined change envelope (also known as change boundaries or prespecifications), for example, for a purpose not covered by the intended use or for more specific purposes than claimed by the manufacturer; or
- While not complying with the manufacturer’s ACP.

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

In this 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
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, Article 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 authority 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. Of note is that the US Food and Drug Administration (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 the 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, there is no standard available for guiding 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 for,
• What caused a substantial 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 levels22 are during a given timeframe, and
• 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, for example, 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 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 predetermined ACP require an update of the technical documentation, including the clinical evaluation and a new conformity assessment.
Assume a natural or legal person (an individual or company, respectively) wants to place an existing device on the market in the EU either by changing its intended purpose or by modifying it in a way in which compliance with the applicable requirements may be affected. In those cases, the person will 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 an EU 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 change the intended purpose or 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 included in the scope of the medical device definition, for example, 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 for in-house use.

**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 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.

---

**Figure 6. Illustration of AI change types**

- **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
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 interaction problems, such as a lack of operator situational awareness (sensing uncertainties/limited situational awareness), that is, 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 may require the human to be taken out of the loop because of the user’s limited decision making capacity, 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.

**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 requires a human in the loop to drive the car safely. It might have to travel 100,000 km before the car has an accident; the human may by then have put too much confidence (overtrust) and automation bias 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 medical domain can inspire. For example, findings from a radiology study have shown that the human-device combination’s accuracy might improve when a computer-aided detection algorithm identifies more than one choice to the radiologist. 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 combined 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 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. On the one hand, 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. Inappropriate levels of trust in automation may cause suboptimal performance.

**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, that is, did the doctor or the user make a mistake? Was it the AI, incorrect or unforeseen input data, or a malicious actor, for example, hackers or disgruntled employees?

Transparency may also be needed 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 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. 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 it is we want to be interpretable, that is, 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 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 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 we can understand or rationalize post hoc, a phenomenon known as emergence. For example, a doctor can review the output of an algorithm reporting 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.

Explained variance, that is, “Given a blank sheet, what would be the minimum input data needed to receive this decision?” is on the opposing end of counterfactuals, that is, “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.

“Inexplicable” devices are not unusual because health care 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 through randomized, controlled studies that electroconvulsive 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 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. 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, that is, AI with no actionable outcome, may not require transparency toward the health care provider or patient but requires sufficient explicability 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 2016/679 (GDPR) 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 techniques, for example, to improve a sensor or optics within a device), a software device must inform users of that fact, 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. However, 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 it can mean different things in different contexts to different people.

To consider the ethical aspects of software that changes itself through learning, the European Parliamentary Research Service 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 health care, in which medical experimentation requires manufacturers to be explicit about the experimental nature of health care 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. These differ based on the organization's goals.
and operating contexts. Such frameworks have limits, for example, because many AI systems comprise a trade-off between algorithmic fairness and accuracy.\textsuperscript{44,50} On the one hand, a fair algorithm should provide the same benefits for the group while protecting it from discrimination. On the other hand, an accurate algorithm should make a prediction that is as precise as possible for a certain subgroup, for example, according to age, gender, smoking history, previous illnesses, and so on. When an algorithm lies on the trade-off 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 who 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 health care 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 health care.\textsuperscript{52} 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.\textsuperscript{53} From a legal perspective, bias is any prejudiced or partial personal or social perception of a person or group.\textsuperscript{54} It is beyond this article’s scope to discuss the differences between the scientific and legal definition; suffice it to say we attribute a broader meaning to the legal definition, mainly because the scientific definition generally is understood to refer to systematic estimation errors.\textsuperscript{55} 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 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.\textsuperscript{56} 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.\textsuperscript{57} The challenge is for computer scientists to get an accurate picture of which group(s) could be 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 nondiscrimination 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. They can use these characteristics to establish 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.

### AI legislation

**European Union**

Legislators across the world are focusing on ethical aspects of AI and the presence of bias. A 2019 heat map published by Anna Jobin shows that the number of published AI ethics guidelines has increased, especially...
people often attribute a broad meaning to ethics. Ethics guidelines from the High-Level Expert Group on Artificial Intelligence, for example, include algorithmic as well as data aspects: Nondiscrimination, 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 carbon dioxide emission regulations, although they are not as specific to AI as digital rights advocates and the European Parliament would like. The European Commission recently published a legislative proposal to address the ethical aspects of AI. The EU is also assessing whether and how to address the liability and intellectual property of AI.

China

The Center for Medical Device Evaluation (CMDE), a division of the Chinese regulatory authority, the National Medical Products Administration (NMPA), issued comprehensive requirements encompassing scrutiny of machine learning devices across their entire device lifecycle. The National Institute for Food and Drug Control, another NMPA division, supplements these requirements with a growing body of standards, for example, to characterize the data sets used for the training or evaluation of AI (see Figure 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, 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)63 published by the World Health Organization, this Chinese initiative is a welcome start.

United States

In the US, the FDA published a discussion paper that focuses on machine learning devices that change during runtime, citing the Software Precertification (Pre-Cert) Program as a possible regulatory pathway for AI. The 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 premarket 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 toward 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. According to personal communication from K. Cobbaert on 8 May 2021, a regulatory sandbox is a framework that provides a structured context for legislators to create and evaluate draft legislation by assessing its impact on a technology, product, service, or an approach and its stakeholders. Where appropriate, such experiment may occur in a real-world environment with appropriate safeguards and under regulatory supervision, while existing legislation is not enforced for a limited time and in a limited part of a sector or area.

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 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.

Among the standardization bodies working in the AI space are:

- International Organization for Standardization;
- International Electrotechnical Commission;
- International Telecommunication Union;
- Institute of Electrical and Electronics Engineers;
- European Committee for Standardization and the European Committee for Electrotechnical Standardization, also known as CEN-CENELEC;
- British Standards Institution; and
- Association for the Advancement of Medical Instrumentation.

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. However, the downside is that companies active in different countries that use competing standards face higher costs, that is, 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 the domain of AI certification. Such hidden agendas bring the risk of more complex standards or those that do not bring value to society, other than to those that sell certification services.

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.
tors 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.

Abbreviations

About the author
Koen Cobbaert, MSc, is senior manager of quality, standards, and regulations at Philips, where he guides and supports the company’s business groups in the EU MDR transition and other EU regulations. He also chairs the Software Work Group at COCIR, which represents medical imaging, radiotherapy, health ICT, and electromedical industries in Europe and is a project team member with the European Committee for Standardization’s CEN/Technical Committee 251 and tech transfer adviser at the Industrial Research Fund. Cobbaert holds master’s degrees in risk management and bioengineering. He can be contacted at koen.cobbaert@gmail.com

*Acknowledgment
This article is based in Software as a Medical Device – Regulatory and Market Access Implications. The full citation for the chapter is:


References and footnotes
5. Logical reasoning: 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. Source: Logical thinking. In: Seel NM (ed). Encyclopedia of the sciences of learning. Springer; 2012. https://doi.org/10.1007/978-1-4419-1428-6_4800 Accessed 4 May 2021.
8. Machine learning: A process using computational techniques to enable systems to learn from data or experience (IEC 23053, 3.16; under development, not accessible as of 10 May 2021).
9. Knowledge representation and reasoning is the field of artificial intelligence dedicated to representing information about the world in a form that a computer system can use to solve complex tasks such as diagnosing...

10. Natural language processing: Information processing based on natural language understanding and natural language generation (CD2 ISO-IEC 22989, 3 September 2020; under development, not accessible as of 10 May 2021).


12. There are two schools of thought in AI – that of Yann LeCun, who thinks we can reach artificial general intelligence via deep learning alone; and that of Gary Marcus, who thinks other forms of AI are needed, notably symbolic AI or hybrid forms. Source: [Not peer-reviewed] Marcus G. Deep learning: A critical appraisal. arXiv. 2018.


18. Psychologist and linguist Steven Pinker 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, that is, 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: 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. Computer programs generally excel at repetitive behavior, but certain AI (e.g., probabilistic AI, soft computing, and fuzzy logic), in order to cope with very noisy input data, will use an execution flow that is highly 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 publication, the Medical Device Innovation Coalition (MDIC), with participation of the FDA, is investigating whether a standard can be created to provide guidance and possibly also requirements on algorithm change protocols.

22. 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: Begoli E, et al. The need for uncertainty quantification in machine-assisted medical decision making. https://www.nature.com/articles/s42256-018-0004-1 Published 7 January 2019. Accessed 4 May 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 were 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. 2 (32) and Art. 52(1) 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 distinguishes human-in-the-loop (HITL, that is, direct control), human-on-the-loop (HOTL, that is, 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.


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.


36. Wikipedia. Self-driving car – Classifications, Levels of driving automation, 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. https://en.wikipedia.org/wiki/Self-driving_car Accessed 4 May 2021.

37. Overtrust is a form of risk compensation.


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

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).”


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 29989 Information Technology: Artificial Intelligence: Artificial Intelligence Concepts and Terminology. Published 31 October 2019.

43. 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 4 May 2021.


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 optical coherence tomography (OCT) AI. It uses OTC 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 seemed to have it wrong, some of cases made the scientists realize that the algorithm had noticed something that the health care 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.
48. Controllers must, at the time of obtaining personal data, 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 (GDPR, Art. 13: Information to be provided where personal data are collected from the data subject. https://gdpr-info.eu/art-13-gdpr/).

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 (GDPR, Art. 22: Automated individual decision-making, including profiling. https://gdpr-info.eu/art-22-gdpr/).


For guidance related to automated decision making and profiling under the General Data Protection Regulation, see: European Union. General Data Protection Regulation (GDPR) – Chapter 4, Article 29, Processing under the authority of the controller or processor. https://gdpr.eu/article-29-processing-under-controller-or-processor/ Accessed 4 May 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 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 can be summarized as: (a) Acceptability, that is, 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; and (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 AI 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.


53. 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.
55. The scientific definition of bias uses the words “tendency” and “statistic,” which alludes to there being a systematic nature.
56. These would include: Women, because it is harder to find a woman who is not childbearing to participate in clinical investigations; minority racial or ethnic groups, because it is harder to find enough people of those subgroups; and the elderly, because there is sometimes greater risk to include them.