7

AI, Explainability, and Safeguarding Patient Safety in Europe

Toward a Science-Focused Regulatory Model

Barry Solaiman and Mark G. Bloom

7.1 INTRODUCTION

This chapter explores the efforts made by regulators in Europe to develop standards concerning the explainability of artificial intelligence (AI) systems used in wearables. Diagnostic health devices such as fitness trackers, smart health watches, ECG and blood pressure monitors, and other biosensors are becoming more user-friendly, computationally powerful, and integrated into society. They are used to track the spread of infectious diseases, monitor health remotely, and predict the onset of illness before symptoms arise. At their foundation are complex neural networks making predictions from a plethora of data. While their use has been growing, the COVID-19 pandemic will likely accelerate that rise as governments grapple with monitoring and containing the spread of infectious diseases. One key challenge for scientists and regulators is to ensure that predictions are understood and explainable to legislators, policymakers, doctors, and patients to ensure informed decision making.

Two arguments are made in this chapter. First, regulators in Europe should develop minimum standards on explainability. Second, those standards should be informed by the computer science underlying the technology to identify the limitations of explainability. Recently, several reports have been published by the European Commission and the National Health Service (NHS) in the United Kingdom (UK). This chapter examines the operation of AI networks alongside those guidelines finding that, while they make good progress, they will ultimately be limited by the available technology. Further, despite much being said about the opaqueness of neural networks, human beings have significant oversight over them. The finger of liability will remain pointed toward humans, but the technology should advance to help them decipher networks intelligibly. As computer scientists enhance the technology, lawmakers should set minimum standards that are leveled-up progressively as the technology improves.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

7.2 WEARABLES IN HEALTH CARE

Wearables are devices designed to stay on the body and collect health data such as heart rate, temperature, and oxygenation levels.[1] Smartwatches, chest belts, clothing, ingestible electronics, and many others are converging with the internet-of-things (IoT) and cloud computing to become powerful diagnostics for more than seventy conditions.[2] The technology has advanced rapidly, with GPUs, CPUs, and increasing RAM being adopted, opening possibilities for deep learning.[3] Despite these advances, adoption remains low in the health care setting overall, being in the early stages of the Gartner Hype Cycle.[4] Nevertheless, the trend is moving toward greater adoption. The COVID-19 pandemic in 2020 may accelerate the development of telemedicine, monitoring patients remotely, predicting disease, and mapping the spread of illnesses.[5] An example of the technology's use can be seen in England under an NHS pilot program where patients were fitted with a Wi-Fi-enabled armband. This monitored vital signs remotely, such as respiratory rates, oxygen levels, pulse, blood pressure, and body temperature. AI was able to monitor patients in real-time, leading to a reduction in readmission rates, home visits, and emergency admissions. Algorithms were able to identify warning signs in the data, alerting the patient and caregiver.[6] This example aligns with a broader trend of adoption.[7] The largest NHS hospital trusts have signed multiyear deals to increase the number of wearables used for remote digital health assessments and monitoring.[8] This allows doctors to monitor their patients away from the hospital setting, both before and after medical procedures.

[1] Aras D. Dargazany et al., Wearable DL: Wearable Internet-of-Things and Deep Learning for Big Data Analytics-Concept Literature and Future, 1 Mobile Info. Systems 4 (2018).

[2] NHSX, Artificial Intelligence: How to Get it Right – Putting Policy into Practice for Safe Data-Driven Innovation in Health and Care, 18 (Oct. 2019), www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf; Sara Gerke et al., Ethical and Legal Issues of Ingestible Electronic Sensors, 2 Nature Electronics 329 (2019).

[3] Sourav Bhattacharya et al., From Smart to Deep: Robust Activity Recognition on Smartwatches Using Deep Learning, IEEE (2016), https://userpages.umbc.edu/~nroy/courses/shhasp18/papers/From%20Smart%20to%20Deep%20Robust%20Activity%20Recognition%20on%20Smartwatches%20Using%20Deep%20Learning.pdf.

[4] NHSX, supra note 2, at 20; Department of Health and Social Care (UK), The AHSN Network: Accelerating Artificial Intelligence in Health and Care (2018), https://wessexahsn.org.uk/img/news/AHSN%20Network%20AI%20Report-1536078823.pdf.

[5] Fight Covid-19 through the Power of the People, Stan. Med. (2020), https://innovations.stanford.edu.

[6] Moni Miyashita & Michael Brady, The Health Care Benefits of Combining Wearables and AI, Harv. Bus. Rev. (2019), https://hbr.org/2019/05/the-health-care-benefits-of-combining-wearables-and-ai.

[7] Such adoption may lead to unintended consequences, such as unregulated yet sophisticated apps marketed as low-level medical devices which may lead to doctors becoming overburdened with requests. See Helen Yu, Regulation of Digital Health Technologies in the EU: Intended versus Actual Use, in Innovation and Protection: The Future of Medical Device Regulation (I. Glenn Cohen et al. eds., 2021).

[8] Laura Donnelly, NHS Experiment in AI Will See Whole City Offered Virtual Hospital Appointments and Diagnosis by Chatbot, Telegraph (Jan. 23, 2020), www.telegraph.co.uk/news/2020/01/23/nhs-experiment-ai-will-see-whole-city-offered-virtual-hospital/.

7.3 HUMAN NEURAL NETWORKS?

Underpinning such technologies is complex computer science. A device can predict illness, but it cannot explain why it made a prediction, which raises several legal issues. A targeted legal strategy cannot be realistically devised without understanding the technology driving it. Lawyers are unlikely to become master coders or algorithm developers, but they can have a reasonable understanding of where most efforts are needed. By examining what drives AI, more technically aware discussions can be generated in the legal sphere.

AI is an umbrella term used for different forms of "machine learning." This includes "supervised" and "unsupervised" learning, which entails making predictions by analyzing data.[9] The former involves predefined labels used to assign the data to relevant groups, whereas the latter searches for common features in the data to classify it. A subset of "machine learning" is "deep learning," which consists of artificial neural networks (ANNs) used for autonomous learning. There are various architectures, but the primary example here is of a deep supervised learning network with labeled data.[10] Such networks are the most numerous in operation and can illustrate how deep learning works and where the legal issues may arise.

Figure 7.1 depicts a neural network. An ANN begins with an "input layer" on the left.[11] The example is an image of a cerebellum, which the ANN converts into many "neurons" (represented by the grid of squares). Each neuron is assigned a value (for black and white images) called the "activation" number. The number could, for example, be higher for brighter neurons (where the cerebellum is) and lower for darker neurons (outside the cerebellum). Every neuron is represented in the input layer. The example shows four neurons, but the ANN will have as many neurons as there are pixels in the image.

The example also shows two hidden layers in the middle, but there will be numerous in practice. In reality, the layers are not hidden to the programmer, but their numerousness makes the ANN virtually undecipherable – much like a human brain. The activations in the input layer (the black circles) will influence what is activated in the first hidden layer (the light grey circles) which will influence further activations. At the end is the output layer with several choices (Cerebellum, frontal lobe, or pituitary gland). The ANN gives the highest value to its choice (here, Cerebellum, the dark grey circle). Between the neurons are connections called "weights" (represented as lines) whose values are determined by a mathematical function. The sum of the weights in one layer determines which neurons are

[9] Also "reinforcement" learning. Stuart Russell & Peter Norvig, Artificial Intelligence: A Modern Approach 830 (3rd ed. 2010).

[10] See, e.g., deep Boltzmann machine, spike neural networks. Aras, 7.

[11] But What Is Neural Network?, YouTube (Oct. 5, 2017), www.youtube.com/watch?v=aircAruvnKk; Russell & Norvig, supra note 9; Ron Sun, Connectionism and Neural Networks, in The Cambridge Handbook of Artificial Intelligence (Keith Frankish & William M. Ramsey eds., 2014).

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

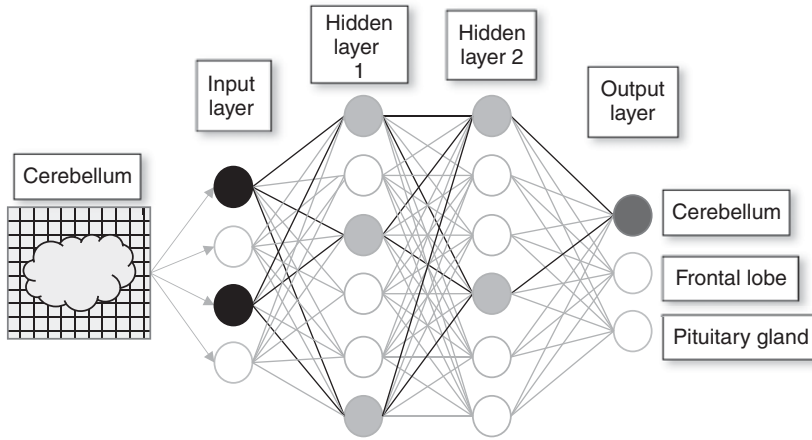

FIGURE 7.1: Example of an ANN

activated in the next layer. For example, the sum of the weights in the input layer has activated the first, third, and sixth neurons in the first hidden layer. Humans can also influence those activations by adding a "bias" to alter the value required for an activation.

In practice, the numerousness and complexity of the connections create an undecipherable matrix of distinct weights and biases. The choice of output cannot be explained, which is where the term "black box" algorithms arises. Despite this, humans play a central role. They give the network training data consisting of many prelabeled images of cerebellums, pituitary glands, and frontal lobes. The network is trained on that data. The process of data moving from left to right is called "forward propagation," and the weights between the neurons are initially random, which produces random outputs. To correct the ANN, a validation data set is used with labels indicating the correct answer. In response, the ANN works backward (back-propagation) from the output layer, through the hidden layers to the input layer, adjusting the weights and biases as it moves along. The network becomes more accurate through repetition.

Deep supervised learning networks are well suited to diagnostics. Inputs, such as scans, are in a standardized format, which is a useful source of structured input data, training, and validation. The process becomes highly accurate because of the numerous hidden layers and connections. However, the black-box nature of an ANN should not be overstated. Humans have significant involvement, labeling data, giving it to the network, providing feedback, computing biases, interpreting data, and putting it into practice.

Most studies of wearable data have focused on supervised learning architectures.[12] However, data derived from wearables is often unlabeled and unstructured, which

[12] See, e.g., Oscar D. Lara et al., A Survey on Human Activity Recognition Using Wearable Sensors, 15 IEEE Commc'n Surveys & Tutorial 1199 (2012).

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

benefits unsupervised learning that identifies patterns to make predictions.[13] These techniques raise more complex legal issues because humans are less involved. They are a work in progress at present, but they will become more prominent.[14] Indeed, there are increasing studies that analyze wearable data using unsupervised ANNs. One study proposes an unsupervised ANN to classify and recognize human activities.[15] It was able to recognize human activities through a combination of data obtained from magnetometers and accelerometers in wearables.[16] Another study analyzed data from 3D accelerometers on the wrist and hip, a skin temperature sensor, an ECG electrode, a respiratory effort sensor, and an oximeter amongst others.[17] The unsupervised network yielded 89 percent accuracy in detecting human activities (walking, cycling, playing football, or lying down).[18] Another approach has analyzed gestures to detect daily patterns that might indicate when older persons require assistance.[19]

These are a small sample of studies increasingly utilizing unsupervised learning architectures in wearables. The underlying point is that such technologies are being used more frequently, which raises legal issues surrounding explainability. At the same time, humans must still train the networks. The processes within the hidden layers are complex to decipher, but humans pretrain and oversee the process.[20] Consequently, while the legal implications must be deciphered, the autonomous nature of these systems should not be overstated.

7.4 EXPLAINABILITY AND THE LAW

Explainability refers to ex-ante explanations of an ANN's functionality, and ex-ante or ex-post explanations of the decisions taken, such as the rationale, the weighting, and the rules.[21] It requires that humans can understand and trace decisions.[22] However, the regulation of an ANN is as complex as its operation, which is problematic in health care. While shortcomings in explainability of AI systems will not necessarily lead to liability, it is one important factor. The key point of

[13] Aras, 5–6; Stuart Russell & Peter Norvig, supra note 9, at 695.
[14] Aras, 15; Stanford.
[15] Lukun Wang, Recognition of Human Activities Using Continuous Autoencoders with Wearable Sensors, 16 Sensors 189, 2–3 (2016).
[16] Id. at 15.
[17] Miikka Ermes et al., Detection of Daily Activities and Sports with Wearable Sensors in Controlled and Uncontrolled Conditions, 12 IEEE Transactions on Information Technology in Biomedicine 20, 21 (2008).
[18] Id. at 24–5.
[19] Alessandra Moschetti et al., Towards an Unsupervised Approach for Daily Gesture Recognition in Assisted Living Applications, 17 IEEE Sensors Journal 8395, 8402 (2017).
[20] Sourav, 2.
[21] Sandra Wachter et al., Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, 7 Int'l Data Privacy L. 76, 78 (2017).
[22] European Commission, Ethics Guidelines for Trustworthy AI: High-Level Expert Group on Artificial Intelligence 18 (Apr. 8, 2019).

interaction between explainability and liability is at the fact finding or evidence stage. It may be difficult to factually prove the harm caused by a neural network because one cannot explain how a certain input resulted in a specific output, and that a deficiency resulted due to that process.[23] The circumstances in which explainability becomes important in liability analyses are broad.

Problems may arise where harm is caused to a patient because the doctor did not follow the appropriate standard of care.[24] Price notes how, in the current climate, the risk of liability for doctors relying on AI recommendations is significant because the practice is "too innovative to have many adherents."[25] Algorithm developers might be liable as well. However, in Europe, the laws are incoherent. The Product Liability Directive (1985/374/EEC) holds manufacturers liable for defective products. Proving that an ANN was defective requires technical expertise, but even experts cannot explain the hidden layers of a network.[26] There is also the problem of ANNs being autonomous and changing. While the European Union has taken a strict approach on manufacturers being liable for the safety of products throughout their lifecycle, it acknowledges that the Directive should be revisited to account for products that may change or be altered thereby leaving manufacturers in legal limbo.[27]

There are also medical device regulations, but half of developers in the United Kingdom do not intend to seek CE Mark classification because it is uncertain whether algorithms can be classified as medical devices.[28] There are medical device conformity assessments, but there are no standards for validating algorithms nor regulating adaptive algorithms.[29] Also, while manufacturers must carry out risk

[23] Expert Group on Liability and New Technologies, Liability for Artificial Intelligence and Other Emerging Digital Technologies, European Commission 1, 54 (2019), https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en/format-PDF; European Commission, White Paper on Artificial Intelligence – A European Approach to Excellence and Trust, 13 (2020), https://templatearchive.com/ai-white-paper/.

[24] W. Nicholson Price II et al., Potential Liability for Physicians Using Artificial Intelligence, 322 JAMA 1765, 1765 (2019).

[25] W. Nicholson Price II, Medical Malpractice and Black-Box Medicine, in Big Data, Health Law, and Bioethics 301 (I. Glenn Cohen et al. eds., 2018).

[26] European Commission, supra note 22, at 13.

[27] European Commission, Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics (2020), https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligence-feb2020_en_1.pdf; this is known as the "update problem." See I. Glenn Cohen et al., The European Artificial Intelligence Strategy: Implications and Challenges for Digital Health, 2 Lancet Digital Health e376, e377 (2020), www.thelancet.com/action/showPdf?pii=S2589-7500%2820%2930112-6; on "system view" approach to regulation, see Sara Gerke et al., The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device, 3 Digital Me. 1 (2020); Timo Minssen et al., Regulatory Responses to Medical Machine Learning, J. L. & Biosciences 1, 6 (2020).

[28] Regulation on Medical Devices (Regulation 2017/745) (EU); In Vitro Diagnostic Medical Device Regulation (IVDR) (Regulation 2017/746) (EU); NHSX, How to Get It Right, supra note 2, at 22.

[29] Id.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

assessments before products are placed on the market, they quickly become outdated because ANNs continuously evolve.[30] For doctors, they may be negligent when advising patients based on AI recommendations that later cause harm. There are also questions about whether a person can consent to flawed medical advice from an ANN. These challenges are recognized in Europe where several reports were published in 2019 and 2020.

7.5 GUIDELINES

In the European Union, there are Guidelines, a White Paper, and an Assessment List regarding AI geared toward developing a future regulatory framework. On the guidelines, the EU Commission set up an "independent group" which released the Ethics Guidelines for Trustworthy AI in 2019 seen as a "starting point" for discussions about AI premised on respect for human autonomy, prevention of harm, fairness, and explicability.[31] The White Paper (which builds upon the Guidelines) was published in 2020 and outlines an approach to AI based on "excellence and trust."[32] It notes that while AI can improve prevention and diagnosis in health care, black box algorithms create difficulties of legal enforcement.[33] The Assessment List for Trustworthy Artificial Intelligence (ALTAI) is a self-assessment list published in July 2020.[34]

7.5.1 *Guidelines, Explainability, and the GDPR*

In the Guidelines, the principle of "explicability" is of primary relevance. It requires that AI processes and decisions are transparent and explainable to those involved.[35] The Guidelines emphasize that this may not always be possible with black box algorithms and, "in those circumstances, other explicability measures (e.g., traceability, auditability, and transparent communication on system capabilities) may be required."[36] Auditability and transparent communication are likely within easiest reach from a technical standpoint. The accuracy of the training data used can be verified, and the specific tasks undertaken by humans developing the network can be checked. Traceability is the greatest challenge owing to the hidden layers.

The Guidelines highlight several principles that may help in realizing explainability. First, "human agency," which refers to humans understanding AI systems

[30] European Commission, supra note 27, at 6.
[31] European Commission, supra note 22, at 3.
[32] European Commission, supra note 23.
[33] European Commission, supra note 23, at 1, 10.
[34] European Commission, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment (2020), https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
[35] Id. at 13.
[36] Id.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

and challenging them.[37] AI can shape human behavior and should support informed decision making.[38] The issue is whether a doctor is liable for advice given that was informed by AI recommendations. Of relevance is Article 22 of the General Data Protection Regulation (GDPR) concerning automated decision making and profiling which protects individuals from decisions "based solely on automated processing, including profiling, which produces legal effects."[39] The Information Commissioner's Office (ICO) in the United Kingdom requires that individuals must have the right to obtain human intervention, express their point of view, an explanation of the decision and the ability to challenge it.[40]

Taken at its most extreme, there would be an automatic infringement of a medical decision based solely on the automated processing of an ANN, and a right to an explanation. However, it has been argued that a "right to explanation" does not exist under the GDPR, but rather a limited right to be "informed" of system functionality.[41] In other words, a right only to ex ante explanations of system functionality at the data collection stage, rather than ex post explanations of the decisions that have been made once the data has been propagated and an output generated.[42] Furthermore, a right to explanation has existed for many years in different jurisdictions but has not led to greater transparency because copyright protections have precluded algorithms from being revealed.[43] The general distinction is that persons might be entitled to know of specific data used in a neural network, but are not entitled to know the weights, biases and, statistical values.[44] The extent of the right is very narrow and would, in any case, be limited to those bringing a claim rather than general laws on explainability setting minimum standards.

Further, it would be a rare scenario indeed for a decision to be made "solely" by AI as required under Article 22. In practice, AI is used to supplement informed decisions rather than make them. It is also unlikely that AI outputs can result solely from "automated" processing because humans are always involved in some capacity.[45] Most fundamentally, the wording of Article 22 requires that automated processing has "legal effects" on the individual. However, an ANN will not interfere with the right not to consent, nor to withdrawing consent once it has been given. Although,

[37] Id. at 16.
[38] Id.
[39] GDPR 2016/679 and the UK Data Protection Act 2018 (DPA).
[40] Information Commissioner's Office, Right Not to Be Subject to Automated Decision-Making (2020), https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-law-enforcement-processing/individual-rights/right-not-to-be-subject-to-automated-decision-making/ [hereinafter ICO].
[41] Wachter et al., supra note 21, at 79, 90; Further, an individual's right to know about how personal data is evaluated, is significantly curtailed by ECJ jurisprudence. See Sandra Wachter & Brent Mittelstadt, A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI, 1 Colum. Bus. L. Rev. 1, 6–7 (2019).
[42] Id. at 82.
[43] Id. at 86.
[44] Id. at 87.
[45] Id. at 92.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

the law might protect those relying on wearable tech giving flawed advice that would interfere with their right to informed consent.

Matters are further complicated by an interrelated provision in the GDPR concerning "profiling," which is any "automated processing of personal data" used to predict aspects concerning a person's health.[46] Wearables may combine individual health data with broader user data to provide individualized advice. Users relying on them would be unable to assess why the advice was given nor challenge it, which undermines the aims of "human agency" in the Guidelines. Additionally, nothing in the law appears to preclude automated processing where the individual consents.[47] The law could protect individuals by requiring a minimum level of explainability in such cases.

A related matter is "human oversight and autonomy." This is most practically achieved through "human-on-the-loop" or "human-in-command" approaches.[48] The former requires that humans can both intervene in designing a system and monitor it. The latter refers to holistic oversight over a network. The Guidelines recommend that the less oversight a human has, the more extensive testing and stricter governance is required.[49] However, for neural networks to work, humans must intervene and monitor a system, both granularly and holistically. Without such oversight, the neural network would produce "garbage" outputs. Networks can be tricked easily, and even slight changes to the data can cause them to fail.[50] The Guidelines, therefore, overstate the significance of these principles.

Other principles are "technical robustness and safety" and "human oversight and autonomy." Networks could be required to change procedures or ask for human intervention before continuing an operation when encountering a problem. A network should indicate the likelihood of errors occurring, be reliable, and have reproducible outputs.[51] This requires adequate transparency, which entails principles of "traceability" and "communication."[52] Traceability means documenting outputs of the ANN, the labeled data, the datasets, and data processes.[53] Communication means revealing the AI's capabilities and limitations.[54] Returning to the GDPR, Article 22(3) requires "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."

Two matters arise here. First, what human involvement means. Second, when should humans get involved? The former could mean humans replacing automated decisions without algorithmic help; a human decision taking into account the

46 GDPR, art. 4(4); see also ICO, supra note 40.
47 GDPR, art. 22(2)(C), art. 9(2).
48 European Commission, supra note 22, at 16.
49 Id.
50 Jory Heckman, DARPA: Next Generation Artificial Intelligence in the Works, Federal News Network (Mar. 1, 2018), https://federalnewsnetwork.com/technology-main/2018/03/darpa-next-generation-artificial-intelligence-in-development/.
51 European Commission, supra note 22, at 17.
52 Id. at 18.
53 Id.
54 Id.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

algorithmic assessment, or humans monitoring the input data based on a person's objections and a new decision made solely by the network.[55] It could also mean that a data controller must provide ex-ante justifications for any inferences drawn about the subject's data to determine whether the inference was unreasonable.[56] A risk-based approach could determine the latter. Thus, the riskier the recommendation by an ANN, the more checks required.[57] However, this would be limited to procedural rather than substantive validation, such as appropriately training doctors for using AI.[58] Further, a risk-based approach would still be unable to assess the reasons for AI recommendations.

Much remains undetermined regarding what these factors mean in practice for explainability. The White Paper recognizes that these principles are not covered under current legislation and promises feedback later.[59] For now, it proposes distinct forms of human oversight such as blocking AI systems not reviewed and validated by humans; allowing systems to operate temporarily as long as human intervention occurs afterward; ensuring close monitoring of networks by humans once they are in operation and that networks can be deactivated when problems arise; or imposing operational constraints on networks during the design phase.[60] Such oversight could assist in finding inaccurate input data, problematic inferences, or other flaws in the algorithm's reasoning.[61] It could form part of procedural evaluations of black-box algorithms noted by Price.[62] However, a key question is how such factors may apply in practice, which is why the Commission also released an Assessment List (ALTAI). The ALTAI list contains two questions on explainability, but they are minimalist. The first asks whether the decision of a neural network was explained to users. The second asks whether users were continuously surveyed about whether they understood the decisions of a network.[63] There are other potentially useful questions regarding human oversight and the other principles noted above, but it is the NHSX approach that is of most practical significance.

7.5.2 *Practical Implementation*

The NHS Code of Conduct for Data-Driven Health and Care Technology may provide a practical solution. Principle 7 focuses on explainability. It states: "Show what type of algorithm is being developed or deployed, the ethical examination of

[55] Sandra Wachter et al., Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, 31 Harv. J. L. & Tech. 842, 873 (2018).
[56] Wachter & Mittelstadt, supra note 41, at 7.
[57] By developers and independent external auditors. Price, supra note 25, at 295, 301.
[58] Id. at 304.
[59] European Commission, supra note 22, at 9.
[60] Id. at 21.
[61] Wachter, supra note 55, at 37.
[62] Price, supra note 25, at 305.
[63] European Commission, supra note 34, at 14–15.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

how the data is used, how its performance will be validated and how it will be integrated into health and care provision."[64] The outputs should be explained to those relying on them, the learning methodology of the ANN should be transparent, the learning model and functionality specified, its strengths and limitations and compliance with data protection.[65]

To assist developers, there is a "how-to" guide detailing what is expected when developing AI.[66] Four processes are relevant here. First, reporting the type of algorithm developed, how it was trained and demonstrating that adequate care was given to ethical considerations in the input data.[67] For this, a "model card" or checklist approach is proposed for explaining those aspects of the ANN.[68] Second, provide evidence of the algorithm's effectiveness through external validation, communicating early with NHSX on the proposed method of continuous audit of inputs and outputs, and how they were determined.[69] Third, explain the algorithm to those relying on their outputs, detail the level of human involvement, and develop languages that are understandable to the layperson.[70] Fourth, explain how a decision was "made on the acceptable use of the algorithm in the context of it being used."[71] This may involve speaking to patient groups to assess their thinking on the acceptable uses of AI and monitor their reactions to gauge acceptance of the technology.[72]

The Code is significant because it indicates how minimum standards for explainability might operate in the context of an ANN. However, it is undetermined how the factors might be realized or whether a uniform approach would work for all neural networks. A pilot Trustworthy AI Assessment List has been proposed in the Commission's Guidelines with questions on traceability and explainability.[73] The questions on traceability concern detailing the method of programming and testing – those on explainability concern the ability to interpret outputs and ensuring that they can be explained. The questions are useful but remain idealistic for deriving sense from the hidden layers. The technological limitations mean that other ideas in the Guidelines are more practicable at present. This includes a "white list" of rules

64 NHSX Code of Conduct, https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology.
65 Id.
66 Id. at 29.
67 Id. at 31.
68 Id. at 31; Margaret Mitchell et al., Model Cards for Model Reporting, FAT* '19: Conference on Fairness, Accountability, and Transparency 1, 3 (Jan. 2019).
69 NHSX, How to Get It Right, supra note 2, at 32; this approach aligns with Leong Tze Yun's recommendation that AI systems should be systemically examined and validated; see Gary Humphreys, Regulating Digital Health, 98 Bulletin of the World Health Organization 235, 235 (2020), www.who.int/bulletin/volumes/98/4/20-020420.pdf.
70 Id.
71 Id.
72 Id.
73 European Commission, supra note 22, at 24–31.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press

that must always be followed and "black list" restrictions that must never be transgressed.[74]

While such requirements could provide minimum standards for explainability, there are some aspects of neural networks that remain unexplainable. If networks do not provide insight into their continuously evolving reasoning, it will be impossible to achieve detailed insight arising from any checklist. For this reason, researchers are developing new technologies surrounding "algorithmic transparency." This includes auditing techniques and interactive visualization systems.[75] It is beyond the scope of this chapter to explore these in detail, but one example involves the creation of a "deep visualization" toolbox that examines the activation of individual neurons.[76] Working backward, researchers can map out different neurons and determine which one influences the other. The activated neurons can be viewed in real-time to see which parts of an image the neuron is highlighting.[77] As this technology develops further, lawyers and policymakers should remain alert to incorporating standards developed in this field into the explainability requirements of guidelines and regulations. One day, they could form part of the minimum standards for explainability.

7.6 CONCLUSION

The foundations for setting minimum standards concerning explainability have now been established. However, there are shortcomings in AI-enhanced technology, such as wearables, which undermine informed decision-making for doctors, patients, and others. This is problematic because wearables will become ever more heavily relied upon for a wide variety of medical purposes. Further, doctors and patients ought to know why neural networks produce specific outputs. In time, scientists will develop more sophisticated models of explainability. Regulators, doctors, patients, and scientists should work together to ensure that those advances filter into the relevant guidelines as they develop – a gradual and flexible "leveling up" that keeps apace with the science. In this manner, lawyers and policymakers should take responsibility for better understanding the technology underlying those systems. As such, they should become more familiar with and knowledgeable about neural networks, the use of input data, training data, how data propagates, and how "learning" occurs. This will be key for creating standards that are relevant, sound, and justified. While laws and guidelines in the future will indicate the path to be pursued, some matters will take concerted interdisciplinary efforts to resolve.

[74] Id. at 21.

[75] Information Commissioner's Office (UK), Big Data, Artificial Intelligence, Machine Learning and Data Protection 86 (2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

[76] Jason Yosinski et al., Understanding Neural Networks through Deep Visualization, Deep Learning Workshop, 31st International Conference on Machine Learning (2015).

[77] Id.

https://doi.org/10.1017/9781108975452.008 Published online by Cambridge University Press