ORF ISSUE BRIEF

SEPTEMBER 2020
ISSUE NO. 401

# ‘Reasonable Explainability’ for Regulating AI in Health

YUKTI SHARMA
ABHINAV VERMA
KRISSTINA RAO
VIVEK ELURI

**ABSTRACT** Emerging technology is slowly finding a place in developing countries for its potential to plug gaps in ailing public service systems, such as healthcare. At the same time, cases of bias and discrimination that overlap with the complexity of algorithms have created a trust problem with technology. Promoting transparency in algorithmic decision-making through explainability can be pivotal in addressing the lack of trust with medical artificial intelligence (AI), but this comes with challenges for providers and regulators. In generating explainability, AI providers need to prioritise their accountability to patient safety given that the most accurate of algorithms are still opaque. There are also additional costs involved. Regulators looking to facilitate the entry of innovation while prioritising patient safety will need to look into ascertaining a reasonable level of explainability considering risk factors and the context of its use, and adaptive and experimental means of regulation.

Attribution: Yukti Sharma, Abhinav Verma, Krisstina Rao, and Vivek Eluri, "‘Reasonable Explainability’ for Regulating AI in Health," *ORF Issue Brief No. 401*, September 2020, Observer Research Foundation.

ISBN 978-81-947783-5-6

## INTRODUCTION

Artificial intelligence (AI) models across the globe have come under the scanner over ethical issues; for instance, Amazon's hiring algorithm reportedly discriminates against women,[1] and there is evidence of racial bias in the facial recognition software used by law enforcement in the United States (US).[2] While biased AI has various implications, concerns around the use of AI in ethically sensitive industries, such as healthcare, justifiably require closer examination.

Medical AI models have become more commonplace in clinical and healthcare settings due to their higher accuracy and lower turnaround time and cost in comparison to non-AI techniques. These systems can now rival clinician's performance in diagnostic applications, for instance, the detection of atrial fibrillation from an electrocardiogram (ECG)[3] or predicting the onset of sepsis before clinician recognition.[4] With its potential to plug gaps in health systems—such as resource shortages, strained tertiary health setups, lagging research and medical institutions—medical AI is a frontrunner among innovations of the Fourth Industrial Revolution. Its increased use, however, has uncovered a 'trust' problem with the technology—for instance, AI seems to perform much better while detecting skin cancer on Caucasian skin and misses potentially malignant lesions on darker-skinned people.[5] The automated decision-making AI offers could also exacerbate the underrepresentation of ethnic minorities[6] and women[7] in traditional medical research. Such biased outcomes can exacerbate discriminations inherent in health systems, taking public health further away from its welfare function. Overcoming this trust deficit will require AI to be more transparent with its inner workings. But this is easier said than done.

The underpinning theme here is the opacity of AI algorithms. The most efficient and complex AI algorithms are notoriously opaque, more commonly referred to as 'black boxes', offering little reasoning, if at all, as to how they arrive at conclusions without being programmed to do so. In addition to discriminatory outputs that result from bias, AI output is often counterintuitive to trained clinician output.

When first marketed to clinicians as a supercomputer, IBM Watson was perceived as incompetent when its output contradicted recommendations of clinicians. Discovering that the algorithm was too complex to explain its output further exacerbated the lack of trust that users and patients had in AI-based medical procedures.[8] Overcoming this opacity requires investing in building explainability[a] through product development processes that emphasise the use of ethically sourced and treated data, and generating post hoc explainability through the use of secondary algorithms. However, this is a double-edged sword—more complex AI would need more

[a] 'Explainability' is the ability to understand the inner mechanics and functioning of a system and explain it in human terms. This is important to understand how an algorithm works and subsequently identify 'why' and 'how' a decision was reached. AI models have very sophisticated and often convoluted decision mechanisms owing to the extensive data that is processed in the models. With the non-transparent internal encoding and inherent complexity, complete explainability is yet not possible to achieve. 'Reasonable explainability' is the ability to provide reasoning behind certain decisions or predictions.

time and resource investment to make them explainable, but they are also more accurate due to their ability to engage with complex variables.

Regulating this black box is thus a complex and multifaceted issue, especially since most jurisdictions have yet to succeed in comprehensively regulating even simple and mostly explainable AI. The role of the regulator is to establish a balance between many competing interests, the most complex of which involves upholding patient safety against a potential adverse event in the context of the most efficient innovation being unexplainable. Further, the regulator must also consider possible ramifications of the level of explainability (or unexplainability) relative to accuracy and efficiency, especially when it comes to the imputation of liability in case of adverse events. This is all the more essential in fault-liability regimes where developers can escape legal liability by adhering to prescribed norms.

## THE PROMISE OF AI IN HEALTHCARE

As AI gets integrated into clinical workflows and competes with the gold-standard of doctor efficiency by becoming more complex and multi-layered, there is fear that the strong demand for transparency and oversight might stonewall any progress in healthcare applications of AI, especially in areas where medical understanding and abilities are limited. For instance, doctors detect anomalies in a patient's cardiac activity (ischemia or rhythm disturbances) merely by observing a handful of features in the ECG waveform components. They must rely on these limited features because of the high-pressure settings they operate in with rigid time constraints. Their capacities are even more limited if they have to observe multivariate features in a more complex Halter monitor.[9] Algorithms, by contrast, can systematically analyse every heartbeat far more comprehensively, and even identify subtle microscopic variations that can serve as early signs for major cardiac issues.[10] This may be outside human ability.

AI is also steadily responding to the needs of patients who require time-critical care, wherein the absence of rapid-diagnosis technology that AI leverages might lead to their diseases being undiagnosed.[11] The time-sensitive leverage of medical AI becomes even more relevant when large parts of medical practice frequently reflect a mixture of empirical findings and inherited clinical culture.

Furthermore, AI's promise is most prominent in areas where medical research has shown little success. For instance, the failure of the decade-long search for neuroprotective interventions against Alzheimer's disease shows current limitations of medical knowledge, which is where AI has transformative potential.[12] The potential of AI to respond to critical challenges of clinical knowledge and its implementation are thus well established, and, when combined with its ability to supplement strained public health system resources, indicate that its advent into healthcare is due.

### The need for regulation

Concerns over AI's lack of explainability closely flow from the murky accountability frameworks that surround it. AI is recognised for its ability to identify and predict patterns in data that unassisted humans might not be able to. Counterintuitive outcomes are traceable for specific AI—it is possible to track with reasonable certainty why an AI made a particular decision, what factors affected it, and where liability needs to be placed in the case of a mishap, be it biased outcomes or wrongful prognosis. However, complex AI lacks transparency regarding its decision pathways. This also means that any accountability framework cannot pinpoint to the exact element of the algorithmic process that faltered to create an adverse outcome. In the absence of an accountability framework, liability, when needed to be established, is understood as a medley of subjective factors, such as the potential of harm, the possibility of bias, and the possibility of correction or compensation.[13]

The opacity of an algorithm can be a result of intentional secrecy, technical illiteracy or unintelligible complex mathematical representations.[14] The first two kinds require regulatory standards and guidelines for safeguarding its users along with equipping clinicians with technical know-how, whereas exposing the third kind can be used to prevent or rectify errors, mitigate biases, and develop trust in the algorithm's decision-making process. As in scientific and academic research, open access enables peer review to ensure that only accurate, relevant conclusions contribute to the sector's discourse.[15] Explainability can enable similar accountability for outputs of AI models.

While the need for explainability through 'transparency' has been articulated by international forums like the Organisation for Economic Co-operation and Development (OECD)[16] and G20,[17] the unambiguous right to seek meaningful information about the existence, logic and projected consequences of automated decision-making systems is still not established, prompting debate around the necessity of explainability and its limits. Consequently, the US Defense Advanced Research Projects Agency has launched an Explainable AI (XAI) program aimed at producing explainable models to enable human users to understand and appropriately trust the system.[18] Other jurisdictions like France have taken a stricter stance by explicitly stipulating that a 'black box' algorithm cannot be used.[19]

It will not be long before cost and quality efficiencies bring advanced machine learning closer to addressing complex clinical decision support systems as well. As the Indian public health infrastructure embraces digital health, it is crucial to solve for explainability.

## COMPLICATIONS

### General transparency versus explainability

Given that the regulation for AI through an iterative process of clinical evaluation and performance monitoring is not inexpensive, ensuring patient safety while not over-regulating is critical. As articulated in software medical device regulations,[20] the degree of regulation required is assessed against the device's risk category, which is a function of the significance of the information it provides on the healthcare condition and the healthcare condition itself. An AI software that informs supply chain management for vaccination, therefore, need not be subject to the same level of scrutiny as a device that diagnoses pathology confirming the presence of a malignant tumour.

The yardstick for defining an optimal level of evidence for safety and performance is also applied to explainability. Relying on the critical distinction between interpretability (making something clear) and completeness (revealing the entirety of mathematical operations and parameters in the system), the 'right' level of explainability is relative to the context.

Researchers at Georgia Tech developed a machine-learning system that comprises of neural networks capable of action and simultaneously translating it into an 'explanation'.[21] Called 'rationale generation,' an AI agent (character) plays a video game (Frogger) and provides rationales in plain English to justify its actions. The result is an impressive Frogger-playing AI that verbally formulates explanations like "I'm moving left to stay behind the blue truck" every time it moves.[22] While rationale generation is a critical principle in developing responsible AI, using a separate machine learning algorithm that generates an explanation in natural language is complex and nearly unattainable for advanced algorithms.

A critical challenge to generating explainable AI is creating models that are interpretable and complete in their transparency offering. The most accurate explanations are not easily interpretable to people, and conversely, the most interpretable descriptions often do not provide predictive power.[23] This points to the ethical dilemma of creating persuasive systems versus transparent systems. Oversimplistic explanations may not do justice to the complex system. An effective guide to explainability emerges from an understanding of its context—for whom the AI is being made explainable and why.

Research shows four categories of factors influence the 'form' of explainability—audience or recipient factors (who is receiving information and what they need to know); impact factors (aligning level of explainability with a degree of risk); regulatory factors (rights or regulations that the application of AI engages with); and operational factors (like user trust and safety certification that influence the level of explainability determined to be necessary). Further, attention to context emphasises the degree of explainability—global (making the algorithm explainable in its entirety) versus local (ability to explain a particular algorithmic decision).[24]

A contextual understanding of explainability is useful in juxtaposing the need for transparency against the costs of generating it. Certain transparency risks might be limited to revealing source data, but others might need to articulate the decision pathway to establish the factors that most heavily correlate with a specific outcome.

### Explainability-Accuracy trade-off

Deep learning is at the forefront of machine learning for healthcare solutions. Its accuracy allows it to perform better than traditional AI methodologies and makes it especially useful in clinically high-stakes settings.[25] This was seen in the case of Mount Sinai's 'Deep Patient' in 2015. By applying unsupervised deep feature learning to EHRs[b] of about 700,000 patients, it substantially improved the prediction performance for severe diabetes, schizophrenia and various forms of cancers.[26]

With their ability to classify objects in an image with increased accuracy, these algorithms have found applications in medical imaging based specialties, such as radiology, dermatology and oncology,[27] demonstrating success in operating faster and more accurately than their human counterparts. Algorithms can now assess diagnostic images of the retina for a variety of 50 different retinal diseases and suggest which patient needs urgent medical attention at high accuracy.[28,29] However, due to the black box problem, their internal functioning and certainty about how they reach their output is unknown, further adding to the trust deficit, uncertain liability norms and complications of integrating technology in a largely human-driven field.

A sobering perspective comes from understanding the human-reliant gold standard of medical diagnosis, which itself is imperfect. Physician-led medical diagnosis is a complex process of iterative hypothesis creation and evidence generation, despite which it is nearly impossible to achieve 100 percent accuracy. With the time-sensitive nature of critical diagnoses, diagnostic iterations for certainty are not practical. Therefore, medical errors, especially diagnostic errors, are still a significant concern in the human-led systems. India, for instance, witnesses 5.2 million medical errors annually,[30] but patients still trust physicians more than machines.[31]

AI in medical diagnosis can address this problem. Apart from rapid diagnostic patterns learnt from large data sets, it can circumvent common cognitive biases that are otherwise prevalent in physician-led diagnoses and instead conduct treatment decisions based on normative standards laid down by expert clinicians.

The accuracy-explainability trade-off is seen in how trust governs AI adoption in highly regulated industries like healthcare. Assuming the epistemic position of an algorithm (or the validity of its output) is established, a case of 'disagreement' between an algorithm's output and a clinician's diagnosis (the gold standard) is quickly resolved by tracking the decision pathway of the algorithm or providing an 'explanation'. The inability to extract an explanation from the algorithm will result in an impasse—is it clinically ethical to trust a diagnostic route that can be interpreted or one that overlooks interpretability and favours data-driven accuracy? At the same time, opaque AI-based decisions also overlook a primary premise of patient treatment—complete transparency and comprehensive information being available to the patient to make an informed choice.[32] However, in traditional medical settings and scenarios, it is the doctor that plays the intermediary role of processing the AI-outcome, analysing it and communicating the prognosis to the patient. This begs the

[b] Electronic Health Records (EHRs) are the digital/electronic version of patient's medical records.

question of whether AI explainability should be perceived from the perspective of the user-centricity of the healthcare professional or the common patient.

The lack of explainability for high-stakes solutions is only permissible if one is willing to take a leap of faith with the output of complex AI solutions. Healthcare places an equal onus on medical advancement (for its efficiency) and the role of the clinician (for generating trust in a certain medical solution, and some accountability herein). AI is only likely to prevail if it responds to and integrates both of these functions played by current medical interventions. Ongoing research into making deep learning models explainable, sheds some light on local explainability methods that can be integrated by design or as post hoc mechanisms to derive explanations for outputs.[33] It also highlights the need for research in exploring this further as AI learns to make more complex decisions.

## ACCESSIBLE XAI AND THE COSTS OF EXPLAINABILITY

While the benefits of making AI algorithms explainable include higher trust in and accountability of the technology product, explainability itself is not inherent to the design of AI-based technology. The need to generate an explanation for an output tends to increase with the consequence of the output, creating what is called a sliding scale of explainability, representing its trade-off with the accuracy of the algorithm. Yet it could be argued that despite the short-term costs imposed on innovators, XAI has long-term legal benefits that might be prudent for healthcare innovators to invest in.

Expenses like design costs (explicitly designing an explanation function in the algorithm, which caters to the contextual need for explanation and cannot be standardised) and the creation of audit logs (storing algorithmic decisions for a specified period after they are created to provide local explanations on decisions made) could directly deter innovators from investing in explainability. Furthermore, opportunity costs in the form of lower accuracy and violation of trade secrets make explainability a less accessible option for innovators.[34] Revealing proprietary algorithms and their inner workings to consumers and regulators can lead to a loss of the competitive advantage innovators have spent considerable time and resources to attain. All this might achieve little because pure explainability of the inner workings of AI does not ensure understandability by the common patient. In such a scenario, explainability only works to breed trustworthiness (in healthcare facilities and professionals) and transparency (for certification and vigilance agencies).

The objective of the regulator should be to propel a fledgling industry that has substantial potential benefits for the healthcare system by making it economically viable and reducing entry barriers. From this lens, imposing high costs to generate explainability and transparency will be counterproductive to the purpose defined. However, ascertaining a 'reasonable' or outcome-focused level of explainability for the technology involved that focuses on ethical documentation and submissions to regulators can provide a useful lens to guard patient safety.

## RECOMMENDATIONS

Given the bias in training data for AI algorithms, it could be argued that the thrust towards responsible AI[35] comes less from beneficiaries or users and more from the institutions of markets and social justice that will be at the forefront of dealing with its consequences. There are two conflicting human tendencies that cannot be truly quantified but still need to be traded-off for a patient-centric view on explainability—the automation bias or the inclination to place blind trust in automated decisions, and the value of human trust. Some user research indicates that users are likely to forego better healthcare to have a human, rather than an AI, care provider.[36] Resolving the dilemmas inherent to the pursuit of explainability is thus in the best interest of regulators that prioritise patient safety as well as their trust.

The philosophy of explainability that places the onus on the innovator of AI-assisted technology to promote 'transparency' is highlighted in the guiding principles of the OECD and G20 and imbibed in the design philosophy of leading innovators (such as Google's 'Be accountable to people' and Microsoft's 'Transparency as an approach to AI').[37] Innovators have found ingenious ways to address the explainability-accuracy trade-off without raising their innovation costs, with many embracing a 'glass box' model to AI development. Typically, these are simple-to-train models and can quantify uncertainty in their predictions. Simple glass box models can perform just as well as more complicated neural networks,[38,39] and are thus a potentially more ethical alternative. However, explainability includes transparent visibility over the input data and algorithm design, thereby making trade secrets an inefficient way of protecting investments in AI innovation. Therefore, the role of the regulator must also then extend to developing comprehensive intellectual property regimes to protect AI applications and datasets, something that the World Intellectual Property Organization is considering.[40]

Further, researchers have found ways to add interpretability constraints to deep learning models, which has led to more transparent computations. These interpretability constraints have not come at the expense of accuracy, even for deep neural networks for computer vision.[41] If accuracy and interpretability prove to be a false dichotomy, eliminating black-box models from high-stakes decisions is a call regulators will have to take. Meanwhile, a comfortable balance of the two is possible, and a consultative policy should help secure it.

Regulatory frameworks have attempted to address this. The EU's General Data Protection Regulation (GDPR) that governs personal data places the 'right to an explanation' in the hands of the data principal. Articles herein specifically call out the use of personal data in automated decision-making, highlighting the right to be opted out of consequences of any decision based on an automated decision-making process as well as demand a meaningful explanation of the logic involved when the decision affects those whose data it concerns. However, interpretations of the GDPR limit its legal binding and do not concern non-personal data or data that is anonymised.[42]

Below are some principles that can be considered by regulators while resolving the explainability conundrum.

- **Addressing explainability requirements based on the risk involved**

When designing a safety fence for transparency to operate, regulators are better off acknowledging the high costs inherent to generating higher transparency. Given that a transparency norm that emphasises 100 percent explainability is unattainable, adopting a risk-based explainability approach to determine a reasonable level of explainability is more realistic. The International Medical Device Regulators Forum risk-classification for medical devices is a useful starting point to ascertain the risk involved based on how serious the health condition in concern is and how significant the AI output is to influencing the healthcare decision.[43] Such an approach would delineate the risk based on the scope of impact of the technology; one will be less concerned with how a chatbot recommended a doctor on a telemedicine app than how a machine learning algorithm predicted early-onset Alzheimer's.

- **Providing explanations based on input, process and output norms**

While general transparency about the algorithm's function is alluded to when speaking of 'trust' in artificial systems, explanations are typically sought in the case of an undesirable outcome generated by technology to detect the source of the problem and to establish accountability in cases of liability. In cases of algorithmic predictions, for instance, generally understanding the process of how an algorithm reaches an output is unsatisfactory if a patient has been advised to undergo an expensive preventive surgery as a result of an algorithm's early prediction of cancer. In the case of surgery mishap, the patient will need to be assured of the validity of the algorithm prediction given factors it learned to prioritise in its prediction patterns from its data and how it has succeeded in doing so in the past. The surgeon and hospital involved will need to be assured of the same factors, in addition to understanding if the data it trained on was relevant to the population served by them and hence had validity in its prediction. An insurer would go a step further to inquire about how the algorithm has learned to be extremely cautious with one segment of the population and relatively lax with another segment given their overall life expectancy.

Therefore, in generating explainability, it is important to know if a specific input factor influenced an algorithm's output and perhaps prevented it from making one decision against another.[44] A transparency framework can thus place requirements on the stage at which the evidence is received—input factors (training, testing, operational data), decision-making factors (how a particular input factor relates to a decision), and outcome factors (counterfactuals to justify a decision trajectory).

As benchmarks for explainability are then set up, it would be of critical importance to train users of the AI-based solution (medical practitioners, nurses) on the limitations of explainability, and the risks that may not be explainable, which they will need to communicate with patients.

- **Structural innovations to regulate for explainability sandboxes**

For a regulatory system that is yet to establish precedence with privacy and safety, mandating reasonable transparency requires clear foresight into the kinds of risks that need to be mitigated and how harmful they really are. Given the application of this technology to healthcare, the cost of not being able to mitigate risk is high and not up for experimentation. A regulatory sandbox allows live, time-bound testing of novel products, innovations and technologies under regulatory oversight, and appropriate safeguards.[45] These experimental regulation mechanisms can not only reduce barriers to entry but also allow regulators to collect insights for further regulatory action if necessary before the product is released. Sandboxes can be useful in exploring the fence for 'reasonable explainability'—what is the real risk involved, who will need the explanation and what are viable design or post hoc measures that will respond to the need.

## CONCLUSION

The need for transparency in new technology like AI, especially in healthcare, can be understood in two ways—it either needs to demonstrate how its accuracy is equal to or better than that of a clinician (focusing on accuracy of AI output), or how its bias is the result of biased input and not its process itself (therefore focusing on how it can be resolved). Explainable AI, or AI that is transparent enough to be able to demonstrate the trajectory of its output, is one way of addressing the need for transparency, but represents complex dilemmas for those looking to create conditions for its success.

Although the 'trust' problem in AI may not be that different from trust in any other experience technology (like the internet) and is likely to reduce as users and patients are more familiar with the technology itself,[46] it can prevent the entry of life-saving technology to strained innovation ecosystems such as those in developing countries. As regulatory systems in these countries grapple with the outputs of a thriving innovation ecosystem that holds high promise for safety-first industries like healthcare, acknowledging the complexities inherent to regulation is pivotal to its success. For medical AI, this means examining the need for trust, the stakeholders who need it, and the implications of not being able to trust new technology.

The precedent set by G20, OECD and GDPR provides a useful starting point in the pursuit of explainability. However, contextual guidelines through regulation that merit a reasonable level of explainability and indicate the evidence required to ascertain the same will need to be realised for medical AI to be attractive to innovators. These guidelines can be most beneficial when matched to the risk of the concerned technology and further contextualised to the need for an explanation. In exploring a reasonable level of explainability, regulators who adopt innovative structures such as sandboxes are likely to be ahead of the curve in being a facilitator for innovation as well.

## ABOUT THE AUTHORS

**Yukti Sharma** is a tech innovation and public policy professional with a specialisation in product development and strategy. She is a Young India Fellow.

**Abhinav Verma** is a lawyer with a specialisation in International Law, and a public policy professional working on health systems strengthening.

**Krisstina Rao** has worked in healthcare and education to design grassroots program and advocate for responsive policy. She is an Amani Social Innovation fellow.

**Vivek Eluri** is a healthcare and technology innovation professional. He is a Young India Fellow.

The authors are a project team with the International Innovation Corps, University of Chicago Trust, supported by The Rockefeller Foundation in developing digital health and AI strategies for public health systems in India.

## ENDNOTES

1. Jeffery Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women", *Reuters*, October 11, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

2. Drew Harwell, "Federal Study Confirms Racial Bias Of Many Facial-Recognition Systems, Casts Doubt On Their Expanding Use", *The Washington Post*, December 20, 2019, https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/.

3. AY Hannun et al., "Cardiologist-Level Arrhythmia Detection And Classification In Ambulatory Electrocardiograms Using A Deep Neural Network". *Nature Medicine* 25, no.1 (2019): 65-69. doi:10.1038/s41591-018-0268-3. https://www.nature.com/articles/s41591-018-0268-3

4. Nemati, Shamim et al, "An Interpretable Machine Learning Model For Accurate Prediction Of Sepsis In The ICU", *Critical Care Medicine* 46, no. 4 (2018): 547-553. doi:10.1097/ccm.0000000000002936. https://pubmed.ncbi.nlm.nih.gov/29286945/

5. Adewole S. Adamson and Avery Smith, "Machine Learning And Health Care Disparities In Dermatology", *JAMA Dermatology* 154, no. 11 (2018): 1247. doi:10.1001/jamadermatol.2018.2348.

6. Meghan E McGarry and Susanna A McColley, "Minorities Are Underrepresented in Clinical Trials of Pharmaceutical Agents for Cystic Fibrosis", *Annals of the American Thoracic Society* 13, no. 10 (2016): 1721-1725. doi:10.1513/AnnalsATS.201603-192BC

7. Institute of Medicine (US) Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical Studies, *Health Consequences of Exclusion or Underrepresentation of Women in Clinical Studies*, by Carol S. Weisman and Sandra D. Cassard, Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies, Volume 2: Workshop

and Commissioned Papers, Washington D.C., 1999. https://www.ncbi.nlm.nih.gov/books/NBK236583/

8. Vyacheslav Polonski, “People Don’t Trust AI: Here’s How We Can Change It”, *The Conversation*, January 10, 2020.. https://theconversation.com/people-dont-trust-ai-heres-how-we-can-change-that-87129

9. Ziad Obermeyer and Thomas H. Lee, “Lost In Thought — The Limits Of The Human Mind And The Future Of Medicine”, *New England Journal Of Medicine* 377, no. 13 (2017): 1209-1211. doi:10.1056/nejmp1705348.

10. Zeeshan Syed et al., “Computationally Generated Cardiac Biomarkers For Risk Stratification After Acute Coronary Syndrome”, *Science Translational Medicine* 3, no. 102 (2011): 102ra95. doi:10.1126/scitranslmed.3002557.

11. Alex John London, “Artificial Intelligence And Black-Box Medical Decisions: Accuracy Versus Explainability”, *Hastings Center Report* 49, no. 1 (2019): 15-21. doi:10.1002/hast.973.

12. London, “Artificial Intelligence And Black-Box Medical Decisions: Accuracy Versus Explainability”, at p. 18

13. Finale Doshi-Velez et al., “Accountability Of AI Under The Law: The Role Of Explanation”, *SSRN Electronic Journal* (2017). doi:10.2139/ssrn.3064761.

14. Jenna Burrell, “How The Machine ‘Thinks’: Understanding Opacity In Machine Learning Algorithms”, *Big Data & Society* 3, no. 1 (2016): 205395171562251. doi:10.1177/2053951715622512.

15. Adam Felman, “What To Know About Peer Review”, *Medical News Today*, March 29, 2019. https://www.medicalnewstoday.com/articles/281528.

16. OECD Ministerial Council, *Recommendation Of The Council On Artificial Intelligence (OECD/LEGAL/0449),* 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

17. G20, *Ministerial Statement On Trade And Digital Economy*, 2019, Tsukuba: http://www.g20.utoronto.ca/2019/2019-Ministerial_Statement_on_Trade_and_Digital_Economy.pdf.

18. Defense Advanced Research Projects Agency, “Explainable Artificial Intelligence”, *Darpa.Mil*, 2020, https://www.darpa.mil/program/explainable-artificial-intelligence.

19. “Protection Des Données Personnelles - Séance En Hémicycle Du Mercredi 7 Février 2018 À 21H30 - Intervention De Mounir Mahjoubi - Nosdéputés.Fr”. *Nosdeputes.Fr*, 2020, https://www.nosdeputes.fr/15/intervention/147090.

20. US Food & Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/ Machine Learning (AI/ML)- Based Software as Medical Device (SaMD): Discussion Paper”, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

21. "Explainable AI - Rationale Generation". *GVU Center*, 2020, https://gvu.gatech.edu/research/projects/explainable-ai-rationale-generation.

22. Will Douglas Heaven, "Why Asking An AI To Explain Itself Can Make Things Worse", *MIT Technology Review*, January 29, 2020, https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse/.

23. Leilani H. Gilpin, et al., "Explaining Explanations: An Overview Of Interpretability Of Machine Learning", *Arxiv.Org*, 2020, http://arxiv.org/abs/1806.00069v3.

24. Valérie Beaudouin et al., "Flexible And Context-Specific AI Explainability: A Multidisciplinary Approach", *SSRN Electronic Journal*, 2020. doi:10.2139/ssrn.3559477.

25. Niall O' Mahony et al., "Deep Learning Vs. Traditional Computer Vision", Imar Technology Gateway, Institute of Technology, Tralee, Ireland, https://arxiv.org/pdf/1910.13796.pdf.

26. Riccardo Miotto et al., "Deep Patient: An Unsupervised Representation To Predict The Future of Patients From The Electronic Health Records", *Scientific Reports* 6, no. 1 (2016). doi:10.1038/srep26094.

27. Titus J. Brinker et al., "A Convolutional Neural Network Trained With Dermoscopic Images Performed On Par With 145 Dermatologists In A Clinical Melanoma Image Classification Task", *European Journal Of Cancer* 111 (2019): 148-154. doi:10.1016/j.ejca.2019.02.005: "For e.g. owing to fine grain variability in the appearance of skin lesions, identifying skin cancer is a challenging task that is primarily diagnosed visually followed by dermoscopic analysis, a biopsy and histopathological examination but a single neural network trained on dermascopic images using only pixels and disease level inputs performed on par with 145 dermatologists in a clinical melanoma image classification task"

28. Jeffrey De Fauw et al., "Clinically Applicable Deep Learning For Diagnosis And Referral In Retinal Disease". *Nature Medicine* 24, no. 9 (2018): 1342-1350. doi:10.1038/s41591-018-0107-6.

29. Thomas Grote and Philipp Berens, "On The Ethics Of Algorithmic Decision-Making In Healthcare", *Journal Of Medical Ethics* 46, no. 3 (2019): 205-211. doi:10.1136/medethics-2019-105586.

30. Shahid Akhter, "5.2 Million Medical Errors Are Happening In India Annually: Dr Girdhar J. Gyani", *EThealthworld*, August 2, 2016, https://health.economictimes.indiatimes.com/news/industry/5-2-million-medical-errors-are-happening-in-india-annually-dr-girdhar-j-gyani/53497049.

31. Chiara Longoni and Carey Morewedge, "AI Can Outperform Doctors. So Why Don't Patients Trust It?", *Harvard Business Review*, October 30, 2019, https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it.

32. Brinker, Titus J. et al. "A Convolutional Neural Network Trained With Dermoscopic Images Performed On Par With 145 Dermatologists In A Clinical Melanoma Image Classification Task"

33. Rui Aguiar, "An Overview Of Model Explainability In Modern Machine Learning", T*owards Data Science*, December 4, 2019, https://towardsdatascience.com/an-overview-of-model-explainability-in-modern-machine-learning-fc0f22c8c29a.

34. Valérie Beaudouin et al., "Flexible And Context-Specific AI Explainability: A Multidisciplinary Approach"

35. Term used to encompass ethical, transparent and accountable use of AI technologies that is consistent with expectations, values and norms. Dominic Delmolino and Mimi Whitehouse. *Responsible AI: A Framework For Building Trust In Your AI Solutions*. Accenture, 2018, https://www.accenture.com/_acnmedia/pdf-92/accenture-afs-responsible-ai.pdf. Accessed 13 July 2020.

36. Chiara Longoni and Carey Morewedge, "AI Can Outperform Doctors. So Why Don't Patients Trust It?"

37. Explainable AI: The Basics. The Royal Society, 2019: 30, https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf. Accessed 13 July 2020.

38. Rich Caruana et al., "Intelligible Models For Healthcare: Proceedings Of The 21Th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining - KDD '15" ACM Press (2015). doi:10.1145/2783258.2788613.

39. Cynthia Rudin and Berk Ustun, "Optimized Scoring Systems: Toward Trust In Machine Learning For Healthcare And Criminal Justice", *Interfaces* 48, no. 5 (2018): 449-466. doi:10.1287/inte.2018.0957.

40. WIPO Secretariat, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*. WIPO, 2020, https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf.

41. Cynthia Rudin and Joanna Radin, "Why Are We Using Black Box Models In AI When We Don't Need To? A Lesson From An Explainable AI Competition", *Harvard Data Science Review* 1, no. 2 (2019). doi:10.1162/99608f92.5a8a3a3d.

42. Valérie Beaudouin et al., "Flexible And Context-Specific AI Explainability: A Multidisciplinary Approach"

43. IMDRF Software as a Medical Device (SaMD) Working Group. "*Software As A Medical Device": Possible Framework For Risk Categorization And Corresponding Considerations*. International Medical Device Regulators Forum, 2014: 6-7, http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf.

44. Sandra Wachter et al., "Why A Right To Explanation Of Automated Decision-Making Does Not Exist In The General Data Protection Regulation", *International Data Privacy Law* 7, no. 2 (2017): 76-99. doi:10.1093/idpl/ipx005.

45. Ivo Jenik, *Briefing On Regulatory Sandboxes*, United Nations Secretary-General'S Special

Advocate For Inclusive Finance For Development, https://www.unsgsa.org/files/1915/3141/8033/Sandbox.pdf.

46. William H. Dutton and Adrian Shepherd. “Trust in the Internet as an Experience Technology”, *Information, Communications & Society* 9, no. 4 (2006). https://doi.org/10.1080/13691180600858606

ORF OBSERVER RESEARCH FOUNDATION

Ideas • Forums • Leadership • Impact

20, Rouse Avenue Institutional Area, New Delhi - 110 002, INDIA
Ph. : +91-11-35332000 Fax : +91-11-35332005
E-mail: contactus@orfonline.org
Website: www.orfonline.org