REPORT 8 OF THE COUNCIL ON SCIENCE AND PUBLIC HEALTH (A-25)
Explainability of Artificial/Augmented Intelligence and Machine Learning Algorithms
(Reference Committee E)

EXECUTIVE SUMMARY

BACKGROUND. In continuance of the American Medical Association's (AMA) interest in the oversight and regulation of augmented intelligence (AI) and machine learning (ML) algorithms in the medical system, the Council on Science and Public Health has initiated a report to examine the concept of explainability in the context of AI/ML algorithms.

Briefly, "explainable AI" (XAI) describes AI/ML-enabled algorithms whose decisions would be understandable to an expert in the field and can be evaluated for accuracy or other external factors. At its core, the implementation and use of XAI is a question of physician autonomy. In instances where algorithms can explain their decision-making, they are useful tools for scraping huge data sets and recognizing patterns that may have been imperceptible to the clinician. In those cases, those algorithms are augmenting the physician's skillset, and the outputs can be viewed as recommendations, which can be discarded if the physician's training and expertise disagree with the conclusion. However, if clinical AI algorithms are not explainable, the clinician's training and expertise is removed from decision-making, and they are presented with information they may feel compelled to act upon without knowing where it came from or being able to assess accuracy of the conclusion.

METHODS. English language articles were selected from searches of PubMed and Google Scholar using the search terms "AI explainability," "black box algorithm," and "white box algorithm." Additional articles were identified by manual review of the reference lists of pertinent publications. Web sites managed by government agencies and applicable organizations were also reviewed for relevant information.

DISCUSSION. Generally, humans have low trust of AI, particularly when it is involved in decision-making. One proposed approach for rectifying this lack of trust is through AI devices providing explanations for how it arrived at a conclusion. While XAI is highly appealing for building trust, it is less clear that it is implementable for medical applications. A common fallacy when evaluating AI/ML algorithms is to assume that they approach problems and datasets in the same way that a human would. Algorithms cannot feel, intuit, or infer, but rather perform a series of highly complex calculations. This report discusses several key concepts around explainability, including the motivations for its use, the technical feasibility, regulatory approaches used in both medical and non-medical fields, and recommendations for moving towards more trustworthy AI development.

CONCLUSION. Ironically, the concept of explainability is hard to explain. It is complex, nuanced, and asks profound questions on human cognition and the meaning of trust. The appeal of explainability is clear – particularly in medicine, where decisions can have life or death consequences. Being told a computer has decided you do not qualify for treatment would be disturbing for patients and their physicians. Physicians have trained for decades to utilize context and see the whole patient's clinical picture, and a binary "yes/no" output from a black box may feel restrictive and lacking nuance. Instead, medicine may wish to leverage its experience with other unexplainable phenomena, and push for explainability while requiring safety and efficacy.

REPORT OF THE COUNCIL ON SCIENCE AND PUBLIC HEALTH

CSAPH Report 8-A-25

Subject: Explainability of Artificial/Augmented Intelligence and Machine Learning Algorithms

Presented by: John T. Carlo, MD, MS, Chair

Referred to: Reference Committee E

INTRODUCTION

In continuance of the American Medical Association's (AMA) interest in the oversight and regulation of augmented intelligence (AI) and machine learning (ML) algorithms in the medical system, the Council on Science and Public Health have has initiated this report to examine the concept of explainability. To keep the focus of this report narrow, it is the Council's intent to regularly examine issues relevant to AI/ML's intersection with science and public health and develop policy recommendations as necessary. This report will also serve as an opportunity to define key concepts related to AI in a field where groups are using different definitions of the same term, leading to confusion.

Briefly, "explainable AI" (XAI) describes algorithms whose decisions would be understandable to an expert in the field and can be evaluated for accuracy or other external factors. AI/ML tools in the medical setting can take a variety of forms, ranging from those which interpret images, make diagnostic recommendations, or have a conversation with a patient using a large-language model (LLM). All of these tools would theoretically be able to be developed using an XAI framework. An example of this includes an algorithm trained to distinguish between pictures of wolves and huskies to help track endangered species.[1] The program became very good at detecting wolves, but when researchers probed deeper, they discovered that their AI had instead learned to identify pictures of snow, which just so happened to be present in every picture of a wolf (see Appendix 1). Had the algorithm been required to present its explanation, any expert would have been able to discard the results as identifying snow. When translating this example to the medical ecosystem, some experts have questioned whether physicians and patients will ever fully be able to trust or implement AI decision-making tools in the clinic if the output is not explainable for fear of misdiagnosis or improper treatment plans.

METHODS

English language articles were selected from searches of PubMed and Google Scholar using the search terms "AI explainability," "black box algorithm," and "white box algorithm." Additional articles were identified by manual review of the reference lists of pertinent publications. Web sites managed by government agencies and applicable organizations were also reviewed for relevant information.

Action of the AMA House of Delegates 2025 Annual Meeting: CSAPH Report 8
Recommendations Adopted as Amended, and Remainder of Report Filed.

BACKGROUND

Prior to the 2000s, most augmented/artificial intelligence (AI) applications were a series of "if, then" statements in which an end user could conceivably recreate the program's logic, albeit in a more time-intensive manner.[2] In 2000, researchers began to develop what was known as "machine learning" (ML), which was a much more complex network of equations that began to obscure the logic used to arrive at conclusions, even to the designers of ML programs.[3] While orders of magnitude more powerful than previous iterations of AI tools, the inability for many ML tools to describe their decision-making process has resulted in significant discussion as to the value of XAI.

At its core, the implementation and use of XAI places physician autonomy in question. In instances where algorithms can explain their decision-making, they may be useful tools for assessing huge data sets and recognizing patterns that are imperceptible to a clinician. In those cases, those algorithms are augmenting the physician's skillset by consolidating large volumes of data, with outputs which can support a physician's decision-making or be discarded if the physician's training and expertise disagree with the conclusion. However, if clinical AI/ML algorithms are not explainable, the clinician's training and expertise is removed from decision-making, and they are presented with information they may feel compelled to act upon without knowing where it came from or being able to assess quality and accuracy of the conclusion.

To help readers conceptualize the concepts of XAI, this report discusses several simplified, hypothetical scenarios. However, to best convey the utility of XAI, these hypotheticals will often describe instances where AI performed poorly or resulted in negative patient outcomes. This focus, however, should not be construed as a blanket criticism of AI in the clinic, but rather highlight opportunities where the physician voice can be used to push for development of safe, responsible, and impactful products for patients.

TRUSTWORTHINESS OF OUTPUTS

Briefly, trustworthiness describes what a person (or in this case, an AI tool) has done to demonstrate that they can be trusted. Historically, AI developers have used accuracy or other performance measures as a proxy for trustworthiness, but experts have argued that this alone is insufficient.[4] Physicians and their patients may have many concerns about AI in medicine beyond accuracy, including privacy or fairness. One of the steps proposed by ethics researchers and regulatory bodies to demonstrate AI trustworthiness is through explainability.[5]

Take for example, a hypothetical AI/ML algorithm that provides recommendations on whether to prescribe antibiotics for patients with a bacterial infection. If that algorithm were not explainable, also known as a "black box" algorithm, then an output may be a simple 'do not treat' versus 'treat'. In this situation, the physician may feel pressured to defer to the AI/ML algorithm, as a specialized tool designed to determine the appropriateness of an antibiotic and accept the recommendation. If the patient were to ask why they were not getting treatment, the physician would have a challenging time communicating the reasoning due to the lack of explainability of the AI/ML algorithm, undercutting the physician's role as a trusted expert.

By contrast, if that algorithm were explainable, also known as a "white box" or "glass box" algorithm, then the same hypothetical output would provide more context and may read "When compared to 5000 patients of similar age, sex, weight, social history, body temperature, and presence of headache, the patient is 80 percent likely to have a mild viral infection and a five

percent chance of having a severe bacterial infection. Given the risk for adverse side effects and antibiotic resistance, it is not recommended to initiate antibiotics." This explanation provides the physician with information about the size of population being compared, the demographics being compared, the inputs it considered, and the risks it balanced – but potentially contains several mistakes, inaccuracies, or extraneous information or misses important context to the patient case. When presented with this explanation, a physician could recognize that a culture is most useful to differentiate between a viral and bacterial infection, or that locally there has been a significant prevalence of viral infections increasing the likelihood of a viral over bacterial infection where antibiotics may be warranted. In this instance, a physician could give this AI recommendation lower value in their differential diagnosis, but could still glean some useful insights, such as recognizing that in a similar patient population, their patient is generally perceived as low risk for severe infection.

Generally, humans have low trust of AI, particularly when it is involved in decision-making. In one study, even when participants were explicitly told that an automated decision-making tool was performing better than them, 81 percent of participants still chose to ignore the tool's inputs.[6] The researchers posit that this response may come from an innate feeling that humans use "perfected automation," meaning humans perform the same calculations as the AI/ML tool but without error. Thus, the majority of AI tool users will only ever incorporate an AI tool's output into their own decision making, rather than solely rely on it.[7] As such, explainability for AI/ML systems may assist to overcome this hesitance for use. Even when the stakes are far lower than someone's health, studies have found that people are more likely to accept an algorithm's movie recommendation if it comes with an explanation as to why (such as, it is similar to other movies you have watched), and even more when the reason for the recommendation is easily understandableand conveys high value information (it has your most frequently viewed actor in it).[8]

In the hypothetical example of an algorithm recommending against prescribing antibiotics, if the patient were to develop a severe bacterial infection, then the importance of explainability for trustworthiness is further amplified. If the algorithm were a black box algorithm and simply got the recommendation wrong, trust is fractured. Studies have found the human response to AI/ML algorithms is that a single failure from a black box algorithm causes a person to severely underestimate the algorithm's level of accuracy in the future, often fully disregarding recommendations in perpetuity.[6] However, with a white box algorithm, even though the logic may be flawed, the user can see the decision-making process, understand what its limitations are, and may still feel comfortable using it in the future, albeit with more caution.[9] This trend also matches the perceptions of patients; in studies where AI is a black box, or otherwise removes their physician's experience and training from their decision-making, patient trust decreases and concerns about liability increase.[10-12]

It is unclear how the use and recommendations for AI will be disclosed to patients or documented. Current AMA policy calls for "a risk- and impact-based approach that considers the unique circumstance of AI and its use case. The need for transparency and disclosure is greater where the performance of an AI-enabled technology has a greater risk of causing harm to a patient." In instances where AI tools are aiding in diagnosis but do not provide explanations (or the explanation is inadequate for non-experts), disclosure of an AI recommendation into a patient's medical record could have serious implications such as insurance reimbursement or on the patient-physician relationship, particularly in instances where the physician does not agree with the AI recommendation that their patient can view.

This tension further escalates as the conditions of interest become more serious. For example, in the mid-2010s, the electronic health record vendor Epic released a module for the detection of

sepsis which utilized a proprietary, black box model. When external researchers probed the performance of the model, they found that while it did detect seven percent of true sepsis cases that clinicians missed, the software had a 67 percent false negative rate across all sepsis cases.[13] Further, the model had a high false-positive rate, generating electronic alerts on 18 percent of *all* hospitalized patients (compared to the true positive sepsis rate of seven percent), resulting in significant "alert fatigue" in clinicians. After working with the model for prolonged periods of time, it is possible that clinicians would begin to ignore the alerts and instead rely on traditional clinical signs for sepsis diagnosis.

Technical Feasibility

While XAI is highly appealing for building trust in this new technology, it is less clear that it is implementable for medical applications. A common fallacy when evaluating AI/ML algorithms is to assume that it approaches problems and datasets in the same way a human would. While the field of human cognition is vast and complicated, two simple techniques that humans use to assess information and make decisions are linear logic and inference.[14] Briefly, humans often rely on linear logic (if A is true, then B is true) to understand information, then use intuition and experience to fill in (or infer) where there are information gaps. For example, if a 50-year old patient with a family history of colon cancer were to present with occult blood in their stool, a physician is likely to infer from a constellation of risk factors, signs and symptoms to assess that the patient is at high risk of having colon cancer. They would then use a confirmatory test, such as a colonoscopy with biopsy, to create a linear logic chain to confirm their diagnosis.

AI/ML algorithms do not process information in the same way humans do.[15] Thinking is a purely biologic process and cannot currently be replicated by any artificial technique. As such, AI/ML cannot infer. In a typical ML model, every single input is assigned a specific weight, and the output is a sum or other computation of those weights. Weights are based on datasets used to "train" the algorithm and are often dynamic and/or non-linear. These systems are generally referred to as "neural networks" to invoke the imagery of the hundred billion neurons and their interconnectivity similar to the human brain.[16] To help visualize the complexity of these connections, a neural network diagram for a simple AI/ML algorithm used to differentiate between 10 unique digits (0 through 9) from handwriting samples is included as Appendix 2.[17] Due to the complexity of the computations, even the developers of AI/ML algorithms may not fully comprehend how their programs arrived at an output when the process is too complex, and would thus argue that explainability is an unrealistic expectation humans have of AI/ML.[18]

Building Bridges to Explainability

While there may never be a way to truly convert the computation of an AI/ML algorithm into a comprehensible form for human interpretation, there are promising techniques being developed that do provide the user with *more* information to build confidence in AI, albeit incomplete. Some of these techniques include Locally-Interpretable Model-agnostic Explanation (LIME),[19] Gradient-weighted Class Activation Mapping (Grad-CAM),[20] and Occlusion Sensitivity (OS).[21]

Consider a hypothetical AI/ML algorithm that has been designed to detect lung tumors. In this example, the physician inputs a chest MRI for their patient, and the program returns a score for the likelihood a malignant tumor has been detected. In instances where the score is high, the obvious follow-up is: how can you tell? What factors is the algorithm using to distinguish between malignant and benign? In this hypothetical, the algorithm is only trained to calculate a "likelihood of tumor" value from millions of interconnected calculations and report "malignant" if the value is above a pre-determined value, or "benign" if the value is below.

To bridge the gap between the complex AI/ML algorithm and an explainable, interpretable result, a technique such as OS can be utilized. OS analysis highlights the area of an image that is being utilized in a computation for human assessment of accuracy.[21] Using the OS technique, the same MRI is entered into the algorithm hundreds more times, but each time with a slight modification – typically by systematically removing a subunit of the image, like a single pixel. Then, the "likelihood of tumor" output is compared to the original image; if the modified image has a lower likelihood value than the original image, you can infer that the missing pixel was strongly associated with the tumor. This process is repeated multiple times, with each run removing a different pixel and assessing its impact on the tumor likelihood value. Then, a heat map can be generated and overlaid on the original MRI to highlight what regions of the image were most strongly associated with the tumor by the AI. The physician can then interpret the findings, use their experience, and evaluate whether the highlighted area is in fact likely a tumor. Appendix 3 visualizes a hypothetical workflow for how OS explanations in a program identifying dog breeds from an image can improve useability, accuracy, and trust.

Another model, LIME, uses a similar approach, in which it slightly changes the inputs and assesses how the output changes. However, using LIME, the model then attempts calculate a linear regression between how each new input and the matching output.[19] This allows LIME to then generate a series of numerical weights to convey how important each changed input may be. These weights could be used to generate a similar heat map to OS if the input was an image, but it also allows the user greater flexibility in the inputs it considers, including text. For example, if LIME were used in the above MRI example, it may be possible for it to additionally tell the physician how important factors from the patient's medical record (age, weight, sex, etc.) were to its interpretation of the MRI. While approaches like LIME may be more flexible, they do rely on simplifying the complex calculations of an AI/ML algorithm to simple linear equations, which may make it more prone to failure the further an individual case gets from the typical cases used in the algorithm's training set.

OS and other explainability methods have been utilized for applications such as identifying prostate cancer in histopathological samples, distinguishing lung diseases from a radiograph, predicting risk for psoriatic arthritis from the electronic health record, and more.[22-24] Beyond heat maps or other visualization tools, some tools can also provide simple written outputs (such as "clean margin") or link to data from its training set that the tool found to be most similar to the input. This provides an opportunity for the AI/ML user to utilize their own expertise in deciphering the accuracy of the output. While true explainability may never be possible for AI algorithms, models like OS, LIME, and Grad-CAM are promising efforts to make these systems more trustworthy. However, some experts in the field question: is explainability the right bar to hold AI to?[25]

Other Black Boxes in Medicine

When looking at other aspects of medicine, it is not uncommon to find black boxes or unclear processes that physicians and patients trust. For example, the analgesic mechanism for acetaminophen is not fully known and debated frequently in the literature, yet it is widely available over-the-counter in the United States.[26] Similarly, many genetic tests rely on genome wide association studies (GWAS), which often do not have a known, underlying biologic mechanism, but correlate certain genetic mutations with an increased risk of disease.[27] Yet these, and many other aspects of medicine, are routinely utilized in practice despite not truly being explainable, because they have been found to be safe and effective using rigorous scientific testing.

Randomized clinical trials (RCTs) are recognized as the gold standard or high level for evidence development in medicine, whether the intervention is explainable or not.[28] By carefully controlling variables and often utilizing a placebo, RCTs allow researchers, physicians, and regulators to best assess safety and efficacy of an intervention – but notably they do not necessitate a known mechanism of action, but often a theoretical hypothesis. When evaluating a new drug, the U.S. Food and Drug Administration (FDA) prefers, but ultimately does not require, a drugmaker to know how their drug works; they simply must prove that it is safe and effective for a specific disease in its target population before approval.[29] True explainability may never be achievable for AI, but there is no reason to believe that an individual AI tool could not be found to be safe and effective using an appropriately designed RCT model. This raises the question as to why humans intrinsically view explainability for AI to be more important than in other black boxes or unexplainable processes in medicine, and if that will always remain a barrier to trust, but that philosophical debate reaches beyond the scope of this report.

Using explainability as a strict requirement for clinical adoption could additionally exclude applications of AI which rely on noticing patterns for which current medical knowledge cannot explain. The hypothetical examples described thus far in this report (antibiotic prescribing recommendations and image interpretation) describe devices that aim to improve or build upon current best practices in medicine. However, there are many researchers actively seeking to discover *new* methods or treatments from the vast amounts of medical data available. For example, one study used a ML algorithm to diagnose patients with type-2 diabetes mellitus (T2DM) based on recordings of their speech – a completely novel approach.[30] The authors of the study hypothesized several potential causes, such as the influence of blood glucose levels on vocal cord elasticity, or pitch modulation caused by myopathy, but as of writing, this correlation would be considered unexplainable using current medical knowledge. Despite that limitation, the authors reported over 70 percent accuracy in detecting T2DM just from audio recordings of a mere 11 words. Given the low level of invasiveness, low cost and prevalence of smart phones, similar tools could be desirable for routine screening applications despite the inability for patients and their physicians to comprehend how T2DM changes the voice. In those instances, transparency around the relative risks and benefits could be more useful for developing trust, which could be derived from a RCT.

CURRENT REGULATORY APPROACHES

In Medicine

As described in BOT 01-I-24, "Augmented Intelligence Development, Deployment, and Use in Health Care," the regulatory landscape for AI in the United States is inconsistent, and relevant health care regulatory agencies do not currently have a comprehensive strategy for oversight of AI. The FDA has been reviewing and approving algorithm-based devices since 1995, with over 1000 devices that utilize AI/ML being approved as of January 2025.[31] Applications for these devices vary, including triage and diagnostics, and cross multiple specialties.

In June 2024, the FDA, in collaboration with Health Canada and the United Kingdom's Medicines and Healthcare Products Regulatory Agency (MHRA), released their 10 guidelines for "good machine learning practice," which have been listed in Appendix 4.[32] These guidelines are non-binding, but give insight as to how regulators are thinking about AI oversight. None of the guidelines specifically mention explainability, however recommendations 7 ("Focus Is Placed on the Performance of the Human-AI Team"), and 9 ("Users Are Provided Clear, Essential

Information") are generally supportive of the concept. Of note, under recommendation 9, device manufacturers are suggested to provide "the basis for decision-making when available." Interestingly, this phrasing could have two interpretations: (1) decision-making is not always present in AI tools, but an explanation is required whenever it is; or (2) explanations of decision-making are preferred, but ultimately not required if they are too complex to derive or no external explainability model is available. While ultimately the specific interpretation of this recommendation is moot, as they are non-binding, it does underscore the level of uncertainty in potential regulations for explainability moving forward.

In January 2025, the FDA released a draft Guidance for Industry that has yet to be formalized at the time of writing.[33] In it, the FDA further describes the types of data they wish to see in submissions from AI tool developers. Within this guidance, the FDA describes explainability as a "risk control" to mitigate potential harm. Additionally, they further expound on explainability and visualization tools such as those described in this report, stating "[…] explainability tools or visualizations can be valuable in increasing model transparency and a user's confidence in a model's output and could be developed as part of the user interface. However, if not well designed and validated for the target user group, explainability tools or visualizations could also significantly mislead users. Therefore, sponsors should develop and validate explainability metrics and visualizations through appropriate testing." However, the guidance does not ultimate require explainability for a device submission.

In a scoping review of FDA-approved AI devices from 1995 to 2023 (692 total devices), researchers found that only 46 percent of device sponsors provided the FDA with the results of performance studies, and only 37 percent provided information on their testing sample size.[34] Given these gaps in disclosed information, the researchers concluded that "[their] current findings suggest that evaluation [of explainability] cannot be comprehensively conducted across approved FDA devices." A similar study, investigating 104 FDA-cleared AI tools to aid in medical imaging interpretation, found that less than half provided an explanation of their output.[35]

At the 2024 Interim Meeting of the House of Delegates, the AMA adopted policy stating that regulation should be "a risk-based approach where the level of scrutiny, validation, and oversight should be proportionate to the overall potential of disparate harm and consequences." While a lack of explainability does not increase the risk of a tool per se, removing a physician's ability to contextualize information from the output using their expertise should be considered a risk factor.

In 2024, the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology (ASTP/ONC) enacted policies to advance AI transparency through electronic health record (EHR) regulation, requiring disclosure of source attributes, data elements, and decision-making roles in AI technology embedded in EHRs. While not specifically focusing on explainability, ASTP/ONC's intent is to empower physicians to make informed choices, ensuring AI tools enhance rather than override clinical judgment. Eplainability is not explicitly mandated, yet ONC's emphasis on transparency will likely foster trust by clarifying how predictive models operate and assist physicians in interpreting AI outputs. These policies are intended to promote responsible AI adoption, reinforcing physician autonomy and incentivizing the development of fair, effective, and safe AI-driven tools in healthcare.

Some manufacturers and scholars have raised concerns that by disclosing or otherwise visualizing an explanation of AI computational processes, they may be exposing their intellectual property (IP) to competitors.[36,37] There remains ambiguity as to what is patentable with regards to AI-enabled medical devices. For example, the software driving the algorithm is generally patentable, as it is considered a finished product. However, the Supreme Court found in *Gottschalk v. Benson* that

mathematical formulas are generally not patentable as they represent abstract concepts, which leave algorithms unprotected.[38] In a January 2025 report on AI and copyrightability, the U.S. Copyright Office concluded that current laws and regulations adequately address AI copyright concerns, and did not recommend any legislative changes.[39] The recent release of the Chinese-based AI DeepSeek has highlighted the difficulty in protecting IP in the rapidly growing AI/ML space. Briefly, the American-based company OpenAI claim that DeepSeek developers used the outputs of OpenAI models to reverse engineer or otherwise train a model that would be a market competitor.[40] While explanations of AI/ML tool outputs make the outputs more trusted, they may also make the underlying system more vulnerable to rival companies.

As such, other fields which have utilized algorithms (like banking and finance) rely on a "trade secrets" model for protecting their IP, in which algorithms are deemed proprietary and are hidden from the user. This is a similar approach to how food manufacturers protect their recipe yet disclose their ingredients for food products. However, if this approach were to continue, medical AI developers may be pitting innovation against a patient's right to transparency and autonomy in their medical decision-making.

Outside of Medicine

Given the rapid expansion of AI, other fields grapple with similar issues of explainability in their regulatory oversight. For example, the Equal Credit Opportunity Act of 1974 requires that financial institutions provide written descriptions explaining why they made an adverse decision (such as denying a loan application), and explicitly protecting certain traits (such as race or sex) from being the basis for those decisions. However, as financial institutions began incorporating more and more automated AI tools in their decision-making, explanations for adverse decisions became increasingly more abstract, and many worried that protected traits were being used by black box algorithms.

In response, the Consumer Financial Protection Bureau (CFPB) released a memo in 2023 clarifying that even in instances where black box AI tools "[made] it difficult — if not impossible — to accurately identify the specific reasons for denying credit or taking other adverse actions," customers are still legally owed an explanation of those specific reasons they were denied, and thus does not permit unexplainable algorithms to be used.[41] The CFPB went further, stating that even estimations or proxies of the AI tool's logic may not be acceptable if they are not specific enough. For example, if a financial institution did not know the logic their AI tool used to make decisions, a simple explanation of "the applicant has insufficient income" would be deemed inadequate. This approach mirrors legislation in states such as Colorado, New York, California, and Connecticut, which limit insurance companies' ability to use unexplainable, black box algorithms when making insurance coverage determinations.[42]

Outside the United States, the European Union's 2018 General Data Protection Regulation (GDPR) is generally regarded as one of the first attempts at comprehensive regulations of AI and other digital technologies and is the basis for many international regulations. The GDPR contains several regulations for the development and use of algorithms, but its position on explainability is less clear. Under Article 15 of the GDPR, it states that algorithms are required to disclose "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[43] Some scholars have interpreted this text to mean that the GDPR establishes a "right to explanation," however this right has yet to be asserted and adjudicated in a European court.[44]

Additional Considerations

When regulating the explainability of AI, it is critical to establish both who is owed an explanation, and where the explanation comes from. In the medical context, there are several potential audiences, such as the physician, the patient, or external groups such as payors. If, for example, a "right to explanation" was proposed – who has the right? In the United States, the Health Insurance Portability and Accountability Act (HIPAA) generally establishes that patients have a right to access their clinical data.[45] However, as discussed above, there are significant gaps in the ability of most AI tools to explain their outputs in a layperson fashion, and most models to approximate explanations (such as OS, LIME, and Grad-CAM) are targeted to a physician-type expert for contextualization and action.

When discussing the disclosure of test results, the Code of Medical Ethics states that "[test] results [should be] conveyed sensitively, in a way that is understandable to the patient/surrogate, and the patient/surrogate receives information needed to make well-considered decisions about medical treatment and give informed consent to future treatment[.]" In a hypothetical situation where a physician receives an AI tool's explanation, but then uses their own words to convey that information to their patient, it is unclear if that would suffice under some scholarly interpretations of a GDPR-styled "right to explanation."

CURRENT AMA POLICY

The AMA maintains extensive policy on AI generally. Board of Trustees (BOT) Report 15-I-24, "Augmented Intelligence Development, Deployment, and Use in Health Care," provided a comprehensive overview of the regulatory landscape and the AMA's history in AI governance. A brief summary of relevant sections of AMA policies are as follows (full text of policy available at the end of this report):

H-480.931, "Assessing the Intersection Between AI and Health Care"

- "Health care AI must be designed, developed, and deployed in a manner which is ethical, equitable, responsible, accurate, and transparent."
- "Health care AI requires a risk-based approach where the level of scrutiny, validation, and oversight should be proportionate to the overall potential of disparate harm and consequences the AI system might introduce."
- "Clinical decisions influenced by AI must be made with specified human intervention points during the decision-making process. As the potential for patient harm increases, the point in time when a physician should utilize their clinical judgment to interpret or act on an AI recommendation should occur earlier in the care plan. With few exceptions, there generally should be a human in the loop when it comes to medical decision making capable of intervening or overriding the output of an AI model."
- "Medical specialty societies, clinical experts, and informaticists are best positioned and should identify the most appropriate uses of AI-enabled technologies relevant to their clinical expertise and set the standards for AI use in their specific domain."
- "Purchasers and/or users (physicians) should carefully consider whether or not to engage with AI-enabled health care technologies if […] information is not disclosed by the developer. As the risk of AI being incorrect increases risks to patients (such as with clinical applications of AI that impact medical decision making), disclosure of […] information becomes increasingly important..

- "Individuals impacted by a payor's automated decision-making system, including patients and their physicians, must have access to all relevant information (including the coverage criteria, results that led to the coverage determination, and clinical guidelines used)."

H-480.939, "Augmented Intelligence in Health Care"

- "Oversight and regulation of health care AI systems must be based on risk of harm and benefit accounting for a host of factors, including but not limited to: intended and reasonably expected use(s); evidence of safety, efficacy, and equity including addressing bias; AI system methods; level of automation; transparency; and, conditions of deployment."
- "Physicians should not be penalized if they do not use AI systems while regulatory oversight, standards, clinical validation, clinical usefulness, and standards of care are in flux."
- "[Our AMA will advocate that] AI is designed to enhance human intelligence and the patient-physician relationship rather than replace it."

H-480.940, "Augmented Intelligence in Health Care"

- "[Our AMA will seek to promote] development of thoughtfully designed, high-quality, clinically validated health care AI that […] is transparent[.]"

CONCLUSION

Ironically, the concept of explainability is hard to explain. It is complex, nuanced, and asks profound questions on human cognition and the meaning of trust. While this report primarily focused on hypothetical instances where AI/ML performs poorly to highlight the opportunities and challenges for explainability, the responsible usage of well-designed AI/ML tools has already had a profoundly transformative impact on medicine, building efficiency in practice, increasing the breath of data integration, and increasing communication capabilities, such as the use of a LLM. The appeal and need for some level of explainability for AI/ML tools in medicine is clear, where decisions can have life or death consequences. Physicians are trained to utilize a broad compilation of information for medical decision-making, including context and seeing the whole patient's clinical picture; a binary "yes/no" output from an AI/ML tool is restrictive and lacks nuance.

However, explainability struggles in practice in part due to the disconnect between how humans and computers process information. AI does not think, nor can it guess, infer, or intuit, all of which are core processes for how a physician would make a clinical determination. Some models, such as occlusion sensitivity, are being developed to allow for insight into the inner workings of an AI tool, but they generally still require expert interpretation, and risk being an oversimplification of the true computational process.

Medicine is experienced in handling unexplainable phenomena and utilizing data through research and evaluation verify safety and efficacy. Understanding the mechanism of action of a drug, biologic process, or otherwise is crucial for building trust, troubleshooting, and advancing medical practice, but it generally has not been the barrier to clinical *entry*. By the same token, however, physicians should feel confident that the tools they use in the clinic are safe, based on sound science, and can be discussed appropriately with their patients, so they can engage in shared decision-making.

RECOMMENDATIONS

The Council on Science and Public Health recommends that the following be adopted and that the remainder of the report be filed:

1. To maximize the impact and trustworthiness of augmented intelligence and machine-learning (AI/ML) tools in clinical settings, our AMA recognizes that:
    a. Explainable AI with safety and efficacy data should be the expected form of AI tools for clinical applications, and exceptions should be rare and justified and require at minimum safety and efficacy data prior to their adoption or regulatory approval.
    b. To be considered "explainable," an AI device's explanation of how it arrived at its output must be interpretable and actionable by a qualified human ~~trained expert~~. Claims that an algorithm is explainable should be adjudicated only by independent third parties, such as regulatory agencies or appropriate specialty societies, rather than by declaration from its developer.
    c. Explainability should not be used as a substitute for other means of establishing safety and efficacy of AI tools, such as through randomized clinical trials.
    d. Concerns of intellectual property (IP) infringement, when provided as rationale for not explaining how an AI device created its output, does not nullify a patient's right to transparency and autonomy in medical decision-making. While intellectual property should be afforded a certain level of protection, concerns of infringement should not outweigh the need for explainability for AI with medical applications. (New HOD Policy)

2. That our American Medical Association will collaborate with experts and interested parties to develop and disseminate a list of definitions for key concepts related to medical AI and its oversight. (Directive to Take Action)

3. That policies H-480.931, "Assessing the Intersection Between AI and Health Care," H-480.939, "Augmented Intelligence in Health Care," and H-480.940, "Augmented Intelligence in Health Care" be reaffirmed. (Reaffirm HOD Policy)

Fiscal Note – less than $1000

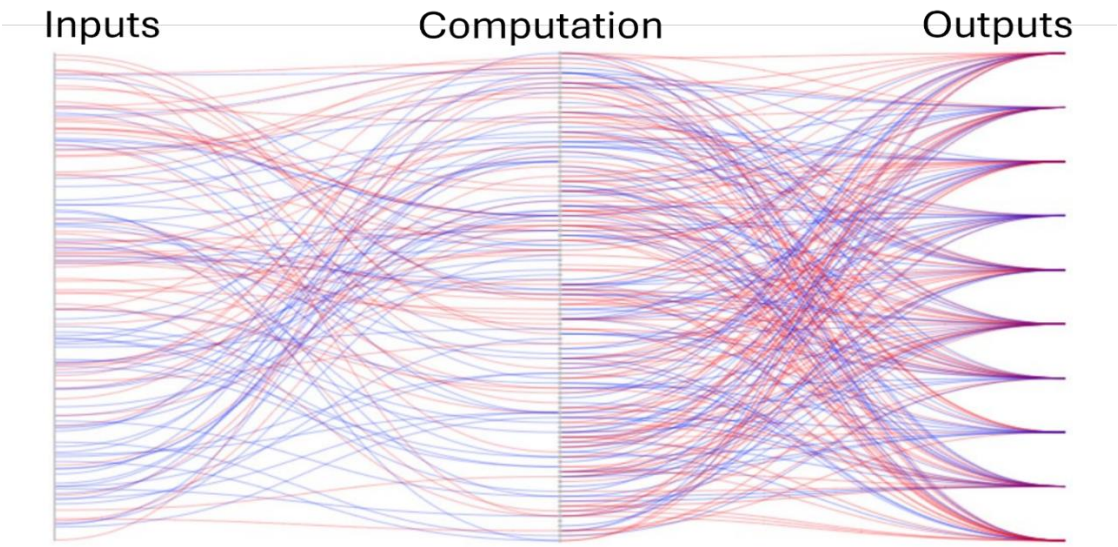
APPENDIX

Appendix 1 – Sample AI Explainability

(a) Husky classified as wolf (b) Explanation

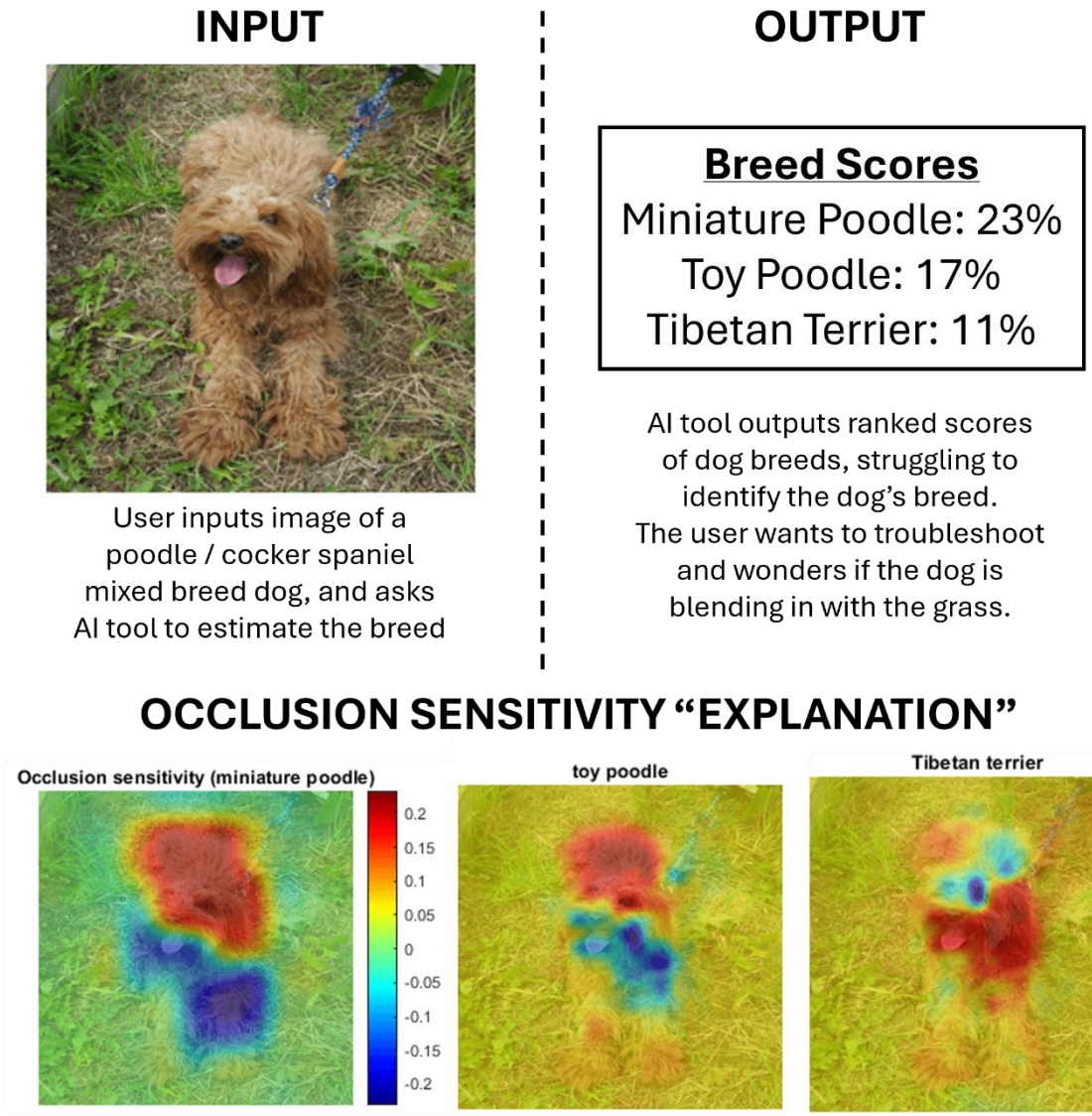
Appendix 1 – A sample of an AI system processing an image (left) and providing an explanation (right) for how it determined whether the input was a picture of a husky or a wolf. In this case, the system mistook a husky for a wolf due to the presence of snow, which was present in most images of wolves used for the AI's training set.

Appendix 2 – Visualization of Neural Networks



Appendix 2 – Visualization of the complexity in a sample neural network used to detect the digits 0 through 9 from handwriting samples. Each line on the left side represents an input considered by the algorithm, and each position on the right side represents a potential output (the digits 0 through 9). The middle visualizes the interconnectivity and how the algorithm sorts inputs into outputs. Image adapted from https://www.i-am.ai/neural-numbers.html.

Appendix 3 – Sample Occlusive Sensitivity Workflow



Apendix 3 – Hypothetical workflow of a user using an AI/ML tool to identify a dog breed. By using occlusion sensitivity, the user can identify that the tool was mischaracterizing key body features, thus explaining the poor results but allowing the user to modify their input to improve accuracy. Images adapted from https://www.mathworks.com/help/deeplearning/ug/understand-network-predictions-using-occlusion.html.

Appendix 4 – FDA Good Machine Learning Practice for Medical Device Development: Guiding Principles

Taken from: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

1. **Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle:** In-depth understanding of a model's intended integration into clinical workflow, and the desired benefits and associated patient risks, can help ensure that ML-enabled medical devices are safe and effective and address clinically meaningful needs over the lifecycle of the device.
2. **Good Software Engineering and Security Practices Are Implemented:** Model design is implemented with attention to the "fundamentals": good software engineering practices, data quality assurance, data management, and robust cybersecurity practices. These practices include methodical risk management and design process that can appropriately capture and communicate design, implementation, and risk management decisions and rationale, as well as ensure data authenticity and integrity.
3. **Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population:** Data collection protocols should ensure that the relevant characteristics of the intended patient population (for example, in terms of age, gender, sex, race, and ethnicity), use, and measurement inputs are sufficiently represented in a sample of adequate size in the clinical study and training and test datasets, so that results can be reasonably generalized to the population of interest. This is important to manage any bias, promote appropriate and generalizable performance across the intended patient population, assess usability, and identify circumstances where the model may underperform.
4. **Training Data Sets Are Independent of Test Sets:** Training and test datasets are selected and maintained to be appropriately independent of one another. All potential sources of dependence, including patient, data acquisition, and site factors, are considered and addressed to assure independence.
5. **Selected Reference Datasets Are Based Upon Best Available Methods:** Accepted, best available methods for developing a reference dataset (that is, a reference standard) ensure that clinically relevant and well characterized data are collected and the limitations of the reference are understood. If available, accepted reference datasets in model development and testing that promote and demonstrate model robustness and generalizability across the intended patient population are used.
6. **Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device:** Model design is suited to the available data and supports the active mitigation of known risks, like overfitting, performance degradation, and security risks. The clinical benefits and risks related to the product are well understood, used to derive clinically meaningful performance goals for testing, and support that the product can safely and effectively achieve its intended use. Considerations include the impact of both global and local performance and uncertainty/variability in the device inputs, outputs, intended patient populations, and clinical use conditions.
7. **Focus Is Placed on the Performance of the Human-AI Team:** Where the model has a "human in the loop," human factors considerations and the human interpretability of the model outputs are addressed with emphasis on the performance of the Human-AI team, rather than just the performance of the model in isolation.
8. **Testing Demonstrates Device Performance during Clinically Relevant Conditions:** Statistically sound test plans are developed and executed to generate clinically relevant device performance information independently of the training data set.

9. Considerations include the intended patient population, important subgroups, clinical environment and use by the Human-AI team, measurement inputs, and potential confounding factors.
10. **Users Are Provided Clear, Essential Information:** Users are provided ready access to clear, contextually relevant information that is appropriate for the intended audience (such as health care providers or patients) including: the product's intended use and indications for use, performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model. Users are also made aware of device modifications and updates from real-world performance monitoring, the basis for decision-making when available, and a means to communicate product concerns to the developer.
11. **Deployed Models Are Monitored for Performance and Re-training Risks are Managed:** Deployed models have the capability to be monitored in "real world" use with a focus on maintained or improved safety and performance. Additionally, when models are periodically or continually trained after deployment, there are appropriate controls in place to manage risks of overfitting, unintended bias, or degradation of the model (for example, dataset drift) that may impact the safety and performance of the model as it is used by the Human-AI team.

REFERENCES

1. Husky or Wolf? Using a Black Box Learning Model to Avoid Adoption Errors. https://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learning-model-to-avoid-adoption-errors/. Published 2017. Accessed.
2. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy.* 2020;92(4):807-812.
3. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface.* 2018;15(141):20170387.
4. Li B, Qi P, Liu B, et al. Trustworthy AI: From Principles to Practices. *ACM Comput Surv.* 2023;55(9):Article 177.
5. Starke G, Ienca M. Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics.* 2024;33(3):360-369.
6. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. *International Journal of Human-Computer Studies.* 2003;58(6):697-718.
7. Dzindolet MT, Pierce LG, Beck HP, Dawe LA. The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors.* 2002;44(1):79-94.
8. Herlocker JL, Konstan JA, Riedl J. Explaining collaborative filtering recommendations. Proceedings of the 2000 ACM conference on Computer supported cooperative work; 2000; Philadelphia, Pennsylvania, USA.
9. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI.* 2024;3:e53207.
10. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' Perceptions Toward Human–Artificial Intelligence Interaction in Health Care: Experimental Study. *J Med Internet Res.* 2021;23(11):e25856.
11. Zhang Z, Citardi D, Wang D, Genc Y, Shan J, Fan X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. *Health Informatics Journal.* 2021;27(2):14604582211011215.
12. Ploug T, Sundby A, Moeslund TB, Holm S. Population Preferences for Performance and Explainability of Artificial Intelligence in Health Care: Choice-Based Conjoint Survey. *J Med Internet Res.* 2021;23(12):e26611.
13. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine.* 2021;181(8):1065-1070.
14. Reyna VF, Brainerd CJ. Dual Processes in Decision Making and Developmental Neuroscience: A Fuzzy-Trace Model. *Dev Rev.* 2011;31(2-3):180-206.
15. Emmert-Streib F, Yli-Harja O, Dehmer M. Artificial intelligence: A clarification of misconceptions, myths and desired status. *Frontiers in artificial intelligence.* 2020;3:524339.
16. Fjelland R. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications.* 2020;7(1):10.
17. I Am AI. Neural Numbers. https://www.i-am.ai/neural-numbers.html. Accessed November 20, 2024.
18. Yampolskiy RV. Unexplainability and Incomprehensibility of AI. *Journal of Artificial Intelligence and Consciousness.* 2020;07(02):277-291.

19. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision.* 2020;128(2):336-359.
21. Zeiler M. Visualizing and Understanding Convolutional Networks. Paper presented at: European Conference on Computer Vision/arXiv2014.
22. Gallo M, Krajňanský V, Nenutil R, Holub P, Brázdil T. Shedding light on the black box of a neural network used to detect prostate cancer in whole slide images by occlusion-based explainability. *New Biotechnology.* 2023;78:52-67.
23. Ullah N, Khan JA, Falco ID, Sannino G. Bridging Clinical Gaps: Multi-Dataset Integration for Reliable Multi-Class Lung Disease Classification with DeepCRINet and Occlusion Sensitivity. Paper presented at: 2024 IEEE Symposium on Computers and Communications (ISCC); 26-29 June 2024, 2024.
24. Lee LT-J, Yang H-C, Nguyen PA, Muhtar MS, Li Y-CJ. Machine Learning Approaches for Predicting Psoriatic Arthritis Risk Using Electronic Medical Records: Population-Based Study. *J Med Internet Res.* 2023;25:e39972.
25. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health.* 2021;3(11):e745-e750.
26. Ohashi N, Kohno T. Analgesic Effect of Acetaminophen: A Review of Known and Novel Mechanisms of Action. *Frontiers in Pharmacology.* 2020;11.
27. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nature Reviews Methods Primers.* 2021;1(1):59.
28. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the Gold Standard — Lessons from the History of RCTs. *New England Journal of Medicine.* 2016;374(22):2175-2181.
29. Mechanism matters. *Nature Medicine.* 2010;16(4):347-347.
30. Kaufman JM, Thommandram A, Fossat Y. Acoustic analysis and prediction of type 2 diabetes mellitus using smartphone-recorded voice segments. *Mayo Clinic Proceedings: Digital Health.* 2023;1(4):534-544.
31. U.S. Food and Drug Administration. FDA Issues Comprehensive Draft Guidance for Developers of Artificial Intelligence-Enabled Medical Devices. https://www.fda.gov/news-events/press-announcements/fda-issues-comprehensive-draft-guidance-developers-artificial-intelligence-enabled-medical-devices. Published 2025. Accessed January 23, 2025.
32. U.S. Food and Drug Administration. Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles. Published 2024. Accessed November 11, 2024.
33. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing. Published 2025. Accessed February 7, 2025.
34. Muralidharan V, Adewale BA, Huang CJ, et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *npj Digital Medicine.* 2024;7(1):273.
35. McNamara SL, Yi PH, Lotter W. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. *npj Digital Medicine.* 2024;7(1):80.

36. Hilty R, Hoffmann J, Scheuerer S. Intellectual property justification for artificial intelligence. 2020.
37. Anjos L. Rethinking Algorithmic Explainability Through the Lenses of Intellectual Property and Competition. 2023.
38. Diaz GC. Contested Ontologies of Software: The Story of Gottschalk v. Benson, 1963-1972. *IEEE Annals of the History of Computing.* 2016;38(1):23-33.
39. United States Copyright Office. Report on Copyright and Artificial Intelligence. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf. Published 2025. Accessed February 7, 2025.
40. Criddle C, Olcot E. OpenAI says it has evidence China's DeepSeek used its model to train competitor. *Financial Times.* January 29, 2025.
41. Consumer Financial Protection Bureau. Adverse action notification requirements in connection with credit decisions based on complex algorithms. https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/. Published 2023. Accessed November 21, 2024.
42. Zhang D. Insurers' AI Use for Coverage Decisions Targeted by Blue States. *Bloomberg Law.* November 30, 2023.
43. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). In:2016:1-88.
44. E. Kaminski M, Sandeen SK, Rademacher C, Ohly A. Research Handbook on Information Law and Governance. In: *Chapter 15: The right to explanation, explained.* Edward Elgar Publishing; 2021.
45. U.S. Department of Health and Human Services. Individuals' Right under HIPAA to Access their Health Information 45 CFR § 164.524. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access/index.html. Published 2024. Updated January 4, 2024. Accessed November 11, 2024.