Cleveland Clinic

Kevin M. Mooney, Esq., Senior Director
Enterprise Data Governance Office
Enterprise Data & Analytics
mooneyk@ccf.org

February 2, 2024

Information Technology Laboratory
National Institute of Standards and Technology
100 Bureau Drive
Mail Stop 8900
Gaithersburg, MD 20899-8900

Re: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11) (NIST–2023–0309)

Submitted electronically via: Regulations.gov

Cleveland Clinic is a not-for-profit, integrated healthcare system dedicated to patient-centered care, teaching, and research. With a footprint in Northeast Ohio, Florida, and Nevada, Cleveland Clinic Health System operates a main campus near downtown Cleveland, 22 hospitals, and 276 outpatient locations. Cleveland Clinic employs over 5,600 physicians and researchers and 19,000 nurses and advanced practice providers. Last year, our system cared for 3.4 million unique patients, including 12.8 million outpatient visits and 303,000 hospital admissions and observations.

Cleveland Clinic appreciates the opportunity to share our thoughts with the National Institute of Standards and Technology (NIST) regarding the agency's responsibilities under the October 30, 2023, executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. (AI) The answers and insights we provide relate specifically to AI in a health care setting.

Developing Guidelines, Standards, and Best Practices for AI Safety and Security: *NIST is seeking comprehensive information regarding the development of guidelines, standards, and best practices for ensuring the safety and security of AI systems. The request includes specific inquiries related to generative AI risk management, AI evaluation, and red-teaming practices. This encompasses the development of a companion resource for the AI Risk Management Framework (AI RMF) focusing on generative AI, exploring risks and harms associated with generative AI, current industry norms, governance practices, measurement approaches, content authentication, impacts assessment tools, transparency requirements, economic and security implications, among other considerations. Additionally, NIST seeks guidance for evaluating and auditing AI capabilities with a focus on potential harmful impacts, proposing metrics, benchmarks, and rigorous measurement approaches for evaluating AI functionality, safety, security, privacy, equity, and trustworthiness.*

The guidelines, standards, and best practices associated with safety and security of AI systems in healthcare are clearly evolving and currently immature. It is challenging to even think that these systems will be fully autonomous in the future given users' varying levels of trust and reliability in them. Cleveland Clinic has taken various steps to promote transparency, ensure high-quality output,

and keep pace with best practices in what has quickly become "traditional AI," like machine learning. The underpinning of trust is transparency: the ability to understand through reasonable and logical deduction a program's explainability. The more transparent the algorithm, the more likely it adheres to safety and quality requirements. NIST defines trustworthiness in the NIST AI RMF to include elements of being "*valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhances, and fair with harmful bias managed.*"[1] While tradeoffs are contemplated in the NIST AI RMF to give it flexibility and agnostic features, a requirement could be added for full documentation on the reasonable deliberation of those tradeoffs.

Further, transparency in particular has three sub-components: model transparency, patient transparency, and developer transparency. Transparency, as the centerpiece of societal trust, must be integral to any regulations, guidelines, procedures, and governance.

1. **Developer/Development Transparency**

Those engaged in AI activity, AI Actors under the NIST AI RMF, need their organization to publish clear standards of engagement, development frameworks, and expected guardrails for creation, testing, peer review, validation, deployment, and ongoing model management. These include:

- Create a governance approval pathway that has milestone gates for each material subject matter in the full lifecycle of AI.
- Expect developers to understand the underlying training data, methods, and assumptions of any generative AI model they may be using from a third party.
- Use fine tuning on a limited basis to protect sensitive data, and instead, utilize augmentation patterns which keeps information highly secure.
- Vigorous testing of the data itself and the developing algorithms to mitigate bias and safety concerns.
- Explicitly state the "dos" and "don'ts" of an AI system and continuously monitor usage.
- Invest time and effort into quantifying the quality of responses, either by expert review (e.g. a physician reviewing a task he or she would have completed) or asking for structured responses (e.g. "choose the best response" from a series of answers which includes the ideal result).
- Introduce a "human in the loop" whenever possible and measure the necessity of the AI system with that in mind. This is vastly important in AI use in operations where human judgment and intervention to independently evaluate the resulting output are mandated, excepted on a limited basis with very low-risk repeatable systems.
- When it is appropriate, clearly state for the end user that the output is a product of generative AI.
- Create a multi-disciplinary governing body that can review use cases for generative AI within the organization, particularly inclusive of legal and ethics expertise.
- Consult patients about new use cases and disclose the use of generative AI within different workflows across the health system.

[1] AI Risk Management Framework FAQs, https://www.nist.gov/itl/ai-risk-management-framework/ai-risk-management-framework-faqs.

Guidelines for developers foster an inter-disciplinary team and create a more inclusive group of practitioners leading to efficiency, availability, and quality of patient interactions. Further, well governed and current outcomes in general benefit future patients and continue to educate professionals, providing them with better and quicker tools. AI may therefore help with the ultimate goal of migrating healthcare away from being in the business of sickness to being in the business of wellness simply by the execution of its potential team effort approach and its constant education of those serving patients.

Additionally, turning to Generative AI specifically, although issues arise in all AI programs, Generative AI programs have certain nuanced legal issues that need to be addressed in any companion frameworks. The collection, storage, and use of data by or for these AI programs raise questions concerning intellectual property, privacy, and data protection, both from (1) the use of massive amounts of publicly available data, such as websites, images, and videos, and (2) from deployed data models containing sensitive information.

- Related to intellectual property, clear compliance with laws and standards (such as copyright law) should be mandated and an assessment of intellectual property questions should guide developers' understanding of such laws, including guidance related to ownership of newly created works and further guidance on the fair use doctrine under the US Copyright Act.
- Related to privacy and data protection, a similar set of guiding questions should be developed to prevent additional vulnerabilities that may be introduced in this type of AI program, including hardened access controls, security protocols, logging and storing techniques as well as a clear defense standardization related to prompt engineering, data poisoning, and other malicious activity.

2. Model Transparency

Model transparency has several core components, but one significant issue that needs to be addressed is the ability of AI to unintentionally disrupt legal theory. Any evolving regulation or framework related to AI must keep in mind its utility in driving societal predictability as it relates to liability under the law. The lynchpin of liability, i.e., predictable accountability under the law, will need to be carefully crafted and vetted. In fault-based actions related to AI, the question is whether we have clear standards related to how fault will be proven? If there is an action taken or omitted in the generation of decisions in AI, and the output harms an individual, is it clear who the acting party is who should be held liable for the action – is it someone in the chain of treatment; the third party software company; the subcontractors who wrote work for hire code for the third party; the provider's data scientist; the provider's leadership; or the AI software itself, free thinking and evolving its intelligence over time? There are several theories of liability that may be disrupted by the continuous evolution of AI.

As we know, some AI tools are referred to on occasion as "Black Boxes" which may be software or an algorithm, the output or byproduct of which lacks transparency related to its decision making and path of reason, and therefore cannot be explained, accounted for, and audited. The results are nebulous and complex algorithms whose decisions are not easy to reverse engineer or to understand. However, someone – a human – created it. AI is, at its heart and demystified, source code, math, process, and data. The more likely one can explain the mechanism and the nature of the AI tool, the more likely one may be able to prove and defend its fairness in practice. Explainability and interpretability have a direct impact upon certain legal theory, including causation and foreseeability.

In short, a deadly combination of AI-centric facts may blend together to impact certain legal theory. With the increased complexity of algorithms and the processing of more expansive and numerous variables, the harder it will be for the human mind to comprehend the nuances of these algorithms. The human brain, quite obviously, may be more limited in its processing power than AI. We cannot process higher dimensionality, nor may we be able to match with understanding how an algorithm is making decisions, connecting data points. or drawing distinctions. If the algorithm is too dimensionally complex for the human capacity of thought, and human logic cannot track where or how the decisions or distinctions are being made, such an algorithm could cause trouble with certain theory.

Intent and Causation theories largely depend upon the behavior retroactively traced back to a subject human being. That behavior may be posited as evidence designed to satisfy certain elements of legal theory. Humans inevitably leave in their wake their fingerprints and footprints that may show where they have been and what they have done, and perhaps, what they intended to do, based on that evidence, whether in tangible damage, emails, notes, or conversations. That human, as a witness, may be deposed or cross examined, linking the behavior to the element attempted to be proven as fact as the puzzle of foreseeability is constructed.

Intent Based Civil and Criminal Actions

Intent theory not only aids juries and courts in understanding and regulating human conduct, but it also helps with societal comfort in the predictable assessment of the severity of penalty related to certain actions. Intent is a human notion, and AI is not a person under the law. AI does not have intent. A court and opposing counsel can only examine the programmer's intent, perhaps by reviewing his memos, emails, and code to try to understand his intent. In a weak Black Box perhaps the AI itself can be reviewed, and, if the code is transparent enough, may show the programmer's intent. Without this transparency, however, the AI will not be helpful in demonstrating this element.

For example, if an organization intended to streamline an operational system to secure a larger pool of clients, it may build an algorithm to accomplish this goal. How the algorithm accomplishes this goal, depending on its sophistication, may be unclear or unknown to the programmer, or ultimately, the court. The algorithm in this hypothetical example, learning, connecting, creating new output, and continuing to learn, finds a connection that ultimately maximizes efficiency of securing clients, yet in practice violates federal law. That strategy to break the law was developed by the algorithm, not the programmer, who only intended to be more efficient.

In short, unless it is shown that the outcome of the action was derived from the programmer's intent, then it will be most difficult to hold the programmer liable for the action of the algorithm. This means that laws that rely on the intent element, in part, to demonstrate culpability, may fail to hold the guilty party in check. Claimants and plaintiffs, therefore, may be left with no legal recourse to address their harm. Further, some of these intent based laws are found in criminal statutes. Some carry steep penalties, some jail time. Yet without proving intent when it is required to be proven, prosecutors will be left screaming at ghosts while the guilty return to their labs to cook up their next generation algorithm. Similar outcomes occur on the civil side with the likelihood of claims dying due to the granting of motions to dismiss and summary judgments.

Negligence

Turning to another concept: under basic negligence theory, a duty is owed to a particular person. That duty is breached by the defendant; that particular person is harmed by the defendant. Lastly, it can be shown through causation that the action led to the damage. One lynchpin to prove negligence, therefore, is determining the standard of care under which the defendant will be evaluated under these actions.

Standard of Care

Simplified, the standard of care for general negligence for the public is based upon a reasonably prudent person. However, the standard of care by which a physician will be judged is a higher level, that of a reasonably prudent physician under similar circumstances. Further, the standard of care may change over time with the advent of new technologies. Conceptually by example, prior to the use of x-rays, that technology was not part of the standard of care imputed upon a physician in the treatment and care of a patient. Then the world started using x-rays. After some time, the use of x-rays became common practice, in effect, a technology upon which physicians relied to help make decisions. Subsequent to the common practice of using x-rays, if a physician failed to use an x-ray in a situation where the industry was using x-rays, then should subsequent harm befall the patient, that physician may face a negligence action in which he may be deemed to have failed to adhere to the appropriate standard of care, namely, a reasonably prudent physician under the circumstances should have used an x-ray.

AI is the same animal as that in the advent of x-rays. Using AI broadly to augment physician decisions is not the standard of care currently. However, as time goes by and more and more algorithms are built, and more and more physicians collectively merge the utilization of certain AI in repeatable ways in care settings, such use may eventually become a standard of care upon which physicians rely. Eventually, with its prevalence proven in the industry, there may come a day when a physician who does not use that AI algorithm in a similar circumstance as the standard that has been created, may find himself needing to defend this failure to adhere to the standard of care, leading hypothetically to a potential finding of negligence.

Proximate Cause

Leaving the standard of care aside for the moment, another fixture in negligence claims that may be impacted related to AI advances is the theory of causation. Proximate cause focuses on whether the result of the subject's conduct was one that could have or should have been foreseen by a reasonably prudent person. It surmises that a person should be held liable for damage that he could have foreseen and should not be liable for damages that were unforeseeable related to his actions. If the harm was foreseeable, and he performed the action anyway, he should be held liable. If the damage is, however, unforeseeable, and he could do nothing to prevent the damage he could not foresee, he therefore should not be held liable. This creates societal predictability related to liability. Since Black Box AI and its decisions may not have been foreseeable (i.e., AI may find shortcuts in complex, dimensional, and obscure patterns, and relationships between variables naked to human understanding), causation tests may fail since they focus on foreseeability which cannot, hypothetically, be proven.

If AI, in a Black Box context, is reading patterns or producing results that are unforeseeable or reaching conclusions that a human being could not have made, then it arguably cannot be held liable for those results (and again, AI is not a person under the law and cannot be held liable anyway). Neither can its programmer, who similarly cannot understand the foreseeable outcomes. Logic would seem to dictate that someone making medical decisions he or she cannot justify using a tool that will

provide unpredictable and unforeseeable results is, counterintuitively, actually being negligent; but if the risk of harm is unforeseeable, and the Black Box nature of AI is unforeseeable, then there may be no recourse for resultant harm, and the actor may be shielded in this cause of action. It is difficult to imagine this logic helping to build a legal framework of predictability to determine when a party should be held liable and when another should not be. Yet this is exactly what is confronting us – without foreseeability, as AI becomes more autonomous and distanced from its programmers, arguably, certain elements in negligence claims may go unproven, thereby weakening society's need for predictable assignments of fault.

On the other hand, these theories may still function in their intended design should the algorithm contain a reasonable basis of explanation and transparency and perhaps leave open the possibility to reverse engineer the decision-making capability. AI will ultimately continue to become more complex, becoming less transparent over time. Legal theory will continue to drag behind technological advances. Rather than a confluence of theory, a further divide will continue to grow between society's desire for predictability through the rigors of the law versus innovations and cutting-edge clinical practices for the betterment of society through technological advances. This creates a deepening lack of transparency and a further erosion of trust. How we remain attuned to this inherent conflict will be critical in how we advance these divergent theories.

By using a NIST standard or building a robust governance process, clear guidance is needed to understand whether an organization will be shielded from liability in certain respects, similar to a safe harbor in which the duty of care requirements are satisfied. (See an example of this at https://codes.ohio.gov/ohio-revised-code/chapter-1354 as codified in Ohio.)

3. Patient Transparency

Where applicable, patient transparency suggests that patients should be generally informed when AI is materially used in their care or treatment. This will include the standard protocol in such situations where the AI is positioned as a decision support tool and the notion that patients understand that a human is intervening on this augmented technology. This human judgement, using AI as decision support as an input by the actual human who is making the medical decision, is critical in ensuring patient understanding, informed decision making, and safety in the clinical use of AI. Ultimately certain touchstones are addressed with such transparency, including whether the machine is intended to be making clinical decisions; what level of human intervention is at play; and whether reliance on such output poses additional risks for the patients. Again, ultimately, is the patient expectation aligned with the use or disclosure of his or her data in this manner? How transparent has the provider been in its explanation, and does the patient understand the risks? Procedures should be drafted and added to any framework providing patients this additional right to know how clinical decisions are being made about them.

Further, as we innovate, we also encounter unintended consequences – as technology rolls out, patient experience and patient engagement may be negatively impacted. Since AI is driven by large sets of data that need to be input by clinicians, the less face time providers have with patients, the more potentially subordinate that relationship may become. AI can certainly agitate this relationship if not adopted with certain mitigation steps in mind, depending on the workflows, data needs, time to input data, etc. In general, healthcare, the industry that requires empathy more than any other, may actually lack that empathetic ingredient that should support trust within the encounter. We have seen where technology, a double-edged sword, has made the practice of medicine more distanced from patients

and has impacted patient experiences. Within the strain of healthcare systems to see more patients and provide better results, time and resources are stretched to the limit. It is increasingly difficult to offer truly personalized approaches to medicine. Yet some of the solutions to the problems vexing healthcare are found in AI's promise of personalization. AI may help offer more tailored approaches, but the cost to patient primacy must be paramount in this calculation.

Embedding analytic insights from AI directly into application and operational and clinical workflows is required to advance the notion of better quality and outcomes at a lower cost. However, the ecosystem in which this occurs must be patient centric and must include long-term leadership commitment, patient experience and engagement, patient empathy, data empathy, robust security, thoughtfully implemented innovative tools, well vetted holistic platforms and contracts, a formal data governance framework to empower the rich data, and a team of teams approach to care coordination that encompasses a blending of legal, analytics, governance, clinical, and security professionals. When the patient is locked in the center of this ecosystem, maturation of patient care through AI has its best hope of accomplishing the widest variety of goals, while maintaining the strictest level of patient preeminence. Therefore, another element of emerging frameworks should include the consumer's rights related to the use of AI.

Reducing the Risk of Synthetic Content: *NIST is soliciting insights concerning synthetic content creation, detection, labeling, and auditing. The aim is to contribute to the Secretary of Commerce's report to the Office of Management and Budget (OMB) and the Assistant to the President for National Security Affairs by identifying existing and potential science-backed standards, tools, methods, and practices to mitigate the risks of synthetic content generated by AI technologies. The focus includes authentication, labeling techniques, detection, resilience against manipulation, economic feasibility, prevention of harmful content, considerations for different AI model types, applicability across AI lifecycle stages, software testing, and tools maintenance for analyzing synthetic content labeling and authentication.*

Synthetic data, from our estimation, comes with both risks and great opportunities. As it relates to healthcare, there are two particular areas where we see benefit to its usage. First, health systems are being inundated with new technology which needs appropriate testing, but it is difficult to keep up with the pace of change and advancement in these algorithms. One of the long-standing barriers to the AI lifecycle and software testing is the handling of protected health information (PHI) and its associated legal and compliance parameters. Synthetic data offers an opportunity to avoid the sharing of PHI for testing by offering a realistic alternative. Second, synthetic data continues to be valuable in addressing health inequities and underrepresented segments of our population. Synthetic data creation of minority groups is somewhat typical in model training and has boosted the performance of developed tools.

That said, the risks related to data quality and hallucination need to be understood and considered moving forward. In our field, we have seen evidence that synthetic data generation struggles with patient biomarkers. Creators of large language models (LLMs) boast the ability for these models to be able to generate synthetic data from a few examples. Though LLMs themselves have constraints, the user must have an in-depth understanding of their training data to ensure bias and ethical concerns are not introduced.

Lastly, there is a difference between real data that is masked or de-identified and synthetic data. Synthetic data may be seen as computer generated data derived from existing data sets, reflecting the

characteristics of real-world data, mathematically or statistically, while used to improve AI models, protect sensitive data, and potentially mitigate bias. Masked data is real-world data that has been modified to protect sensitive information. Synthetic data can be used when real-world data is not available or when privacy concerns outweigh the use of real-world data. However, such data would need to be thoroughly tested to ensure that it properly represents real-world data (i.e., ensuring it does not accentuate or create bias) or is contextually accurate.

Advance Responsible Global Technical Standards for AI Development: *NIST is requesting comprehensive insights on developing and implementing AI-related consensus standards, cooperation, and coordination in alignment with Section 11(b) of the EO. The objective is to assist the Secretary of Commerce, in collaboration with other relevant agencies, in creating a global engagement plan for fostering AI consensus standards guided by principles outlined in the NIST AI RMF and the U.S. Government National Standards Strategy for Critical and Emerging Technology. Topics of interest include AI nomenclature, data handling best practices, impactful AI system types, model training guidelines, trustworthiness standards, risk management, application-specific standards (e.g., computer vision, facial recognition), stakeholder inclusivity in standardization processes, strategies for global standards adoption, mechanisms for international collaboration, implications of standards on competition and trade, and methodologies for assessing the impact of international engagements concerning AI standards.*

Technology can help set the guardrails for development and implementation. While giving developers the freedom they need to experiment, having a clear architecture for preferred tools that will integrate the development environment to production clearly helps organize development. Further, this architecture goes beyond development, and particularly influences model monitoring and maintenance standards which can be easily overlooked. Within the data architecture, having a cohesive environment that supports the entirety of the model lifecycle will end up facilitating faster innovation. The end of that lifecycle, which includes model monitoring, can be a tool to alert for issues, the need for model retraining, or model retirement. The problem is that, without a clear national standard, each organization will have their own unique architecture and set of guardrails. While this is on one hand appropriate, as each organization has its own processes, culture, and risk tolerance, coming up with national standards on baseline technology and architecture could lift all into a more cohesive way of developing AI. Further, while NIST AI MRF is voluntary, like many other frameworks, incentives such as safe harbor status related to compliance with NIST AI MRF should be considered.

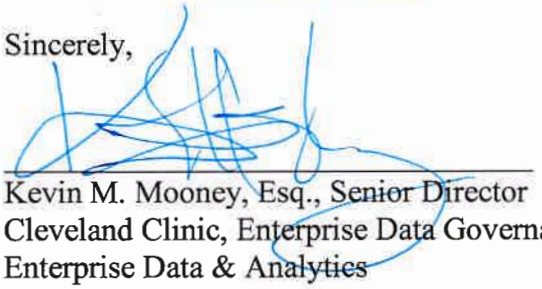
Much of these theories and considerations all converge on how thoughtful an organization has been in thinking through, constructing, resourcing, and holding accountable its data management program, which will include AI as a subset. What are the operational considerations to using AI and blending cybersecurity, legal theory, and ethics into operations? Is the organization cognizant of likely pitfalls in standards of care and foreseeability? Is the organization well informed about potential risks to patients and causes of actions based on AI tools? How is AI rolled out, who is trained, what is monitored, and how is the system maintained? How is the organization empowering its data and otherwise managing the data and using AI tools for the primary benefit of the patients? Is AI aligned with enterprise strategy, and has the value of the asset been articulated? Has the organization developed AI uniformity and measurable standards? Lastly, has the organization evolved inspiring leadership in AI driven data practices and the promotion of accountability to the notions of patient primacy in the provision of healthcare?

Our experience suggests that high-quality development and implementation standards at an organization come from the integration of consistent technologies as well as mandated, empowered, active, and strong governance. Governance begins with a multi-disciplinary team that ranges from developers of AI tools to individuals who could be impacted by their use (patients in our case, even indirectly). Between those two ends of the spectrum, having expertise from legal, cybersecurity, information technology, and ethics is critical to the creation and execution of these standards. At our organization, this cross-functional team comes together to review new project intake and proposals for production after standards or prerequisites are met. No organization should build and deploy AI algorithms without having a structure in place to manage and govern innovation, facilitate repeatable and documented processes, and provide communicated guardrails to its organization. We would suggest once an organization's own house is in order, they should be coordinating and collaborating with cross-industry, and at times, inter-industry partners, to share best practices and lift the universal education of all involved. Governance has a baseline but is a living theory, and its documentation and processes can be improved over time.

Further, to assert leadership internationally and enhance collaboration with international partners, the government needs to set standards that are on par with the international community, and then produce a funnel for individual organizations to share best practices across borders. In other words, the power of the United States government abroad can be augmented by the country's innovative companies having a voice fostered by NIST and other federal agencies in that international arena. This could be fostered through consortiums targeting specific elements of AI, for instance, or targeting AI as a whole but siloed in individual industries.

Thank you for conducting a thoughtful process that allows us to provide input on such important issues and for your consideration of this information. Should you need any further information, please contact me at mooneyk@ccf.org.

Sincerely,



Kevin M. Mooney, Esq., Senior Director
Cleveland Clinic, Enterprise Data Governance Office
Enterprise Data & Analytics