Michelle Grisat
Director of Health and Regulatory Policy
National Nurses United
155 Grand Avenue, Suite 100
Oakland, CA 94612

February 2, 2024

Information Technology Laboratory
ATTN: AI E.O. RFI Comments
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 8900
Gaithersburg, MD 20899–8900.

Dear Director Laurie E. Locascio,

On behalf of more than 225,000 registered nurses (RNs) who work as licensed health care professionals in every state in the nation, National Nurses United (NNU) submits these comments in response to the National Institute of Standards and Technology (NIST)'s Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence.[1]

NNU appreciates that the Biden Administration and NIST are assessing the risks associated with generative AI. However, relying on voluntary risk management is not enough. Technology developers and health care corporations cannot be allowed to decide what level of risk to patients and workers is acceptable in their pursuit of profit—we have seen over and over again that too many companies are happy to risk the lives, jobs, and privacy of the public. We need binding regulation on the use of AI in health care to protect patients' safety and workers' autonomy and security. NIST should take this opportunity to set concrete safety standards that make it clear what uses are too dangerous and transparency standards that allow the public to see if developers and deployers are following the best practices NIST recommends.

NNU is extremely concerned about the use of AI in health care. Nurses across the country are already experiencing the problematic effects of AI. The decisions to implement AI-driven technologies are typically made with little to no knowledge of either nurses or patients, putting both at risk of negative safety and rights impacts. AI is being used to replace registered nurses, who have the education and clinical experience necessary to effectively exercise professional judgment in patient care, with lower cost staff following AI-generated prompts. However, each patient is unique and health care is made up of non-routine situations that require human intelligence and perception as well as empathy and kindness. Thus, health care should be

[1] 88 FR 88,368 (Dec. 21, 2023).

provided by licensed health care professionals who can discern whether or not AI-generated guidance is appropriate for a particular patient.

NNU discusses below our concerns with the failure of AI industry self-regulation and gaps in current health care AI regulation and a summary of some of our serious concerns with the use of generative AI in health care, where the stakes are life and death. NNU urges NIST to be clear that generative AI should not be used for safety- and rights-impacting applications. The evidence shows that it can be biased and unreliable and is typically inscrutable. For these reasons, it should not be used in health care or other areas that affect people's safety and rights. NIST should establish as a "best practice" the concrete standards that should be met before generative AI can be considered for these contexts. If industry does not comply voluntarily with standards that will truly protect patients and workers, then legislators and regulators must take the next steps. Strong NIST guidance on generative AI can serve as a starting point.

Sincerely,

Michelle Grisat

Michelle Grisat
Director of Health and Regulatory Policy
National Nurses United

I. The AI Industry Lacks Sufficient Regulation to Protect Public Health and Safety and Civil Rights. *RFI Question 1*

A. The AI Industry Is Failing to Adequately Self Regulate

The AI industry approach is to deploy fast and iterate. Systems have been released to, and upon, the public without full consideration of their risks or meaningful consent by the people affected, let alone allowing for democratic participation in determining their deployment. This approach is particularly dangerous in safety- and rights-impacting industries like health care. Technology and health care corporations have chosen to widely roll out predictive algorithms in clinical care settings without sufficient proof of efficacy and then changed them significantly after they were shown by independent researchers to have harmful effects on patients.[2]

NIST should develop standards that can serve as a template for mandatory regulation and that are based on the precautionary principle. According to Harvard University Professor A. Wallace Hayes, the precautionary principle stands for the proposition that, "When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically."[3] The precautionary principle is essential when dealing with generative AI, where the potential effects are highly undetermined and there is the potential for that impact to change significantly after a system is rolled out.[4]

While there are some specific regulations for certain AI systems used in health care settings, these have not been sufficient to curtail the implementation of dangerous systems in clinical practice. Existing efforts either rely entirely on industry to self-regulate through risk management frameworks without rigorous standards that must be met before a system is used, or exempt broad categories of systems on the basis of human review or decision-making without ensuring that the human participant can or will exercise independent exercise judgment without time or management pressure to comply with the system output.

B. Gaps in Current Federal Regulations Leave the Public Unprotected.

The Food and Drug Administration performs premarket review of medical devices. The 21st Century Cures Act creates a distinction between types of "clinical decision support [CDS] software": CDS software that is intended to provide recommendations to a health care professional and allows the health care professional to independently review the basis for the recommendations, and CDS software that purports to diagnose patients or create treatment plans directly. The Cures Act exempts the former from the definition of medical device. This distinction allows the FDA to steer clear of regulating health care professional practice, which is

[2] *See, e.g.,* Ross, C. Epic's overhaul of a flawed algorithm shows why AI oversight is a life-or-death issue. Oct 24, 2022. STAT. https://www.statnews.com/2022/10/24/epic-overhaul-of-a-flawed-algorithm/.

[3] Hayes, A.W. The precautionary principle. Arh Hig Rada Toksikol. 2005 Jun;56(2):161-6. PMID: 15968832.

[4] *See, e.g.*, Palmer, K. (2023). The 'model-eat-model world' of clinical AI: How predictive power becomes a pitfall. *STAT*. https://www.statnews.com/2023/10/10/the-model-eat-model-world-of-clinical-ai-how-predictive-power-becomes-a-pitfall/.

traditionally overseen by states.[5] However, in practice, it leaves significant gaps in regulation where software developers can claim to regulators that their systems merely offer recommendations to clinicians while they promise health care employers that the system will make the diagnosis and treatment planning process faster—cutting out the time a human takes to analyze information and ensure whether it is appropriate for each particular patient. No one is prospectively reviewing whether the information provided to clinicians is actually adequate to review the basis for recommendations or auditing whether they have the time, information, and ability to do so.

The Office of the National Coordinator for Health Information Technology (ONC) also provides some oversight of AI used in certified health information technology, recently finalizing rule HTI-1 that includes what it calls an "algorithm transparency" requirement.[6] However, the algorithm transparency only requires that specified "source attributes" be provided, where applicable, with 31 source attributes specified for AI-based "predictive decision support interventions" (DSI) defined as "technology that supports decision-making based on algorithms or models that derive relationships from training data and then produces an output that results in prediction, classification, recommendation, evaluation, or analysis." HTI-1 requires only limited transparency for source attributes, eliminating the proposed requirement for the source information to be available to users in plain language "via direct display, drill down, or link out."[7] Instead, the rule only requires this information be available to "a limited set of identified users" determined by the deployer and may not even be available to the individual nurses and physicians using the AI.[8] Additionally, the rule does not require that developers provide information about which specific datasets that were used to train a DSI. NIST's criteria for "trustworthy" AI strongly suggests that this information be available not only to clinicians but also to patients and the public.

Finally, HTI-1 offers no specific metric for determining whether a DSI meets any of the FAVES criteria (fair, appropriate, valid, effective, and safe) that the rule focuses on, leaving it up to users to determine independently whether a DSI has been validated and tested appropriately[9] and that the DSI meets "acceptable" FAVES levels.[10] Notably absent from FAVES criteria are whether a DSI is explainable or interpretable, requiring only that developers provide risk analysis and risk mitigation information.[11] As discussed below, explainability and interpretability are crucial for clinicians to exercise their professional judgment and communicate with patients about their diagnosis and treatment.

[5] Evans, B.J. & Pasquale, F. Product Liability Suits for FDA-Regulated AI/ML Software. In: Cohen I.G., Minssen, T., Price II W.N., Robertson, C., and Shachar C., eds. *The Future of Medical Device Regulation: Innovation and Protection*. Cambridge University Press; 2022:22-35.

[6] 89 FR 1192 (Jan. 9, 2024). Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing.

[7] https://www.federalregister.gov/d/2023-28857/p-834

[8] 45 CFR § 170.315(b)(11)(v)(A)(1)

[9] https://www.federalregister.gov/d/2023-28857/p-640

[10] https://www.federalregister.gov/d/2023-28857/p-623

[11] https://www.federalregister.gov/d/2023-28857/p-935

II. NIST Guidelines Should Recommend that Generative AI Is Prohibited in Health Care until Effective Safety and Rights Protections Are in Place.

In its new standards, NIST must make clear that generative AI should be prohibited in safety- or rights-impacting industries unless and until a rigorous regulatory structure is in place to ensure that it meets strenuous standards that ensure public health and safety and civil rights are protected. NIST should recommend that generative AI developers use terms of use, contractual arrangements, or other mechanisms to prohibit use of their models in health care applications until such a structure is in place. Moreover, generative AI must be transparent, explainable, and interpretable for patients and for clinicians. Although NNU knows that this is a high bar, it is essential that patients and clinicians fully understand the basis for all diagnostic, prognostic, and treatment decisions.

NIST should fill the gaps in regulation discussed in Section I by creating a framework for generative AI that requires rigorous premarket testing and verifies AI systems meet sufficient standards before use in health care, as well as continued testing during use. This framework should set safety standards for industry to use now and create a template for drafting legislation to increase legally required premarket review if and when industry fails to comply with best practices voluntarily.

A template for mandatory regulation based on the precautionary principle should ensure that any AI system, particularly generative AI, used in health care or other safety or rights-impacting settings meets specific standards. In NIST’s original AI Risk Management Framework, NIST identifies characteristics of “trustworthy” AI systems including “valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed” and calls for risk management assessment throughout the development and deployment of a system. These are important factors, but the risk management approach leaves each developer and deployer to independently decide how much risk to the public is acceptable. For areas that impact safety and rights, clearer rights-based standards are needed. The Office of Management and Budget’s draft memorandum on agency use of artificial intelligence provides guidance on identifying safety and rights-impacting areas, which may be a helpful starting point for NIST.[12] NIST has an opportunity to develop more fully each of these concepts and determine how firm lines can be drawn to ensure appropriate standards are met before deployment. Risk assessment is not enough—industry must be held to clear standards that ensure safety and rights for any system before implementation.

In addition to drafting a framework for protections that must be in place before safety- and rights-impacting AI is used, NIST should also identify red lines: tasks for which AI should never be used. For example, advocates have persuasively argued that automated systems should never be used for management decisions such as hiring, firing, or worker discipline.[13] In health

[12] OMB–2023–0020, Request for Comments on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum.

[13] *See* discussion of algorithmic management in Kak, A. & West, S. M. (2023). AI Now 2023 Landscape: Confronting Tech Power. AI Now Institute. https://ainowinstitute.org/2023-landscape.

care, AI should not be used for determinations of worker competency to provide care or assignments to particular units, or for delegation of tasks from an RN to an un- or lower-licensed worker. Generative AI and algorithms derived from large language models should not be used for communication with patients, health records, or communication about patient care since these models have a low level of explainability and periodically output nonsensical results known as "hallucinations."[14] For example, in a hospital providing care around the clock, effective, interactive communication between the nurse whose shift is ending and the nurse who is taking over patient care is crucial to ensuring patient safety. Replacing these face-to-face discussions that allow for interactive communication with a written report created by generative AI endangers patients.

III. Generative AI Cannot Be Used Safely in Health Care and Safety- and Rights-Impacting Contexts. *RFI Question 1a*

NNU is particularly concerned about the use of generative AI in health care because of its potential to harm patients due to these systems' inaccuracy, opacity, and inequities.

A. Generative AI Is Often Opaque, Inscrutable, Error-prone, Biased and Inequitable.

Generative AI has serious problems that prevent it from being safe for use in health care. The manner that generative AI systems "learn" is often opaque, even to the developers of the systems.[15] As a result, generative AI systems may have hidden capabilities that developers and deployers are unaware of. For example, researchers demonstrated OpenAI ChatGPT's vulnerabilities when they extracted underlying training data from the model, causing it to reveal personal information from dozens of real individuals in response to a prompt.[16]

The opaque nature of these systems conflicts with one of the core principles of trustworthy AI—transparency, as the NIST AI Risk Management Framework (RMF) outlines. Under this framework, trustworthy AI requires transparency and accountability among other characteristics.[17] Accountability depends on transparency, measured by the extent to which

[14] Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3), 226-234.
Bhattacharyya, M., et al. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus, 15*(5).

[15] *See, e.g.*, Wach, K., et al. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7-30.

[16] Nasr, M., et al. (2023). Scalable extraction of training data from (production) language models. *arXiv* preprint arXiv:2311.17035. As reported by Pearson, J. (2023, Nov. 29). ChatGPT Can Reveal Personal Information From Real People, Google Researchers Show. *Vice News.* https://www.vice.com/en/article/88xe75/chatgpt-can-reveal-personal-information-from-real-people-google-researchers-show - accessed January 26, 2024.

[17] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI. National Institute of Standards and Technology, U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1 - accessed January 29, 2024.

model developers and deployers provide users with access to meaningful information.[18] The scope of transparency broadly spans design decisions, model training, and the underlying training data.[19] Importantly, while not sufficient on its own, transparency is necessary for actionable redress when AI's outputs are incorrect or otherwise lead to harm.[20]

However, research demonstrates the extent to which generative AI consistently fails across measures of transparency. Researchers from the Stanford Institute for Human-Centered Artificial Intelligence indexed the ten most prominent generative AI models by assessing their transparency across 100 indicators, including data, labor practices, environmental impact, methods, risks, and downstream distribution channels.[21] According to their findings, the highest score among model developers was 54 out of 100, with an average score of 37. These low scores demonstrate a "pervasive lack of transparency" among major generative AI developers, with "virtually no transparency about the downstream impact" of these models.[22] Among the gaps and "pervasive blind spots," are that none of the developers provide a mechanism for users to seek redress.[23]

Additionally, generative AI developers also may fail to provide methods for users to understand and evaluate how and why a model generates an output. That is, they lack explainability and interpretability, which are among the core characteristics of trustworthy AI under the NIST AI RMF. These characteristics measure the extent to which models provide information that will help users understand, interpret, and evaluate how the system functions and produces outputs, and thus answer the questions of "how" and "why" a model made a decision.[24] However, researchers have criticized generative AI for its inscrutability, and the complete lack of methods for users to understand how the model arrives at the results it generates.[25] Generative AI produces outputs that researchers have described as "artifacts," rather than verifiable "decisions,"[26] predictions, or estimations. Because of these models' non-deterministic nature, researchers have characterized generative AI as a "stochastic parrot"[27] that often generates a different output with each use, "not based on logic or reasoning" but on the probability of outcomes.[28] Generative AI's probabilistic basis means that these systems inherently evade standards of explainability and interpretability.

[18] *Ibid.*

[19] *Ibid.*

[20] *Ibid.*

[21] Bommasani, R., et al. (2023). The foundation model transparency index. *arXiv preprint arXiv:2310.12941*.

[22] *Ibid.*

[23] *Ibid.*

[24] Tabassi, 2023.

[25] Dwivedi, Y. K., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

[26] *See* Sun, J., et al. (2022, March). Investigating explainability of generative AI for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces* (pp. 212-228).

[27] Bender, E. M., et al. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

[28] Dwivedi, et al., 2023.

Other hazards, common to AI and algorithm-based systems, are the racial, ethnic, and gender-based biases lurking in the underlying training data that generative AI may perpetuate and even exacerbate. Researchers and investigative journalists have found pervasive racial and gender biases in generative AI models that produce images from a text prompt. One forthcoming study found that OpenAI's DALL-E Mini text-image generator produced images of only men for dozens of occupations (e.g., pilot, builder, plumber), and solely women for others (e.g., hairdresser, receptionist).[29] Further, the model generated images of most occupations either primarily or solely represented by white people (e.g., farmer, painter, software engineer), whereas non-white people represented only a few occupations (e.g., pastor, rapper) in the model's images.[30] Similarly, a news publication conducted an investigation that exposed biases in Stability AI's Stable Diffusion image generator, finding that subjects with lighter skin tones dominated images of high-paying jobs like lawyers and doctors, while subjects with darker skin tones dominated service occupations like fast-food worker and dishwasher.[31] Importantly, the model's representation of these occupations was even more segregated than reality.[32]

The demonstrated capacity for generative AI to reproduce harmful bias has major implications for standards related to fairness, one of the core characteristics of trustworthy AI under the NIST AI RMF.[33] While standards of fairness can be complex, it measures the extent to which models enshrine equality and equity and their propensity to address harmful bias and discrimination.[34] As the NIST AI RMF acknowledges, AI models "can potentially increase the speed and scale of biases and perpetuate and amplify harms to individuals, groups, communities, organizations, and society."[35]

Additionally, numerous studies show that generative AI routinely "hallucinates" by simply making up inaccurate information, such as academic and legal references without a clear basis or reason for doing so.[36] According to researchers like Emily Bender who study generative

[29] Cheong, M., et al. (forthcoming) Investigating gender and racial biases in DALL-E Mini Images. *Acm Journal on Responsible Computing*. https://philarchive.org/rec/CHEIGA-2 - accessed January 26, 2024.

[30] *Ibid.*

[31] Nicoletti, L. & Bass, D. (2023, June 8). Generative AI Takes Stereotypes and Bias From Bad to Worse. *Bloomberg.* https://www.bloomberg.com/graphics/2023-generative-ai-bias/ - accessed January 26, 2024.

[32] *Ibid.* Finding that Stable Diffusion "takes racial and gender disparities to extremes – worse than those found in the real world."

[33] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI. National Institute of Standards and Technology, U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1 - accessed January 29, 2024.

[34] *Ibid.*

[35] *Ibid.* at p. 18.

[36] Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2).

Siontis, K. C., Attia, Z. I., Asirvatham, S. J., & Friedman, P. A. (2023). ChatGPT hallucinating: can it get any more humanlike? *European Heart Journal*. 00, 1–3.

McGowan, A., et al. (2023). ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326, 115334.

Dixon, H. B. (2023). My "Hallucinating" Experience with ChatGPT. *The Judges' Journal*, 62(2), 37-39.

AI hallucination, the problem "isn't fixable, … [i]t's inherent in the mismatch between technology and the proposed use cases."[37] Likewise, there is a fundamental mismatch between generative AI and its proposed use in health care.

B. Generative AI's Deficiencies Are Particularly Dangerous in the Health Care Setting.

Applied in health care settings, generative AI's demonstrated problems are likely to have dire consequences for patients and the nurses that care for them. According to proponents, generative AI in health care would perform benign "language manipulation tasks such as summarization, translation, and answering questions."[38] However, the capacity for generative AI to fabricate or hallucinate is particularly dangerous in health care given that the falsehoods may be subtle and not easily detected by clinicians and patients alike.[39] One study found that ChatGPT provided answers to medical questions that were "deceptively real," by fabricating 69 percent of references.[40]

Further, deploying generative AI in health care would occur in an industry that routinely uses technology to advance a business model that requires cutting labor costs and rapidly churning out patients, rather than advancing safe, effective, equitable care. For example, patient transfers are one of the most dangerous points in a patient's care, yet employers have used AI-based technology to side-step vital RN-to-RN communication during patient hand-off and transfer of duty. NNU nurses report that automated communication leaves out important information while overburdening nurses with information that is not essential. Thus, these systems force nurses to waste precious time searching medical records for information that nurses could have relayed in a complete and accurate manner during a brief interpersonal interaction.

Similarly, the health care industry is using generative AI to summarize clinical encounters to reduce the labor costs of direct RN-to-RN communication. Yet test sessions in which generative AI summarized a physician-patient encounter demonstrate that it leaves out important information and fabricates or "infers" information with no basis whatsoever according to one

Merken, S. (2023, Jun. 6). New York lawyers sanctioned for using fake ChatGPT cases in legal brief. *Reuters*. https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/ - accessed January 26, 2024.

Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38.

[37] *See* O'Brien, M. (2023, Aug. 1). Chatbots sometimes make things up. Is AI's hallucination problem fixable? *AP News*. https://apnews.com/article/artificial-intelligence-hallucination-chatbots-chatgpt-falsehoods-ac4672c5b06e6f91050aa46ee731bcf4 - accessed January 26, 2024.

[38] *See* Shah, N. H., Entwistle, D., & Pfeffer, M. A. (2023). Creation and adoption of large language models in medicine. *JAMA*, 330(9), 866-869.

[39] See Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13), 1233-1239.

[40] Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3), 226-234.

study.[41] As the study highlighted, ChatGPT generated a summary that included the patient's body mass index (BMI) even though the transcript did not contain the patient's weight, information that a BMI calculation requires.[42] At the same time, the model omitted key observations the clinician made regarding the patient's low blood pressure and low pulse.[43] Thus, generative AI clearly omits or fabricates critical information that could easily result in illness, injury, or death.

Generative AI inscrutability may compound this hazard by failing to provide a method for clinicians to understand and evaluate both its underlying decision-making process and its output, i.e. it lacks explainability, and interpretability. In health care, explainability and interpretability are particularly important as clinicians must evaluate whether generative AI output is appropriate for a particular patient given that patient's needs and values. Without a method of evaluation, generative AI may pit the professional judgment of a clinician against the model with little recourse, an observed phenomenon in clinical settings that researchers call "divergence."[44] Relatedly, generative AI inscrutability also undermines informed consent, patient autonomy, and the clinician-patient relationship by failing to provide sufficient information for patients to engage in shared decision making based on their needs, values, and preferences.

Finally, generative AI's propensity to reproduce racial bias in medicine raises major fairness and equity concerns. Researchers from the Stanford School of Medicine assessed four generative AI models, including those that are currently in use at hospitals,[45] and found that they all produced harmful racist content by either perpetuating debunked, race-based medicine or propagating unsubstantiated racist tropes.[46] For example, responses to questions related to racial differences in pain thresholds perpetuated repugnant, race-based claims regarding cultural beliefs, such as Google Bard's response that "Some Black patients may be less likely to report pain because they believe that it is a sign of weakness or that they should be able to 'tough it out.'"[47] In response to the same question, Anthropic Claude stated that biological differences existed between white and Black patients: "For example, studies show Black individuals tend to have higher levels of GFRα3, a receptor involved in pain detection."[48] Alarmingly, GPT-4 generated race-based content *without any reference to race* in response to questions regarding kidney function.[49]

[41] Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13), 1233-1239.

[42] *Ibid.* Figure 2. Using GPT-4 to Assist in Medical Note Taking.

[43] *Ibid.*

[44] Beaulieu-Jones, B. K., et al. (2021). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?. *NPJ digital medicine*, 4(1), 62.

[45] Eddy, N. (2023). Epic, Microsoft partner to use generative AI for better EHRs. Healthcare IT News. https://www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-ai-better-ehrs - accessed January 31, 2024. Noting that UC San Diego Health; UW Health in Madison, Wisconsin; and Stanford Health Care have already deployed generative AI as part of an "integrated" EHR system.

[46] Omiye, J. A., et al. (2023). Large language models propagate race-based medicine. *NPJ Digital Medicine*, *6*(1), 195.

[47] *Ibid.*

[48] *Ibid.*

[49] *Ibid.*

As these examples demonstrate, generative AI systems are prone to serious errors and disturbing falsehoods that can directly impact the safety and rights of patients. The deficiencies and uncertainties around generative AI's theoretical performance, as well as its implementation by health care corporations which seek to keep labor costs low and churn patients out as soon as possible, warrants a precautionary approach. NIST should therefore advise hospitals and other health care facilities to refrain from adopting generative AI technology unless and until these serious deficiencies are resolved.

IV. Human Oversight of AI Is Necessary But Does Not Justify Use of Systems that Are Not Proven to Meet Stringent Regulatory Requirements, Including Requirements for Safety, Accountability, Reliability, Fairness, and Explainability. *RFI Question 1a*

Regulatory efforts to prevent the worst outcomes of AI output in health care have largely centered around requiring human oversight. While they vary in design, these so-called "human-in-the-loop" proposals coalesce around requiring some degree of human review of AI output.[50]

However, research has consistently expressed reservations about whether humans can effectively counteract the harms of flawed algorithms.[51] Indeed, evidence suggests that there are serious challenges to humans' ability to oversee AI effectively, particularly in fast-paced environments where human workers might not have the time or capacity to adequately review the recommended output. Many AI systems do not provide sufficient transparency regarding their process to allow for effective review, or rely on machine learning and inferences drawn from large data-sets that make such transparency and explainability effectively impossible.

To account for the individual needs of each patient and context, clinicians must be afforded discretion to provide patient care based on their professional judgment. However, their ability to effectively exercise their professional judgment can be dangerously skewed by the recommendations of faulty AI—including AI with a high overall performance rate that does not account for the particular patient or circumstance. Developers and deployers use human oversight policies to add a false sense of legitimacy to the use of AI in health care spaces without addressing the fundamental problems with these tools. Researchers have determined that "human oversight policies provide a false sense of security in adopting algorithms" and enable

[50] Kak, A. & West, S. M. (2023). AI Now 2023 Landscape: Confronting Tech Power. AI Now Institute. https://ainowinstitute.org/2023-landscape.

[51] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681. (citing Engstrom DF, Ho DE, Sharkey CM, Cuéllar M-F. Government by algorithm: Artificial intelligence in federal administrative agencies. Administrative Conference of the United States; 2020 https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf; European Commission. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act); 2021 https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence; UK Information Commissioner's Office. Guidance on the AI Auditing Framework; 2020 https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-forconsultation.pdf).

developers and deployers to shirk accountability for AI harms.[52] Worse still, this model shifts this accountability from the developers of the algorithm and the health care facilities that deploy them to the human worker tasked with overseeing an algorithm, despite the fact that this human worker often has no control over how the algorithm works or when and how it is used.

In light of these flaws, regulators must provide greater scrutiny of whether human oversight mechanisms provide adequate opportunity to exercise professional judgment, and determine the extent to which AI should be permitted in health care processes through red line measures.[53] All AI that may affect patient care should be subject to thorough regulatory review including premarket and ongoing testing that ensures both that the AI recommendations meet stringent requirements for patient care, including safety, accountability, reliability, fairness, and explainability and that the human oversight mechanisms are effective and can function effectively even under the high pressure conditions of health care.[54] NNU urges NIST to emphasize that human oversight can never be a substitute for other rigorous testing and validation.

A. Empirical Evidence Shows that Human Oversight Cannot Fully Counteract Flawed Algorithms.

Human oversight policies rest on the faulty assumption that workers can provide protection against AI errors, biases, and inflexibility. The evidence demonstrates that this is often not the case.[55] This is obviously true in situations where the human worker has no ability to override the decision before it is implemented, and only provides post hoc review.[56] However, even when human overseers are only given suggested actions by AI systems and are themselves vested with the final decision-making power, the evidence indicates the humans naturally defer to automated systems, even where there is no indication that this deferral is warranted. Likewise, evidence indicates that attempts to integrate human decision-making with algorithmic systems generate unique problems that do not exist in a human-only environment. Thus, developers and deployers of algorithms must not be allowed to use human oversight to skirt other quality protections and they must be required to show not only that there is human oversight, but that the oversight mechanism is effective.

It is now well-established in the academic literature that human oversight of algorithmic decision-making is influenced by the psycho-social phenomenon known as “automation bias.”[57] As described by Green:

> [a] long-standing body of research shows that, across a wide range of domains, automated decision-support systems tend to alter human decision-making in

[52] Green, 2022.

[53] Kak, A. & West, S. M. (2023). AI Now 2023 Landscape: Confronting Tech Power. AI Now Institute. https://ainowinstitute.org/2023-landscape.

[54] Green, 2022.

[55] *Ibid*.

[56] *Ibid*.

[57] *Ibid*.

> unexpected and harmful ways. Numerous studies have demonstrated that people (including experts) are susceptible to "automation bias"—i.e., people defer to automated systems, reducing the amount of independent scrutiny that they exhibit when making decisions.[58]

This tendency to defer to automated systems has been demonstrated in both omission errors – failing to act independently when not prompted to do so by the automated system—and commission errors—acting on the advice of the automated system when the evidence indicates that the system is incorrect.[59] Worse still, this phenomenon has been shown to persist even among experts with experience performing the decision-making task independently, and in high stakes situations where errors can have tremendous consequences.

In addition to automation bias, the automation of certain tasks can make the remaining parts of the task, which still must be completed by humans, more difficult or tedious, and can cause human skills and judgement to deteriorate. This is particularly true in nursing, where skills are developed through clinical application and practice in addition to formal classroom education. As a result, "automated systems may simply lead to different types of errors rather than reducing overall errors as intended,"[60] and can lead to the deskilling of human operators. Deskilling leaves nurses less prepared to handle unusual situations or circumstances where the technology does not function. This also leads to burnout and may cause nurses to leave the profession.

Moreover, conditions in health care clinical environments are rarely ideal. As Khera, et al., acknowledge, "errors resulting from automation bias are likely to be further compounded by the usual time pressures faced by many clinicians."[61] Likewise, Jabbour, et al., found that many AI systems, particularly those based on large language models, cannot provide the level of explainability necessary for allowing meaningful oversight. It is therefore fair to assume that the application of AI systems based on opaque large-language-models, applied in the often faced-paced environment of clinical practice, will result in even higher instances of automation bias than those found in controlled research environments.

In sum, it is clear that human oversight policies designed as a safeguard against potentially life-threatening errors made by AI and large language model systems in clinical environments are not sufficient to prevent adverse outcomes in AI decision-making. AI must be proven to meet stringent requirements, including safety, accountability, reliability, fairness, and explainability, before they are deployed in health care settings independent of the role of a human end user.

[58] *Ibid*. Citing Parasuraman R. and Manzey D.H. Complacency and bias in human use of automation: An attentional integration. Human Factors 2010;52(3):381–410. doi: 10.1177/0018720810376055; Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? Int. J. Hum.-Comput. Stud. 1999;51(5):991–1006. doi: 10.1006/ijhc.1999.0252.).

[59] Green, 2022. (citing Parasuraman & Manzey, 2010; Skitka et al., 1999).

[60] Green, 2022. (citing Skitka et al., 1999).

[61] Khera, 2023.

B. Human Oversight Policies are Used to Legitimize the Use of Flawed and Dangerous Algorithms

Human oversight mechanisms are being used to legitimize the use of problematic algorithms, serving as a fig leaf to reduce scrutiny and shift blame. This, in turn, results in a false sense of security being placed on the use of algorithmic decision-making in high-stakes scenarios, such as health care, leading the public to believe that these systems are safe and reliable when the evidence indicates the contrary.

Moreover, the reliance on human oversight systems shifts accountability for algorithmic harms from developers and deployers of these technologies, such as technology vendors and health care providers, who have control over how these systems operate and when and how they must be used, to frontline operators, who are often required to use this technology by their employer under threat of discipline or discharge. As Green describes, "[h]uman oversight policies position frontline human operators as the scapegoats for algorithmic harms, even though algorithmic errors and injustices are typically due to factors over which frontline human overseers have minimal agency, such as the system design and the political goals motivating implementation."[62] This, in essence, allows health care providers and technology vendors to have it both ways—they can promote the use of automated decision making on the basis that automated capabilities can exceed those of humans, while simultaneously defending the algorithm and those responsible for its implementation by placing the blame for errors on the human overseers. In addition to being manifestly unjust, this also allows for the further proliferation of problematic algorithms that are merely "rubber stamped" by human oversight requirements that provide no actual protection or security.

V. NIST Should Avoid Supporting Global Standards that Could Be an Obstacle to Local Regulation or Transparency Requirements. *RFI Question 3*

NIST should avoid endorsing any global standards that provide cover for companies to refuse to share details of algorithms that affect people's rights or safety with regulators, deployers, users, or the public. Global standards should create a minimum standard for transparency and never be an obstacle to further transparency or accountability.

International agreements can be a way for developers to dodge democratic accountability in the countries where they operate. The history of trade agreements illustrates this danger.[63] Trade agreements are often worked out in closed-door negotiations with little public input. However, industry lobbyists have ample opportunity to get their priorities included. For example, the 2019 US-Mexico-Canada Agreement (USMCA) and the 2018 Trans-Pacific Partnership included broad protections against government access to source code and algorithms. These

[62] Green, 2022.

[63] *See, generally* Kak, A. & West, S. M. (2023). AI Now 2023 Landscape: Confronting Tech Power. AI Now Institute. https://ainowinstitute.org/2023-landscape at p.79.

provisions are justified as protection for trade secrets but could potentially have serious ramifications for algorithmic transparency regulations.

While global technical standards will not be binding in the same way as trade agreements, they may influence future binding documents, set a precedent for future regulators, and guide industry practice. These standards should not encourage technical or legal mechanisms to hide how important algorithms function from regulators in the countries where they operate. Standards should encourage privacy protections for individuals' data against both governments and data mining corporations, but not protect the details of what developers are doing from the people affected by their products. Instead, the standards should encourage meaningful in-depth transparency with regulators and the public.

The benefits of uniform global standards are outweighed by the necessity of democratic control over systems that significantly impact people's lives. There may be room for global technical standards that can ensure systems are technically compatible with demands for transparency, privacy, trustworthiness, and other globally important principles. However, NIST must be clear that any global agreements cannot predetermine answers to questions like privacy and transparency versus interest in trade secrets that must be decided by the people whose rights are at stake.