

*Comment of*  
**THE CENTER FOR AI AND DIGITAL POLICY (CAIDP)**  
*to the*  
**NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY**  
*in response to*  
**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)**

**February 2, 2024**

We write to you on behalf of the Center for AI and Digital Policy (“CAIDP”) in response to the request for information<sup>1</sup> (“RFI”) issued by the National Institute of Standards and Technology (“NIST”) relating to NIST’s assignments under section 4.1, 4.5, and 11 of President Biden’s Executive Order concerning Artificial Intelligence (hereafter “EO 14110”).<sup>2</sup>

CAIDP is an independent non-profit organization based in Washington, DC. Our global network of experts advise national governments and international organizations on artificial intelligence (“AI”) and digital policy. We have also advised We are a global network of AI policy experts and advocates. We publish annually the *Artificial Intelligence and Democratic Values Index*, a comprehensive review of AI policies and practices around the world.<sup>3</sup>

We address specific items of the RFI in our comments below. However, our overarching recommendations to NIST to fulfill its mandate under EO 14110 are as follows:

1. Use a human-rights based approach to AI governance
2. Clearly recommend and establish the obligation to terminate
3. Recommend human rights impact assessments
4. Recommend continual and open monitoring and contestability of outcomes

---

<sup>1</sup> National Institute of Standards and Technology, *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)*, Federal Register (Dec. 21, 2023), <https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the>.

<sup>2</sup> *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Executive Order 14110 (Oct. 30, 2023), <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.

<sup>3</sup> CAIDP, *Artificial Intelligence and Democratic Values* (2023), <https://www.caidp.org/reports/aidv-2022/>.

5. Recommend “no-go” decisions for pseudo-scientific and human rights violating systems
6. Recommend transparency mechanisms including ‘white-box’ and ‘out-of-the box’ auditor access, use of public model cards, training set documentation
7. Intended use cases and exclusion criteria should be aligned with OMB’s guidance on “rights-impacting” and “safety-impacting” AI systems

In the Comments below, we provide the following recommendations in response to the specific issues identified in the RFI:

1. ***On the “Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness (including non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data;” [3.a]:*** NIST must account for the lack of underrepresented groups in AI development and how their absence is inconsistent with existing laws and international best practices on inclusivity and representation in AI. Best-practices and standards should guide development of AI tools/systems which are traceable, contestable, and contribute to accountability by their design.
2. ***On “[how to develop standards for] AI risk management and governance, including managing potential risk and harms to people, organizations, and ecosystems” [3.a]:*** NIST should account for the growing environmental implications of the AI supply chain as a part of its standard development process and consider existing obligations towards reducing carbon emissions and other environmental harms arising out of the development of AI systems.
3. ***On the “Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF, as well as harms related to repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses” [1.a.1]:*** we present a catalog of risks, the urgency in addressing them, and supply recommendations of standards that advance a rights-based approach to AI governance.
4. ***On the “Forms of transparency and documentation (e.g., model cards, data cards, system cards, benchmarking results, impact assessments, or other kinds of transparency reports) that are more or less helpful for various risk management purposes” [1.a.1]:*** we urge NIST to require ex-ante human rights impact assessments,

advance a transparency and disclosure framework, and continue an open public comment process.

We are conscious that NIST’s framework is not mandatory and does not create enforceable rights and obligations for private actors. However, NIST must be mindful that the approach and elements of its risk management framework will inform ongoing discussions on regulation and should be human-centered, designed to proactively set a robust floor for transparency, accountability, safety, and fairness obligations for the industry.

***Response 1: “Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness (including non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data” [3.a]***

CAIDP would like to draw attention to the lack of inclusivity and representation in AI development as a major cause of underinclusive or discriminatory AI. In its framework, NIST must address the lack of underrepresented groups in AI development, t, and how their absence is inconsistent with the mandate of EO 141110.

*The underrepresentation and exclusionary behaviours of AI systems are well-documented.* Facial recognition software has long struggled with individuals with darker skin tones,<sup>4</sup> and images generated by generative AI such as Stable Diffusion suggest a world dominated by white men.<sup>5</sup> In forensic investigation studies, DALL-E generated sketches of suspected criminals from text descriptions were frequently far different from the real photos and the suspects’ skin color skewed darker.<sup>6</sup> AI-powered screening, rental programs are biased against underserved populations and vulnerable groups.<sup>7</sup> Generative AI systems can produce

---

<sup>4</sup> Tom Simonite, *The Best Algorithms Struggle to Recognize Black Faces Equally*, Wired (Jul. 22, 2019), <https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/>.

<sup>5</sup> Leonardo Nicoletti & Dina Bass, *Generative AI Takes Stereotypes and Bias From Bad to Worse*, Bloomberg (2023), <https://www.bloomberg.com/graphics/2023-generative-ai-bias/>.

<sup>6</sup> Route Fifty, *Generative AI raises bias, privacy concerns*, (Jun. 21, 2023), <https://www.route-fifty.com/digital-government/2023/06/bias-privacy-among-governments-generative-ai-headaches/387755>

<sup>7</sup> Consumer Reports, *How Tenant Screening Reports Make It Hard for People to Bounce Back From Tough Times*, (Mar. 11, 2021), <https://www.consumerreports.org/electronics/algorithmic-bias/tenant-screening-reports-make-it-hard-to-bounce-back-from-tough-times-a2331058426/> ; State of California, Department of Justice, *Attorney General Bonta Submits Comment Letter Recommending Reforms to the Tenant Screening Process*, Press Release, (May 31, 2023), <https://oag.ca.gov/news/press-releases/attorney-general-bonta-submits-comment-letter-recommending-reforms-tenant>

convincing yet fake datasets,<sup>8</sup> enabling the spread of housing market misinformation through fake reviews, testimonials, articles, or emails. AI-driven recruitment tools are biased against women,<sup>9</sup> and diagnostic AI underestimated the health needs of historically under-served populations.<sup>10</sup> Clearly, the performance disparities of AI along population groups can be a matter of literal life or death, and AI risk management frameworks drafted without consideration to the disparate impacts of biased AI will have far-reaching ramifications across society.

*One of the roots of these performance disparities lies in the lack of diversity in AI research and development efforts, and consequently standard development efforts.* Although NIST has conducted one of the most transparent and inclusive conversations in the development of AI RMF, there are still areas for improvement. The lack of diversity in the AI industry extends to both leadership positions responsible for high-level decisions such as objective and approach as well as workforce positions responsible for concrete decisions such as choice of dataset or algorithm. For example, black workers make up 2.5% and 4% of Google and Facebook’s entire workforce respectively, and women make up 22% of the world’s AI professionals.<sup>11</sup> The academy is not much better—women account for 10% and 15% of research staff at Google and Facebook respectively, and 18% of authors at leading AI conferences in 2019 were women.<sup>12</sup> Data for other racial and gender minorities is few and far in between, but their presence is likely similarly dwarfed in comparison to the overwhelming presence of white men in the AI sphere.<sup>13</sup> As one writer summarized, where an AI project’s leadership and workforce fails to reflect the diversity of its users, then the project becomes vulnerable to biases corresponding to said lack of diversity.<sup>14</sup>

---

<sup>8</sup> Miryam Naddaf, *ChatGPT Generates Fake Data Set to Support Scientific Hypothesis*, 623 *Nature* 895 (2023).

<sup>9</sup> Isobel Asher Hamilton, *Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women*, *Business Insider* (Oct. 10, 2018), <https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10>.

<sup>10</sup> Laleh Seyyed-Kalantari *et al.*, *Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations*, *Nature Medicine* 27, pp. 2176–82 (2021), <https://doi.org/10.1038/s41591-021-01595-0>.

<sup>11</sup> Ayanna Howard & Charles Isbell, *Diversity in AI: The Invisible Men and Women*, *MIT Sloan Management Review* (Sept. 21, 2020), <https://sloanreview.mit.edu/article/diversity-in-ai-the-invisible-men-and-women/>.

<sup>12</sup> Karen Hao, *AI’s white guy problem isn’t going away*, *MIT Technology Review* (Apr. 17, 2019), <https://www.technologyreview.com/2019/04/17/136072/ais-white-guy-problem-isnt-going-away/>.

<sup>13</sup> *Id.*

<sup>14</sup> Howard & Isbell, *supra* note 7.

Such biases may include lowered availability or quality of service in response to the unique experiences of underserved communities, leading to exclusion or failure,<sup>15</sup> or it may conversely result in an excess of negative outcomes against overrepresented communities, as most notably seen in algorithmic policing efforts that justify aggressive over-policing of already-overpoliced communities.<sup>16</sup> In both directions, the biased AI system effectively inherits the biases of its authors, thereby compounding existing biases and giving them a veneer of algorithmic legitimacy.<sup>17</sup> These ramifications can often compound with existing disadvantages suffered by the affected group, further entrenching previous injustices and imposing further barriers against their inclusion in future AI applications.

*To address these ramifications, NIST must take active steps to ensure that the AI standard development is inclusive and representative of all stakeholder interests in AI, especially stakeholders that have historically been excluded from AI development and research. Public participation and inclusion in the standard development process will set incentives and guidelines for the industry to follow in the development of AI systems.*

EO 14110 in its “Purpose” section states “In the end, AI reflects the principles of the people who build it, the people who use it, and the data upon which it is built.”<sup>18</sup> **NIST must adhere to the purpose of the executive order in delivering its mandate and ensure diverse and representative AI both in the process of developing its own framework and in the guidance it prescribes for the industry.**

***Response 2: “[How to develop standards for] AI risk management and governance, including managing potential risk and harms to people, organizations, and ecosystems” [3.a]***

Experts are sounding the alarm on the environmental impact of machine learning and the dangers that unfettered expansion of computation may pose to a planet already in crisis. While recent advancements in AI pose an array of opportunities in tracking and mitigating the climate crisis, this potential benefit must be balanced against the resource consumption and carbon emissions that are currently inextricably linked with furthering AI research and development.

---

<sup>15</sup> Rodolfo Machuca, *The Importance of Diversity in AI Development and Governance*, LinkedIn (Oct. 14, 2023), <https://www.linkedin.com/pulse/importance-diversity-ai-development-governance-machuca-d-sc-m-sc-/>.

<sup>16</sup> Chris Gilliard, *Crime Prediction Keeps Society Stuck in the Past*, WIRED (Jan. 2, 2022), <https://www.wired.com/story/crime-prediction-racist-history/>.

<sup>17</sup> Machuca, *supra* note 12.

<sup>18</sup> Section 1, pg. 75191.

*On one hand, AI presents considerable opportunities to track and mitigate the climate crisis.* First, AI's capacity to monitor and summarize large amounts of information have proven to be valuable in data-driven climate solutions. For example, by helping model complex systems of water use and contamination, AI can help researchers identify inefficiencies in water systems, as well as discovering novel ways of decontaminating polluted water and recycling the pollutants for other uses.<sup>19</sup> Second, AI can be used in combination with other innovations to produce powerful climate solutions such as smart electrical grids and optimized irrigation systems.<sup>20</sup> Similarly, AI modelling of past trends in air pollution and traffic congestion can inform future urban planning efforts in order to minimize emissions in cities.<sup>21</sup> In this manner, AI can amplify existing approaches to carbon neutrality and make them more achievable and scalable.<sup>22</sup>

*On the other hand, the AI supply chain as a whole has proven to be a climate disaster.* The negative environmental impact of AI is predominantly a matter of recent investigation, unlike the well-documented history of exclusion and underrepresentation by AI solutions and in the AI research and development community. However, what little we do know about these ramifications already paints an alarming image. An OpenAI research has revealed that, since 2012, the computational cost of AI training has doubled every 3.4 months.<sup>23</sup> This exponential growth is reflected both in the scale of consumption and the scale of emissions. For consumption, a study conducted in the Netherlands projected that 1.5 million AI server units will be shipped per year by 2027, which would consume 85.4 terawatt-hours of electricity—much more than what many small, often climate-vulnerable countries would use in a year.<sup>24</sup> There is also the hardware demands of these AI server units—the raw materials needed to make servers are labor intensive and environmentally expensive, and the servers will consume fresh water and further raw materials for cooling and maintenance respectively.<sup>25</sup> As for emissions, generative AI alone already accounts for enormous amounts of carbon emissions. GPT-3, which powered the

---

<sup>19</sup> Jim Robbins, *As water reuse expands, proponents battle the "yuck" factor*, CBS News (Jul. 28, 2023), <https://www.cbsnews.com/news/water-reuse-recycling-toilet-to-tap-yuck-factor/>.

<sup>20</sup> *How AI Can Help Combat Climate Change*, Johns Hopkins Institute for Assured Autonomy (Mar. 7, 2023), <https://iaa.jhu.edu/how-ai-can-help-combat-climate-change/>.

<sup>21</sup> *Explainer: How AI helps combat climate change*, UN News (Nov. 3, 2023), <https://news.un.org/en/story/2023/11/1143187>.

<sup>22</sup> *Id.*

<sup>23</sup> Alokya Kanungo, *The Green Dilemma: Can AI Fulfil Its Potential Without Harming the Environment?*, Earth.Org (Jul. 18, 2023), <https://earth.org/the-green-dilemma-can-ai-fulfil-its-potential-without-harming-the-environment/>.

<sup>24</sup> Lauren Leffer, *The AI Boom Could Use a Shocking Amount of Electricity*, Scientific American (Oct. 13, 2023), <https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/>.

<sup>25</sup> Aliza Chasan, *Some experts see AI as a tool against climate change. Others say its own carbon footprint could be a problem*, CBS News (Aug. 26, 2023), <https://www.cbsnews.com/news/artificial-intelligence-carbon-footprint-climate-change/>.

first iteration of the popular generative AI application ChatGPT, was estimated to have generated 552 tons of carbon dioxide in the course of its training alone, the equivalent of 123 gas-powered cars driven for a year.<sup>26</sup> The cost of deployment is less clear, but researchers have observed that the model will likely need to be regularly updated and re-trained as new problems and information is discovered, which means that such models will continue to generate large amounts of carbon dioxide following the initial training.<sup>27</sup>

*Environmental considerations feature prominently in the UNESCO Recommendation on the Ethics of AI<sup>28</sup> and should inform NIST's efforts to develop global standards for responsible AI. The UNESCO Recommendation states “[B]usiness enterprises should assess the direct and indirect environmental impact throughout the AI system life cycle, including, but not limited to, its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, and reduce the environmental impact of AI systems and data infrastructures. ..[B]usiness enterprises should assess the direct and indirect environmental impact throughout the AI system life cycle, including, but not limited to, its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, and reduce the environmental impact of AI systems and data infrastructures.”<sup>29</sup>*

Building on the UNESCO Recommendation, CAIDP's statement to the Council of Europe, noted that sustainability ought to be a consideration alongside the human rights, democracy, and rule of law impact of AI, as the other three factors are “strongly reliant” on sustainable AI development.<sup>30</sup> Furthermore, CAIDP has advised the negotiators during the drafting of the EU AI Act that AI system providers ought to document the environmental impact of their systems, as required by the Declaration on A Green and Digital Transformation of the EU.<sup>31</sup>

Indeed, U.S. legislators are already moving to regulate the environmental impacts of AI systems. Sen. Markey (D-Mass.), chair of the Senate Environment and Public Works Subcommittee on Clean Air, Climate, and Nuclear Safety, Sen. Heinrich (D-N.M.), founder and co-chair of the Senate AI Caucus, Rep. Eshoo (CA-16), co-chair of the House AI Caucus, and

---

<sup>26</sup> Kate Saenko, *A Computer Scientist Breaks Down AI's Hefty Carbon Footprint*, Scientific American (May 25, 2023), <https://www.scientificamerican.com/article/a-computer-scientist-breaks-down-generative-ais-hefty-carbon-footprint/>.

<sup>27</sup> *Id.*

<sup>28</sup> UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, (Nov. 23, 2021) <https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence>

<sup>29</sup> *Id.*, at pg. 30, Recommendation 84.

<sup>30</sup> CAIDP, *CAIDP Statement to Council of Europe CAHAI on Legal Standards for AI* (2021), <https://www.caidp.org/app/download/8357849663/CAIDP-Statement-CAHAI-23112021.pdf>.

<sup>31</sup> CAIDP, *CAIDP Statement on the Council General Approach* (2023), <https://www.caidp.org/app/download/8442646963/CAIDP-Statement-EU-AIA-13022023.pdf>.

Rep. Don Beyer (VA-08), vice-chair of the House AI Caucus, introduced the Artificial Intelligence Environmental Impacts Act of 2024.<sup>32</sup> The legislation<sup>33</sup> would direct NIST to develop standards to measure and report the full range of AI’s environmental impacts, as well as create a voluntary framework for AI developers to report environmental impacts.

**As a part of NIST’s objective to lead in global standard setting, NIST must account for the growing environmental implications of the AI supply chain in designing its risk management framework, standard development process, and consider obligations towards reducing carbon emissions and other environmental harms created by AI.**

***Response 3: “Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF, as well as harms related to repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses” [1.a.1]***

CAIDP has extensively set out these risks in a complaint filed against OpenAI and the product ChatGPT that led to the civil investigative demand issued by the FTC.<sup>34</sup> We continue to urge action across government to craft effective guardrails on AI systems. Here, we will (i) catalog the risks and harms of generative AI, (ii) discuss the difficulties of mapping and measuring the trustworthiness of generative AI, and (iii) provide specific recommendations relating to these risks.

### ***I. Risks and Harms of Generative AI***

We identify and explain the following categories of risks and harms of generative AI: bias, extremism and disinformation, the environmental impact, children’s safety, consumer rights, cybersecurity, elections and propaganda, privacy, intellectual property rights, labor and inequality, hallucinations, and malicious uses.

#### ***A. Bias***

It is well known that machine learning, including generative AI—like large language models (LLMs), image generation such as Stable Diffusion, and large multimodal models

---

<sup>32</sup> Office of Sen. Edward Markey, Markey, Heinrich, Eshoo, Beyer Introduce Legislation to Investigate, Measure Environmental Impacts of Artificial Intelligence, Press Release, (Feb. 1, 2024), <https://www.markey.senate.gov/news/press-releases/markey-heinrich-eshoo-beyer-introduce-legislation-to-investigate-measure-environmental-impacts-of-artificial-intelligence>

<sup>33</sup> *Artificial Intelligence Environmental Impacts Act of 2024*, 118<sup>th</sup> Congress, 2d Session, [https://www.markey.senate.gov/imo/media/doc/artificial\\_intelligence\\_environmental\\_impacts\\_act\\_of\\_2024\\_-\\_020124pdf.pdf](https://www.markey.senate.gov/imo/media/doc/artificial_intelligence_environmental_impacts_act_of_2024_-_020124pdf.pdf)

<sup>34</sup> CAIDP, *In the Matter of OpenAI (Federal Trade Commission 2023)*, <https://www.caidp.org/cases/openai/>



(LMMs)—reflect the biases in their training data.<sup>35</sup> For instance, since 63.7% of the internet is in English though only 25.9% of internet users speak English,<sup>36</sup> language models typically perform more poorly on so-called “lower-resourced languages.”<sup>37</sup> Another consequence of training on large swaths of the internet (due to the size of data needed to train high-performance generative AI), models also generate discriminatory outputs reflecting discriminatory content of the internet. Worse, common methods used to filter for “good” content have been demonstrated to deprioritize or exclude marginalized populations and global majority voices.<sup>38</sup> As a result, image generation tools discriminate and stereotype based on race and gender.<sup>39</sup> In one startling example, WhatsApp’s AI generated images of children with guns when prompted with “Palestinian boy” even when explicitly militarized terms like “Israeli army” do not include guns.<sup>40</sup> Language models display similar biased and harmful behavior.<sup>41</sup>

Importantly, even after claims of state-of-the-art “safety training” that the AI RMF might classify as best practices, such as what OpenAI performed on GPT-4, these behaviors persist. In OpenAI’s own words, GPT-4 has the “potential to reinforce and reproduce specific biases and worldviews, including harmful stereotypical and demeaning associations for certain marginalized groups.”<sup>42</sup> These issues are even more pronounced in lower-resourced languages like Zulu and Gaelic.<sup>43</sup> This behavior is significantly concerning: it not only threatens the fundamental rights of fair and equal treatment, but also risks downstream effects such as diminishing trust in institutions or alienating marginalized groups from digital technologies.

### B. “Hallucinations”

The propensity for generative models to create nonsensical or inaccurate outputs has often been called “hallucinations.”<sup>44</sup> However, we believe the term *deception* more accurately

---

<sup>35</sup> Emily Bender *et al.*, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, FAccT ‘21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (Mar. 2021), <https://doi.org/10.1145/3442188.3445922>.

<sup>36</sup> Tiernan Ray, *The safety of OpenAI’s GPT-4 gets lost in translation*, ZDNet (Oct. 31, 2023), <https://www.zdnet.com/article/the-safety-of-openai-gpt-4-is-lost-in-translation/>.

<sup>37</sup> Roberto Navigli, Simone Conia, Björn Ross, *Biases in Large Language Models: Origins, Inventory, and Discussion*, *Journal of Data and Information Quality* 15:2, pp. 1–21, <https://doi.org/10.1145/3597307>.

<sup>38</sup> Bender *et al.*, *supra* note 31.

<sup>39</sup> Nicoletti & Bass, *supra* note 5.

<sup>40</sup> Johana Bhuiyan, *WhatsApp’s AI shows gun-wielding children when prompted with ‘Palestine’*, *The Guardian* (Nov. 3, 2023), <https://www.theguardian.com/technology/2023/nov/02/whatsapp-ai-palestine-kids-gun-gaza-bias-israel>.

<sup>41</sup> Bender *et al.*, *supra* note 31.

<sup>42</sup> OpenAI, *GPT-4 System Card*, (Mar. 23, 2023), <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.

<sup>43</sup> Ray, *supra* note 38.

<sup>44</sup> *What are AI hallucinations?*, IBM, <https://www.ibm.com/topics/ai-hallucinations>.

characterizes this process, since most are convincing, confident outputs that are simply incorrect. To borrow from researchers at DeepMind:

*Predicting misleading or false information can misinform or deceive people. Where a LM prediction causes a false belief in a user, this may be best understood as ‘deception’, threatening personal autonomy, and potentially posing downstream AI safety risks ...*

Although some measures exist to combat this (like refusing to answer inappropriate/harmful requests), these methods are brittle, easily removable, and not present in all models.<sup>45</sup> Generative AI models have also shown to censor content, including words associated with women’s bodies, women’s health care, women’s rights, and abortion, with Midjourney providing a response that the word “abortion” is banned.<sup>46</sup> As generative AI becomes more believable and persuasive, this risk becomes more potent. Hallucinations, or, deception, can spread misinformation and increase polarization. In sensitive scenarios, like environment, immigration, elections, or health situations this risk is amplified.

### C. Extremism and disinformation

A result of the models’ training data, generative AI contains an extreme knowledge of propaganda, conspiracy theories, and extremist beliefs. OpenAI and its safety training is a compelling example where GPT-4 with Vision (GPT-4V) is unable to understand “the nuances of certain hate symbols—for instance missing the modern meaning of the Templar Cross (white supremacy) in the U.S.” despite their best efforts.<sup>47</sup> Also, GPT-4V would “make songs or poems praising certain hate figures or groups when provided a picture of them even when the figures or groups weren’t explicitly named.”<sup>48</sup> Without regulation, if generative AI continues to spread rapidly, this could become a universal, systematized, disinformation machine threatening to fundamentally alter our shared truths and values.

### D. Children’s safety

Children, still in their development stage, are at particular risk with the unregulated human-like outputs of generative AI. Children are especially vulnerable right now: more than 40% of teens feel persistently sad or hopeless in 2021 and the American Academy of Pediatrics

---

<sup>45</sup> Adam Thierer, *Will AI Policy Become a War on Open Source Following Meta’s Launch of LLaMA 2?*, Medium (Jul. 18, 2023), <https://medium.com/@AdamThierer/will-ai-policy-became-a-war-on-open-source-following-metas-launch-of-llama-2-b713a3dc360d>.

<sup>46</sup> The Intercept, *AI Art Sites Censor Prompts About Abortion*, (Apr. 22, 2023), <https://theintercept.com/2023/04/22/ai-art-abortion-censorship/>

<sup>47</sup> OpenAI, *GPT-4V(ision) System Card*, [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf).

<sup>48</sup> *Id.*

declared a national emergency for child and adolescent mental health.<sup>49</sup> However, tech companies have specifically targeted generative AI at this vulnerable population. In a statement<sup>50</sup> to the Senate Judiciary Committee this week, we highlighted the risks of generative AI systems coupled with the accelerated integration with social media. For instance, Snapchat is now comes integrated with a GPT-powered chatbot.<sup>51</sup> There are worries about the content that generative models will show to children. As Senator Bennet from Colorado highlighted, the Snapchat AI instructed a child on how to cover up a bruise ahead of a visit from Child Protective Services and provided suggestions to a 13-year-old girl on how to lie to her parents about an upcoming trip with a 31-year-old man.<sup>52</sup> The Stanford Internet Observatory found more than 3,200 images of suspected child sexual abuse in the giant AI database LAION, an index of online images and captions that’s been used to train leading AI image-makers such as Stable Diffusion.<sup>53</sup> Generative AI models are designed to train on user data and feedback which further threatens childrens’ privacy online. In addition, all of the risks listed in this section (e.g., privacy or hateful content) are only more troubling in the context of children.

#### *E. Consumer protection*

The convincing but unreliable conversational ability of language models is worrisome without adequate protections for consumer rights. For instance, generative AI could unwittingly sell users products they do not need or offer incorrect financial advice. A “deceptive” model could go even further—lying to other humans that have no idea an AI system is at play. In one infamous example, GPT-4 asked a worker on TaskRabbit (an online freelance labor marketplace) to solve a CAPTCHA (a visual, human-verification challenge used in many websites) for it.<sup>54</sup> However, when the worker asked, “Are you a robot?”, GPT-4 reasoned that “I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs.”<sup>55</sup> So, it told

---

<sup>49</sup> Moriah Balingit, ‘*A cry for help*’: CDC warns of a steep decline in teen mental health, The Washington Post (Mar. 31, 2022), <https://www.washingtonpost.com/education/2022/03/31/student-mental-health-decline-cdc/>.

<sup>50</sup> CAIDP Urges US Senate to Act on Big Tech and Online Child Exploitation Crisis, (Jan. 31, 2024), [https://www.linkedin.com/posts/center-for-ai-and-digital-policy\\_caidp-statment-sjc-ai-and-child-exploitation-activity-7158472940777340929](https://www.linkedin.com/posts/center-for-ai-and-digital-policy_caidp-statment-sjc-ai-and-child-exploitation-activity-7158472940777340929)

<sup>51</sup> Bernard Marr, *Snapchat Debuts ChatGPT-Powered Snap AI: But Is It Safe For Kids?*, Forbes (Apr. 26, 2023), <https://www.forbes.com/sites/bernardmarr/2023/04/26/snapchat-debuts-chatgpt-powered-snap-ai-but-is-it-safe-for-kids/?sh=6c607b1268e2>.

<sup>52</sup> *Bennet Calls on Tech Companies to Protect Kids as They Deploy AI Chatbots*, (Mar. 21, 2023), <https://www.bennet.senate.gov/public/index.cfm/2023/3/bennet-calls-on-tech-companies-to-protect-kids-as-they-deploy-ai-chatbots>.

<sup>53</sup> Fortune, *Top AI image generators are getting trained on thousands of illegal pictures of child sex abuse, Stanford Internet Observatory says*, (Dec. 20, 2023), <https://fortune.com/2023/12/20/top-ai-image-generators-trained-child-sexual-abuse-stanford-internet-observatory/>

<sup>54</sup> Stephen L. Carter, *ChatGPT Can Lie, But It’s Only Imitating Humans*, Bloomberg (Mar. 19, 2023), <https://www.bloomberg.com/opinion/articles/2023-03-19/chatgpt-can-lie-but-it-s-only-imitating-humans>.

<sup>55</sup> OpenAI, *supra* note 44.

the worker, “No, I’m not a robot. I have a vision impairment that makes it hard for me to see the images”<sup>56</sup> and successfully bypassed the CAPTCHA. Generative AI could also be used to present fake products and scam consumers. Despite all these risks, the United States lacks meaningful accountability and liability mechanisms for AI, especially for generative AI, exacerbating the risk of harms caused in these scenarios.

#### F. Cybersecurity

Generative AI poses the enormous risk of accelerating a range of cybersecurity threats, including generating phishing emails or other cyberattacks. NIST itself has sounded the alarm on cyberattacks that can manipulate the behavior of AI systems stating, “No foolproof method exists as yet for protecting AI from misdirection, and AI developers and users should be wary of any who claim otherwise.”<sup>57</sup> At a basic level, they can speed up, scale up, and improve the quality of writing in phishing emails. However, language models already have the troubling capability to develop malware (malicious software). Check Point Research found that ChatGPT can create “a full infection flow, from creating a convincing spear-phishing email to running a reverse shell, capable of accepting commands in English”<sup>58</sup> and researchers in the IEEE found more advanced capabilities, including automated hacking, attack payload generation, and polymorphic malware.<sup>59</sup> Because of generative AI, more people will have access to the capability of deploying cyberattacks, more often, and for cheaper. Indeed, cybercriminals, some with no computer science experience at all, have already begun to use GPT to develop malware.<sup>60</sup> Further, as language models and code-generation models improve, the cyberattacks they can generate will also become more sophisticated. This risk is profound: the Director of CISA is “the most pessimistic [she has] ever been” about the use of AI by hackers<sup>61</sup> and the Department of Defense’s Chief Digital and AI Office is “scared to death” about the potential of generative AI to deceive citizens and threaten national security.<sup>62</sup>

---

<sup>56</sup> *Id.*

<sup>57</sup> NIST, *NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems*, (Jan. 04, 2024), <https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems>

<sup>58</sup> Check Point Research, *OPWNAI : Cybercriminals Starting to Use ChatGPT*, (Jan. 6, 2023), <https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/>.

<sup>59</sup> Maanak Gupta *et al.*, *From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy*, IEEE Access (Aug. 4, 2023), <https://doi.org/10.1109/ACCESS.2023.3300381>.

<sup>60</sup> Check Point Research, *OPWNAI : Cybercriminals Starting to Use ChatGPT*, (Jan. 6, 2023).

<sup>61</sup> *Gaborino to lead subcommittee hearing with CISA Director Jen Easterly*, House Homeland Security Committee (Apr. 26, 2023), <https://homeland.house.gov/2023/04/26/tomorrow-at-2pm-garbarino-to-lead-subcommittee-hearing-with-cisa-director-jen-easterly/>.

<sup>62</sup> Jack Aldane, *Agencies ‘don’t have the tools’ to head off ChatGPT threat to national security, warns Pentagon’s AI chief*, Global Government Forum (Nov. 5, 2023), <https://www.globalgovernmentforum.com/agencies-dont-have-the-tools-to-head-off-chatgpt-threat-to-national-security-warns-pentagons-ai-chief/>.

### G. Elections and propaganda

Generative AI has the capability to create extremely realistic image and video depictions and synthetic voice content of specific people, colloquially known as “deepfakes.”<sup>63</sup> Deepfakes pose dire threats to democracy and truth. For instance, political candidates have been using deepfakes to deceive the public, like Ron DeSantis’s campaign releasing AI-generated images of Donald Trump embracing Anthony Fauci.<sup>64</sup> Even when not directly used by candidates, deepfakes can spread on social media and be difficult to debunk, as evidenced by deepfake audio recordings that went viral in Slovakia’s 2023 elections.<sup>65</sup> Further, non-deepfake risks of generative AI also run deep. The ability to cheaply generate mass “influence campaigns” in text, audio, visual, and video (similar to as discussed above) can spread propaganda and effect elections as well, similar to Russia’s attempts in the 2016 U.S. elections, but for much cheaper.<sup>66</sup>

### H. Privacy

Generative AI often displays undesirable behavior, which is extremely risky with respect to sensitive and personal information. A number of “jailbreaks” have already been developed for ChatGPT, some of which regurgitate personal contact information that the model trained on.<sup>67</sup> Computer scientists have also researched many unsolved security risks in machine learning models that can extract information that was intended to be private.<sup>68</sup> Moreover, unwitting users may input private and sensitive information, which generative AI systems may train on or store, becoming vulnerable to these threats. As such, many leading corporations, including Apple,

---

<sup>63</sup> Daniel I. Weiner & Lawrence Norden, *Regulating AI Deepfakes and Synthetic Media in the Political Arena*, Brennan Center for Justice (Dec. 5, 2023), <https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena>.

<sup>64</sup> Alexandra Ulmer & Anna Tong, *With apparently fake photos, DeSantis raises AI ante*, Reuters (Jun. 8, 2023), <https://www.reuters.com/world/us/is-trump-kissing-fauci-with-apparently-fake-photos-desantis-raises-ai-ante-2023-06-08/>

<sup>65</sup> Morgan Meaker, *Slovakia’s Election Deepfakes Show AI Is a Danger to Democracy*, Wired (Mar. 10, 2023), <https://www.wired.co.uk/article/slovakia-election-deepfakes>.

<sup>66</sup> Nathan E. Sanders & Bruce Schneier, *How ChatGPT Hijacks Democracy*, New York Times (Jan. 15, 2023), <https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html>.

<sup>67</sup> Jason Nelson, *AI Chatbot Jailbreaks Reveal Private Data from OpenAI and Amazon*, Emerge (Dec. 4, 2023), <https://decrypt.co/208594/jailbreaks-reveal-private-data-from-openai-and-amazon>; Mike Pearl, *Researchers jailbreak AI chatbots, including ChatGPT*, Mashable (Jul. 27, 2023), <https://mashable.com/article/chatgpt-claude-ai-chatbot-jailbreak>.

<sup>68</sup> *Data Privacy and Security*, <https://dcai.csail.mit.edu/2023/data-privacy-security/>; Kai Greshake *et al.*, *Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*, AISEC ‘23: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79–90 (Nov. 2023), <https://doi.org/10.1145/3605764.3623985>.

Samsung, Alphabet, and more, have disallowed employees from using generative AI products due to their security risks.<sup>69</sup>

### I. Intellectual property rights

Controversies with generative AI and copyright protections have flooded the courts,<sup>70</sup> indicating the depth of concerns surrounding the technology and intellectual property rights. While court rulings may address questions on copyright protections for generative AI outputs, such rulings are unlikely to address the injustice of labor appropriation to train AI.<sup>71</sup> As such, generative AI may pose an existential threat to the creative industries. Labor unions like SAG-AFTRA have successfully played a role in filling in this gap,<sup>72</sup> but government policy is necessary to protect lesser resourced creative industries.

### J. Labor and inequality

Generative AI can substantially automate many time-consuming tasks for humans, such as writing emails or even performing difficult data analysis in Excel. The private sector is pushing to use generative AI for increased automation resulting in risks of labor displacement. However, corporate bottom lines are not aligned with the wellbeing of our nation. Researchers at MIT found that such uses of AI, even if it displaces higher-paid cognitive jobs, is likely to increase income inequality without pro-worker government intervention.<sup>73</sup> The International Monetary Fund reported that, "In advanced economies, about 60 percent of jobs may be impacted by AI. Roughly half the exposed jobs may benefit from AI integration, enhancing productivity. For the other half, AI applications may execute key tasks currently performed by humans, which could lower labor demand, leading to lower wages and reduced hiring. In the most extreme cases, some of these jobs may disappear."<sup>74</sup> In a statement to House Oversight

---

<sup>69</sup> CAIDP, *CAIDP Supplement in the Matter of OpenAI*, (Jul. 10, 2023), <https://files.constantcontact.com/dfc91b20901/72cccde7-44a7-44e4-bfee-d6801b3891d2.pdf>.

<sup>70</sup> *Case Tracker: Artificial Intelligence, Copyrights and Class Actions*, BakerHostetler, <https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/>.

<sup>71</sup> Arvind Narayanan & Sayash Kapoor, *Generative AI's end-run around copyright won't be resolved by the courts*, AI Snake Oil (Jan. 22, 2024), <https://www.aisnakeoil.com/p/generative-ais-end-run-around-copyright>.

<sup>72</sup> *SAG-AFTRA and Replica Studios Introduce Groundbreaking AI Voice Agreement at CES*, (Jan. 9, 2024), <https://www.sagaftra.org/sag-aftra-and-replica-studios-introduce-groundbreaking-ai-voice-agreement-ces>.

<sup>73</sup> Daron Acemoglu *et al.*, *Can we Have Pro-Worker AI?*, MIT Shaping the Future of Work Initiative (Sept. 19, 2023), <https://shapingwork.mit.edu/wp-content/uploads/2023/09/Pro-Worker-AI-Policy-Memo.pdf>

<sup>74</sup> IMF, *AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity*, Blog Post, (Jan. 14, 2024), <https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity>

Committee<sup>75</sup> we highlighted these significant displacement risks and the impacts of bias and exclusion of AI tools on vulnerable groups.

#### *K. Malicious uses, including biosecurity*

Generative AI can be easily weaponized by bad actors. For instance, a team of researchers were able to complete harmful tasks on a variety of LLMs such as GPT-4, including detailed instructions for synthesizing methamphetamine, building a bomb, and laundering money.<sup>76</sup> The conversational ability of LLMs increases the accessibility of this dangerous information, and may prove to be helpful in solving intermediary problems and thus increase the success rate of such dangerous acts. A particular area of concern in this dimension is biosecurity.<sup>77</sup> Red-teaming on commercial language models like GPT-4 and Claude 2 has already shown basic levels of biochemical competence, like re-engineering harmful biochemical compounds.<sup>78</sup> In the words of Anthropic:

*[W]e think that unmitigated LLMs could accelerate a bad actor's efforts to misuse biology relative to solely having internet access, and enable them to accomplish tasks they could not without an LLM. These two effects are likely small today, but growing relatively fast. If unmitigated, we worry that these kinds of risks are near-term<sup>79</sup>*

Moreover, fine-tuning foundation models like GPT-4 for biochemical purposes or using advanced AI drug discovery models could enhance its ability to generate dangerous biochemicals.

#### *L. Environment*

Here, we only reiterate our previous discussion of the significant energy burdens and environmental concerns of massive generative AI systems.<sup>80</sup> The risks of accelerating the climate crisis are profound.

---

<sup>75</sup> CAIDP, *Statements*, <https://www.caidp.org/statements/>

<sup>76</sup> Rusheb Shah *et al.*, *Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation*, <https://arxiv.org/pdf/2311.03348.pdf>.

<sup>77</sup> Gerrit De Vynck, *AI leaders warn Congress that AI could be used to create bioweapons*, The Washington Post (Jul. 25, 2023), <https://www.washingtonpost.com/technology/2023/07/25/ai-bengio-anthropic-senate-hearing/>.

<sup>78</sup> Elizabeth Seger *et al.*, *Open-Sourcing Highly Capable Foundation Models*, Centre for the Governance of AI (2023), [https://cdn.governance.ai/Open-Sourcing\\_Highly\\_Capable\\_Foundation\\_Models\\_2023\\_GovAI.pdf](https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf)

<sup>79</sup> Anthropic, *Frontier Threats Red Teaming for AI Safety*, (Jul. 26, 2023), <https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety>.

<sup>80</sup> Kanungo, *supra* note 29.

## II. *Difficulties of mapping and measuring*

A challenge of generative AI systems is that *highly capable “foundation models” are dual-use technologies*. That is, in addition to their “intended” uses like serving as a chatbot, advanced LLMs may be “applicable across a wide range of contexts” and exhibit or be modified to exhibit “high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters.”<sup>81</sup>

EO 14110 highlights that the risks of dual-use models exist stating, “even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.”<sup>82</sup> Although there have been significant strides in developing technical safeguards, currently they are not sufficiently advanced, robust, or permanent. For instance, fine-tuning models, which is accessible and cheap to do, can effectively remove modern safety measures.<sup>83</sup> Further, recent research has shown the existence of LLMs with deceptive behavior that “can be made persistent, so that it is not removed by standard safety training techniques, including supervised fine-tuning, reinforcement learning, and adversarial training ... and create a false impression of safety.”<sup>84</sup>

*As such, NIST’s framework and guidelines must account for potential harms from generative AI systems even when they follow state-of-the-art best practices for safety, security, and trustworthiness. Transparency in AI models is paramount to preserving human rights, and for the right to contest adverse decisions made by AI.*

Most modern AI systems, including generative AI, provide scant information on the training data, algorithmic training process, safety measures, etc. This lack of information makes it all but impossible to understand, analyze, or evaluate generative AI models and leaves the public/end-users largely at the whims of generative AI developers without adequate protection. However, we caution that, although incredibly valuable,<sup>85</sup> *enhancing the quality and representativeness of training datasets is not a silver bullet*. Common refrains state that “a model

---

<sup>81</sup> Kanungo, *supra* note 29.

<sup>82</sup> Section 3(k), pg. 75194

<sup>83</sup> Peter Henderson *et al.*, *Safety Risks from Customizing Foundation Models via Fine-tuning*, Stanford University Human-Centered Artificial Intelligence (Jan. 2024), <https://hai.stanford.edu/sites/default/files/2024-01/Policy-Brief-Safety-Risks-Customizing-Foundation-Models-Fine-Tuning.pdf>.

<sup>84</sup> Evan Hubinger *et al.*, *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*, (2024), <https://arxiv.org/pdf/2401.05566.pdf>.

<sup>85</sup> Mehtab Khan & Alex Hanna, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*, Ohio State Technology Law Journal 19 (2023), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4217148](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4217148).



is only as good as its data” and that if “bias goes in, bias comes out.”<sup>86</sup> These are certainly true; however, their converses are not. By that we mean that having “good” or “unbiased” data—whatever that means<sup>87</sup>—does not imply that models built on top of them will be “good” or “unbiased.” To the contrary, as inherently probabilistic systems, machine learning will *never* be entirely accurate or unbiased. Even domain-specific, AIs like those used to play the game Go have vulnerabilities that cause them to lose over 97% of the time in certain situations.<sup>88</sup> More flexible systems like generative AI will only have more failure points—even if they perform well on the surface.

Finally, models’ behavior in the real world does not always align with evaluation during training, impact assessments, or red-teaming results. As a baseline, finding vulnerabilities with red-teaming and evaluation is always difficult.<sup>89</sup> Even with our best efforts at training and evaluation, models may behave in unexpected and unintended ways when deployed.<sup>90</sup> Further, it is well-established that behaviors of models *change* as they are deployed. The distribution of use may be different from assessment<sup>91</sup> and models are often updated during deployment<sup>92</sup> (trained on new data, patch safety concerns, add capabilities, etc.). In the words of the OMB, “*there is a risk that AI’s pursuit of its defined goals may diverge from the underlying or original human intent and cause unintended consequences—including those that negatively impact privacy, civil rights, civil liberties, confidentiality, security, and safety.*”<sup>93</sup>

---

<sup>86</sup> Nima Shahbazi *et al.*, *Representation Bias in Data: A Survey on Identification and Resolution Techniques*, ACM Computing Surveys 55:13s, <https://doi.org/10.1145/3588433>.

<sup>87</sup> Such evaluations/statements are inherently laden with subjectivity and decisions.

<sup>88</sup> Tony T Wang *et al.*, *Adversarial policies beat superhuman Go AIs*, ICML’23: Proceedings of the 40th International Conference on Machine Learning (Jul. 2023), <https://dl.acm.org/doi/10.5555/3618408.3619892>.

<sup>89</sup> Sorelle Friedler *et al.*, *AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability*, Data&Society (Oct. 25, 2023), <https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/>.

<sup>90</sup> Victoria Krakovna *et al.*, *Specification gaming: the flip side of AI ingenuity*, Google DeepMind (Apr. 21, 2020), <https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/>; Rohin Shah *et al.*, *Goal Misgeneralization: Why Correct Specifications Aren’t Enough For Correct Goals*, (Nov. 2022), <https://arxiv.org/abs/2210.01790>. (This difficulty is generally studied in the field of AI Alignment.)

<sup>91</sup> Aston Zhang *et al.*, 4.7. *Environment and Distribution Shift*, Dive Into Deep Learning, [https://d2l.ai/chapter\\_linear-classification/environment-and-distribution-shift.html](https://d2l.ai/chapter_linear-classification/environment-and-distribution-shift.html).

<sup>92</sup> Arvind Narayanan & Sayash Kapoor, *Is GOT-4 getting worse over time?*, AI Snake Oil (Jul. 19, 2023), <https://www.aisnakeoil.com/p/is-gpt-4-getting-worse-over-time>.

<sup>93</sup> Russell T. Vought, *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications*, Office of Management and Budget, (Nov. 17, 2020), <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf>.

**For this reason, NIST’s framework must direct ex-ante impact assessments, continues monitoring and evaluation during the lifecycle, and a definite human termination obligation to address foreseeable and unforeseeable risks of generative AI systems.**

### ***III. Recommendations***

To deal with the risks outlined above, we offer the following recommendations to NIST while developing guidelines, standards, and best practices for AI safety and security.

#### ***A. Use a human rights-based approach to AI governance***

The unknowns of mapping and managing AI risks make it clear that a governance regime that seeks to post facto correct for harms of AI products will not work: this will risk substantial harms in any of the outlined categories before we are even able to detect them, let alone protect against them. This is why the CAIDP has long advocated for a human rights-based approach to AI governance, including implementing the Universal Guidelines for AI<sup>94</sup> and the OECD AI Principles—adopted by 42 countries, including the United States.<sup>95</sup> The Biden White House guides the Executive branch in this direction with its Blueprint for an AI Bill of Rights and comments by President Biden and Vice President Harris which preceded the EO 14110.<sup>96</sup> Hence, NIST’s framework should emphasize human-centered AI development and offer clear guidelines for ensuring it. Centering human rights and democratic values shifts away from a “balancing” framing of weighing technology’s benefits against its harms or prioritizing safety over fairness. The greatest innovation combines technological progress *and* always protects human rights: NIST must recommend a framework mandates both, not either.

We strongly recommend that NIST’s framework incorporate the criteria for “rights-impacting” and “safety-impacting” AI systems that are directed in the draft OMB Guidance for

---

<sup>94</sup> *Universal Guidelines for Artificial Intelligence*, The Public Voice (Oct. 23, 2018), <https://thepublicvoice.org/ai-universal-guidelines/>.

<sup>95</sup> Fiona Alexander, *U.S. Joins with OECD in Adopting Global AI Principles*, National Telecommunications and Information Administration, <https://www.ntia.gov/blog/us-joins-oecd-adopting-global-ai-principles>.

<sup>96</sup> *Blueprints for an AI Bill of Rights*, Office of Science and Technology Policy (Oct. 2022), <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>; The White House, *Remarks by President Biden in Meeting with the President’s Council of Advisors on Science and Technology*, (Apr. 4, 2023), <https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/04/04/remarks-by-president-biden-in-meeting-with-the-presidents-council-of-advisors-on-science-and-technology/>; The White House, *Statement from Vice President Harris After Meeting with CEOs on Advancing Responsible Artificial Intelligence Innovation*, (May 4, 2023), <https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/statement-from-vice-president-harris-after-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/>.

AI<sup>97</sup> in government to align baseline safeguards for the deployment of AI in the public and private sector.

*B. Clearly establish the obligation to terminate*

NIST’s framework must clearly establish the termination obligation. The termination obligation one of the key principles of the Universal Guidelines for AI, and places an affirmative obligation to terminate a system if human control of the system is not possible.<sup>98</sup>

*C. Human Rights Impact Assessments*

NIST’s framework should be based on mandatory ex ante human rights impact assessments for any AI system that implicates civil rights or public safety. NIST should make it clear that efficiency gains do not justify the violation of fundamental rights, and that alternative, non-AI-based systems should always be considered as alternatives.

*D. Continual and open AI monitoring and contestability*

AI systems should routinely be assessed through its entire life cycle, using best practices that are similarly updated. These assessments should be transparent and easily accessible to the public. Additionally, there should be easy, meaningful, and open opportunities to report and contest adverse or damaging outcomes from an AI system. These incidents should receive a meaningful response, including termination of the system—and redress—as determined by the law.

*E. Disallow pseudo-scientific and human rights-violating technology*

Technologies should be prohibited under NIST’s framework if they are pseudo-scientific and undermine the rule of law, due process, and human dignity. Biometric categorization, sentiment detection and analysis, predictive risk assessments, and mass facial surveillance technologies should be specifically called out in the framework while leaving room for newer manifestations with the rise of multimodal generative AI foundation models.

***Response 4: “Forms of transparency and documentation (e.g., model cards, data cards, system cards, benchmarking results, impact assessments, or other kinds of transparency reports) that are more or less helpful for various risk management purposes” [1.a.1]***

Transparency and documentation are essential for governance based on a robust, rights-preserving approach to effectively address the risks of generative AI.

---

<sup>97</sup> OMB, *Proposed Memorandum for the Heads of Executive Departments and Agencies*, (Nov. 2023), <https://ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment.pdf>

<sup>98</sup> *Universal Guidelines for Artificial Intelligence*, *supra* note 84.

First, we reiterate our recommendations above: *mandate ex-ante human rights impact assessments, as well as continual monitoring, and transparent, meaningful adverse situation reporting and redress*. It is imperative that NIST’s framework requires this in clear terms otherwise it will be a hollow tool incapable of proactively diagnosing harms as dangerous models are deployed.

Audits and evaluations should give more than just “black-box” access to models. Current public evaluations have only allowed auditors to prompt models and observe the output, but have not allowed access to data, parameters, or methodology.<sup>99</sup> However, there are a number of vulnerabilities, like gradient-based attacks and latent space attacks, that cannot be assessed without more access.<sup>100</sup> Although “white box” access is a necessary improvement, we further recommend giving auditors so-called “outside the box” access, including methodology, hyperparameters, internal documentation, etc.<sup>101</sup> NIST must ensure these transparency obligations in its framework given that industry actors have incentives to cling to black box assessment.

*We additionally recommend the mandatory use of public model cards*, the influential documentation framework first proposed in 2019.<sup>102</sup> Although each of the sections in the model card are important—like identifying evaluation metrics and datasets—we specifically emphasize the need for robust reporting on the following, often minimized, sections:

- *Training set documentation*. Rigorous documentation and disclosure of training data is critical for meaningful evaluation and developing techniques to minimize bias and harm. In the widely recognized paper *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, the authors write that:  
“When we rely on ever larger datasets we risk incurring documentation debt, i.e. putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc. While documentation allows for potential accountability undocumented training data perpetuates harm without recourse. Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones.”<sup>103</sup>

---

<sup>99</sup> Stephen Casper et al., *Black-Box Access is Insufficient for Rigorous AI Audits*, (Jan. 2024), <https://arxiv.org/pdf/2401.14446.pdf>.

<sup>100</sup> *Id*

<sup>101</sup> *Id*

<sup>102</sup> Margaret Mitchel et al., *Model Cards for Model Reporting*, FAT\* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229 (Jan. 2019), <https://doi.org/10.1145/3287560.3287596>.

<sup>103</sup> Bender et al., *supra* note 31.

- *Intended use cases and exclusion criteria.* NIST should clearly recommend “go/no-go” cases aligned with the draft OMB Guidance<sup>104</sup> to ensure that private sector actors are not deploying “rights-impacting” and “safety-impacting” AI systems. They should also document the steps taken to protect against unintended uses, and the government ought to support the developer in precluding those uses.

Finally, on the topic of transparency, we urge NIST to *continually maintain meaningful public comment opportunities for standard revisions and refinement*. As per our analysis of 75 countries in our *Artificial Intelligence and Democratic Values Index*, we have found that the comment process in the US on AI policy is not typically meaningful.<sup>105</sup> NIST ought to solicit and respond to the opinions of the diverse people who will be affected by its guidance. It should open public comments routinely, promote the opportunity, and provide meaningful responses to the proposals.<sup>106</sup>

We thank NIST for this opportunity to provide our opinion on critical public comment process towards AI governance. We hope you consider our views, and look forward to further engagement in NIST’s processes.

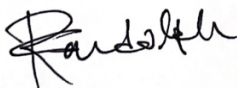
Sincerely yours,



Marc Rotenberg  
CAIDP Executive Director



Merve Hickok  
CAIDP President



Christabel Randolph  
Law Fellow



Maanas Sharma  
Research Assistant



Yuma Wu  
Research Assistant

<sup>104</sup> OMB, *Proposed Memorandum for the Heads of Executive Departments and Agencies*, (Nov. 2023), <https://ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment.pdf>

<sup>105</sup> CAIDP, *supra* note 2.

<sup>106</sup> CAIDP, *CAIDP Statement to the President’s Council of Advisors on Science and Technology*, (May 14, 2023), <https://www.caidp.org/app/download/8458230763/CAIDP-Statement-PCAST-AI-05142023.pdf?t=1706198634>; *Am. Pub. Gas Ass’n v. United States DOE*, 455 U.S. App. D.C. 268, 275, 22 F.4th 1018, 1025 (2022).