UNIDOSUS

STRONGER COMMUNITIES. STRONGER AMERICA.

December 5, 2023

Clare Martorana, U.S. Federal Chief Information Officer
Office of the Federal Chief Information Officer
Office of Management and Budget
725 17th Street NW, Suite 50001
Washington, DC 20503

Submitted electronically via www.regulations.gov

Re: Request for Public Comment on Draft Memorandum—Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence (AI); 2023-24269

Dear Ms. Martorana and to whom it may concern,

On behalf of UnidosUS, we respectfully submit the comments below in response to the Office of Management and Budget's (OMB) Request for Comments on *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum* (hereinafter, "Memo," "OMB Memo," or "draft policy").

UnidosUS is a nonprofit, nonpartisan organization that serves as the nation's largest Hispanic civil rights and advocacy organization. Since 1968, we have challenged the social, economic, and political barriers that affect Latinos through our unique combination of expert research, advocacy, programs, and an Affiliate Network* of nearly 300 community-based organizations across the United States and Puerto Rico. Effective government (including the administration and equitable distribution of goods and services) is vital to the socio-economic advancement of Latinos, who face multiple barriers to opportunity. In sum, this issue is central to our ability to achieve our mission.

As we described at the first AI Insight Forum in September, while AI systems hold enormous promise, they also pose potential threats to civil rights and our democracy absent ethical, constitutionally designed governance. Years of experience with automated and algorithmic decision-making in domains such as credit, housing, and employment demonstrate that unaccountable and opaque uses of technologies can hide bias from view. Because AI capabilities and uses are racing ahead, we require a new social compact between technology and democracy.

UnidosUS proposes a vision for AI that reflects shared **values**, honors a wide range of **voices** in the process, and includes substantial **investment** in the shared benefits of AI. Our three main points—***Values, Voice, and Investment***—are pillars of an interrelated approach to ethical governance. Establishing strong democratic guardrails provides a foundation to realize AI's benefits responsibly. Ensuring concrete means of accountability for impacted communities can drive socially productive innovation, channeling competition to keep human impacts at the forefront of tech development. And

* UnidosUS Affiliate Network, https://www.unidosus.org/about/affiliates/.

achieving meaningful participation requires investment in the capacity of impacted communities and their needs.

In keeping with this framework, below we address the following topics:

- ***The Promise and Perils of AI for Latinos and Communities of Color***
    - *The OMB Memo is a Welcome Step, Yet More Clarity Is Needed to Ensure that Agencies Fully Address Challenges and Gaps in AI Governance*
- ***First Pillar: Values—The Right Decision Matrix Would Fully Acknowledge the Current Limitations of AI and Algorithmic Decisions and Integrate Values with Oversight***
    - *Exempting AI Uses at the Heart of Constitutional Governance Undermines Democratic Norms and Undermine Incentives to Develop Technologies that Are Rights- and Privacy-Enhancing*
- ***Second Pillar: Voice—Impacted Communities Deserve a Powerful Seat at the Table***
    - *Seizing the Opportunity to Democratize AI Governance: An Interactive Model*
        - *A Public Dashboard Would Facilitate Timely Monitoring of AI Progress, Impacts, and Risks*
        - *An Archive of Model Cards Would Support Understanding of Shifts in Model Performance*
        - *Creation of AI Ethics and Impacts Advisory Committees Would Inform and Help Coordinate Agency Efforts*
        - *Inclusive Red Teaming Can Surface Assumptions and Overlooked Risks and Harms*
        - *Leveraging Impact Assessments for Empirical Evidence to Inform Metrics*
    - *Specific Comments Highlight a Need for Additional Direction in Some Areas*
- ***Third Pillar: Investment—Growing an Ecosystem for Equitable Participation, Public Trust, and Innovation Insights***
- ***Conclusion: From Principles to Implementation: Creating an Ethical AI Ecosystem***

The Promise and Perils of AI for Latinos and Communities of Color

Inclusive, ethical AI systems could expand opportunities for Latinos in areas like education, language access, and employment. Personalized learning software, translation tools that break down language barriers, and job matching platforms that support diverse candidates are concrete examples of how AI could support disadvantaged groups and build skills.

Yet the millions of Latinos we represent face multiple risks from the proliferation of automated decision-making and AI systems as well as exclusion from its promise. Gaps in digital access and tech education exclude many Latinos from emerging economic opportunities in technology. Some 35% of Latinos lack home broadband, limiting the development of skills needed for upwardly mobile tech jobs—roles in which Latinos are already severely underrepresented. Latinos are only 8% of the STEM workforce, yet they will be almost 78% of all new workers by 2030.

The evidence on harms is clear. Opaque automated decision models used in high-stakes decisions about matters of criminal justice, lending, and benefits, can entrench historical biases and discrimination, yet both transparency and recourse are limited or nonexistent. Lending and credit access algorithmic discrimination persists despite legal authorities that require lenders to use the "least discriminatory alternative" and bar discrimination. As we saw with pulse oximeters during the pandemic, medical

algorithms and design can also perpetuate health disparities if trained on unrepresentative datasets. To make matters worse, electoral misinformation targeting Latinos mischaracterizes the positions of political parties and candidates, and misinformation in Spanish is permitted to stay online far more than similar statements in English.

Of course, administration of public functions and government programs has been characterized by a long legacy of racial and social inequality. Under the Executive Order on racial equity, agencies were directed by the Biden-Harris Administration to attend to this history. Yet it is not simple to address, even for pandemic-era social and economic supports. Working closely with the Administration, we learned that it required persistence and intention to close gaps in the Child Tax Credit or COVID vaccination rates.

Integration of AI systems into government could eliminate these types of barriers or it could create new, and potentially even more problematic, issues. Even in areas of strength for natural language AI systems, such as translation, because of the digital divide, much of the Internet (and thus, training data for models) is in English, and language barriers and differences in the quality of translations, may persist.

Latinos and other communities of color are also subjected to expansive governmental surveillance technologies. Predictive policing tools trained on flawed crime statistics have been found to disproportionately target low-income neighborhoods of color by falsely correlating race with criminality. Similarly, sentencing algorithms drawing on racially skewed conviction data likewise entrench harsher outcomes for minorities. While constitutional principles like due process, equal protection, and privacy notionally underpin our laws, outdated regulations are failing, even today, to provide adequate accountability for rights-infringing uses of AI and decision models.

Critical tools to alleviate worker displacement, involve impacted communities in decisions on technology, and meaningfully advance equity are underdeveloped. Given its reach and power, and the rush to market powerful products, to deploy AI responsibly and ethically will require new and innovative forms of governance. Systems should anticipate potential harms and include mechanisms for accountability to people they impact—including workers, creators, communities of color and lower-income people, and others left behind and left out by traditional research and the digital divide (and who are thus invisible to the models). Too often, the bias or flaws in models are understood too late—so we must get better at both predicting and preventing foreseeable harms by design, and impacted groups are ideally positioned to tell technologists what they may not know they do not know.

With enough intention, we can choose to govern AI and algorithms in ways that align with our values. Ultimately, the choice that is often posed, for example, between data security and privacy, on the one hand, and effective law enforcement, on the other, is a false one—good privacy by design can make both a reality once appropriate incentives and protections are in place. And the same is true for other aspects of AI systems that undermine fairness and shared values—we can and must ensure that standards for new technology that are fair, transparent, accountable, and explainable are developed with input from impacted communities.

The OMB Memo is a Welcome Step, Yet More Clarity Is Needed to Ensure that Agencies Fully Address Challenges and Gaps in AI Governance

The OMB Memo usefully builds on the ground-breaking and wide-ranging AI Executive Order (EO) work by the National Institute of Standards and Technology (NIST), including its seminal Risk Management Framework (RMF), the AI Bill of Rights from the White House, Office of Science and Technology Policy (OSTP), National AI Advisory Committee (NAAIC), and other Administration and agency policy and legal positions on technology generally and AI and algorithms more specifically.

We deeply appreciate the leadership from the Biden-Harris Administration on AI policy for the federal government and procurement. Both the OMB Memo and Executive Order acknowledge the risks of bias, unfairness, and use cases such as predictive policing methods, identifying them correctly as rights-implicating. Both also call for consultation with impacted communities.

In our view, the draft policy would be improved by the addition of instruction and detail on requirements for mechanisms to facilitate accountability, participation, and inclusive evaluations by impacted groups. For this reason, our comments propose to tighten the connections between these concepts, outlining ways in which AI technologies should be evaluated to standardize approaches. We also describe how agencies can embrace roles for impacted communities in policy formation and generate evidence on empirical and real-world impacts.

The National Institute of Standards and Technology (NIST) is assigned several important responsibilities by the EO, including creation of the AI Safety Institute Consortium, which will play a central role in developing tools to measure and improve AI safety and trustworthiness. The EO provides several other directives for NIST, including, among many other duties to:

- Develop a companion resource to the AI Risk Management Framework (AI RMF) focused on generative AI;
- Develop guidance on authenticating content created by humans and watermarking AI-generated content;
- Create guidelines for agencies to evaluate the efficacy of differential-privacy-guarantee protections, including for AI;
- Launch a new initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities that could cause harm; and
- Establish guidelines and [] to enable developers of generative AI, especially dual-use foundation models, to conduct AI red-teaming tests for deployment of safe, secure, and trustworthy systems.

For each of these responsibilities, NIST should include structured and intentional forms of engagements with impacted communities as a proof-of-concept for how AI standards can benefit from diverse input at each stage of the process. Similarly, OMB and federal agencies should develop formal processes to inform and engage impacted communities, improving the standards while building technical capacity and expertise for civil society and impacted groups. These formalized roles and processes should continue even after agencies' initial responsibilities under the EO are complete.

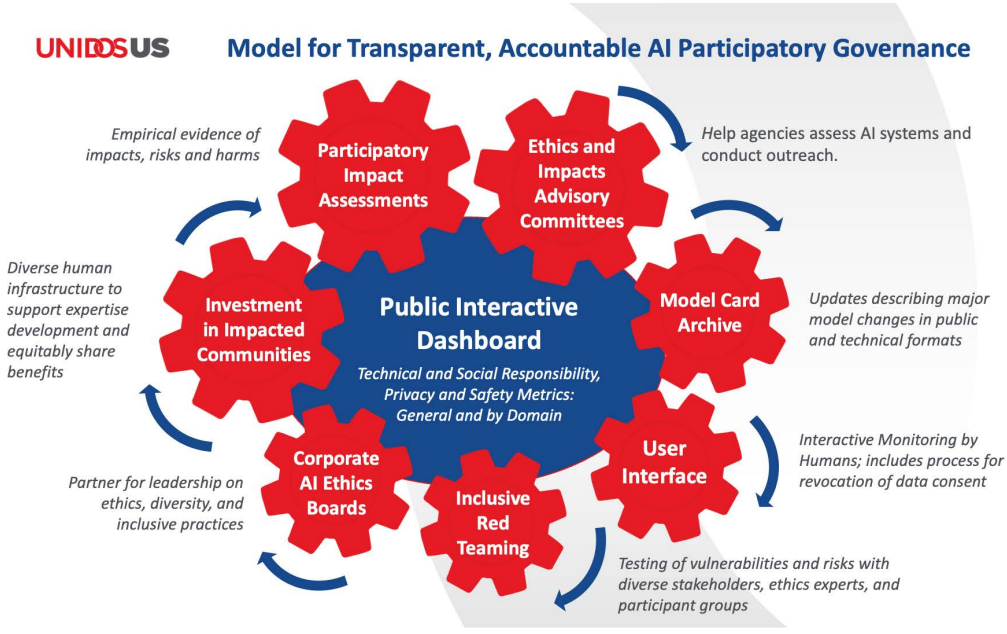
Creating a system of democratized governance to shape standards and systems is also an imperative to earn public trust. An insightful new paper notes that participatory design methods are increasingly at the forefront of AI research and exploration:

> Community-based participatory design is an approach to designing computing technologies with and for different publics, with the aim of forming more equitable relationships between algorithmic systems and often-marginalized publics. [] Computing systems are rarely developed entirely by the publics they serve; and in this way, participatory design is a situated practice of future-making, through which heterogeneous communities collaboratively imagine new sociotechnical futures. While participatory design has a long tradition in shaping the design of computing systems, it has more recently become a means to co-create artificial intelligence (AI) transparency and accountability artifacts, such as model cards, design workbooks, and user agreements. [*Citations omitted.*]

The authors envision five dimensions for the participatory design of user agreements. Applying a comparable vision to the present context, these could be translated as a call for:

1) participatory development of performance standards for models;
2) structures within model designs that anticipate and defend against potential harms;
3) opportunities to provide and revoke informed consent;
4) complaint mechanisms for harms when they occur and a means of redress;
5) disclosures and labeling of limitations and performance; and
6) external information gathering about potential and actual harms to drive iteration on standards.

As described herein, a regulatory ecosystem that fully connects principles in the OMB Memo to participatory design mechanisms that can be informed, iterated upon, and enforced by impacted communities is an essential next step in a democratic vision for AI governance. OMB and NIST can, as appropriate, guide creation of this ecosystem. The model's "gears" are intended to work in conjunction with one another and most (not all) are discussed in detail below.



First Pillar: Values—The Right Decision Matrix Would Fully Acknowledge the Current Limitations of AI and Algorithmic Decisions and Integrate Values with Oversight

The OMB Memo makes an essential contribution in that it will create leadership and structure at the agency level for oversight of AI uses. It designates an official leader and internal committee responsible for cross-coordination and agency leadership on AI, classifies safety- and rights- implicating uses, and requires steps to mitigate these potential risks and harms. To date, both narrow and general model AI adoption by agencies, including longstanding uses of privacy-intrusive data collection and reliance on automated decision-making algorithms for benefits and eligibility determinations, has been *ad hoc* and more-or-less ungoverned by shared principles or safeguards.

But a decision matrix for government use of AI cannot begin with the assumption that automation or other AI capabilities are necessary, fair or relevant. OMB should require agencies to grapple with, rather than gloss over, the serious question of whether an AI system is the best way to accomplish a particular federal function, when the tools may not be a good substitute for decision-making that currently resides with public officials.

There is a very human temptation to be attracted to new tools, but to set aside their limitations, ignore their under-developed state for specific uses, or assume humans will understand how to best apply them. The OMB Memo's definitions section usefully describes this as "automation bias: [t]he propensity for humans to inordinately favor suggestions from automated decision-making systems and to ignore or fail to seek out contradictory information made without automation."

A highly complex system, like AI, that confidently presents analytics as though they are based in fact and embedded with shared values—when they may not be at all—poses a sharp risk for human overestimation and overreliance due to this bias and the false (or real but problematic) efficiencies AI may offer. Given the novelty of the most advanced tools and the history of problematic applications of existing decision-making models, government agencies should demonstrate their safety and fidelity to NIST standards as a first step in governance.

Under the OMB guidance, agency leadership wields considerable discretion in determining what risks matter and how they will be mitigated, and more structure and guidance on appropriate responses to risks and mitigations will certainly be needed. The Memo's lack of clarity on these points raises the possibility that problems will not be acknowledged or adequately addressed.

Moving quickly to achieve transparency, accountability, identify limitations, and map use cases across the government is an excellent idea. But pushing agencies to rapidly integrate poorly tested and evaluated tools for more uses of AI in government, given the need for basic research on transparency, accountability, and other standards, as well as for a delineated process for public input and specific requirements for engagement with impacted communities, is not. We concur with the comments of The Leadership Conference on Civil Rights that "only AI that is proven to be equitable should be developed, acquired, or used," and would urge that forbearance to be exercised across all of the factors identified in the NIST RMF.

Instead of caution, however, throughout the OMB Memo, the agencies are generally encouraged to adopt new AI systems and build human resources to support it, in subtle and not-so-subtle ways. For just a few examples, p. 8 directs that "[a]gencies must improve their ability to use AI in ways that benefit

the public and increase mission effectiveness," while recognizing its limitations. On p. 9, the text notes that "[e]mbracing innovation requires removing unnecessary and unhelpful barriers to the use of AI."

We are concerned that the breathlessness of the conversation around AI and U.S. competitiveness will encourage agency leadership to extend or deepen uses of AI regardless of the substantial—and in some cases—insurmountable problems with specific uses by governments. As Elham Tabassi, the head of NIST's Trustworthy and Responsible AI Program, noted in her remarks at the recent NIST workshop, the measurement science for AI tools is very underdeveloped at the moment. The NIST RMF is also clear that, with regard to AI's risks and capacities, we are in an evolving situation:

> While there are myriad standards and best practices to help organizations mitigate the risks of traditional software or information-based systems, the risks posed by AI systems are in many ways unique. [] AI systems, for example, may be trained on data that can change over time, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to understand. AI systems and the contexts in which they are deployed are frequently complex, making it difficult to detect and respond to failures when they occur. AI systems are inherently sociotechnical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed. These risks make AI a uniquely challenging technology to deploy and utilize both for organizations and within society. Without proper controls, AI systems can amplify, perpetuate, or exacerbate inequitable or undesirable outcomes for individuals and communities. With proper controls, AI systems can mitigate and manage inequitable outcomes.

The task of the OMB Memo for the agencies is to establish "proper controls" over government uses of AI for current and near-future models and uses. We believe the Memo is a solid start, but that its approach is incomplete or lacks important clarity in a number of areas that could benefit from substantially more operational structure for agencies, and that OMB should more fully leverage the work of NIST.

For example, the agencies' assignment under the Memo to achieve "maturity" for AI systems begs the question of how—and who—defines that success and on what grounds. Agencies will need constructive guidance on common technical issues arising from current uses and mitigations for AI systems—and to be informed about helpful developments and technical and sociotechnical challenges that arise in particular contexts and use cases. A robust cross-government conversation and exchange of information will be essential as officials seek to document and more fully understand and communicate about the current state-of-play and its challenges.

Specifically, additional clarity and detailed instruction on how to align success for the "maturity" of AI systems with the requirements and standards from the NIST RMF, alongside new requirements for consultation with impacted communities, as we outline below, would provide a much more developed set of parameters for "success" and could help to generate much more transparent and aligned processes across the government.

In addition, the agencies' Use Inventories and proposed risk management approaches could usefully be organized according to the "AI Risks and Trustworthiness" issues described by NIST, which highlight that AI systems should meet baselines for each of the following factors: 1) Valid and Reliable; 2) Safe; 3)

Secure and Resilient; 4) Accountable and Transparent; 5) Explainable and Interpretable; 6) Privacy-Enhanced; and 7) Fair—with Harmful Bias Managed.

Agencies should evaluate current uses in light of each of these values across their entire portfolio of AI uses, in consultation with NIST and other experts familiar with the evolving science for each of these measures, paying concentrated attention, as the OMB Memo indicates, to risks and safety- and rights-impacting uses. In mapping current uses of AI and algorithmic tools, agencies should also:

- detail and explain the technological limitations of a tool given its use cases and relevant human factors,
- identify the adequacy of any current evaluations of the training data, model design, and impacts, and any mitigations for known and potential risks,
- describe the extent of involvement or consultation with impacted communities (more on this below) on design, risks, impacts, or other aspects of the model or system,
- explain the adequacy and conclusions of external audits and impact assessments that are underway or have been done, and
- fully characterize the socio-technical context at the agency related to human interactions with the technology, evidence on experiences of internal and external users, and other factors.

There are sound reasons for the agencies to take a far closer look at current uses and the lessons those offer before rushing to adopt new ones. Successful systems are more difficult than appears at first glance to build and execute. For example, take the RMF's assignment to make systems, "Fair—with Bias Managed." Goals like achieving a "fair" model, which seems simple enough, can, for technical reasons, sometimes be in tension with the accuracy of an AI model, as Brian Christian explains in his excellent book, The Alignment Problem, in which he provides specific examples of researchers' efforts to grapple with algorithmic bias in parole decision-making.

Perhaps it is for this reason, among others, that NIST's RMF notes that fairness can be a contested, situational, and difficult factor to satisfy—yet fairness along multiple dimensions is an absolutely critical factor to get right in government decisions, both legally and morally. Such decisions will not, therefore, be a function of math alone—human judgment and democratic input, as well as transparency about the tradeoffs to the extent they exist, will be necessary.

Christian also highlights a problematic but intrinsic feature of most AI models and the data sets on which they are trained. Because minorities are, by definition, less represented in data sets, the AI model has less information about these group that it does about a given majority. Further, AI models can draw subtle inferences from data in myriad ways that the how things are now is the way *they should be* in the future—in effect, mistaking what is distributionally true for what is morally or ethically true about human difference and potential.

AI models can also miss intangible considerations essential to human values and decision-making, including in fields like medical ethics that have developed for decades. As Mildred Cho, professor of pediatrics and associate director of the Stanford Center for Biomedical Ethics, recently observed, AI "[d]evelopers are often not from a medical background and haven't spent years thinking about this moral framework—how things like respect and justice and personal principles spill over into medicine."

And there are troubling signs that racial and other forms of bias in multi-modal models such as image generators remain largely unaddressed. Even some of the tools that in use today have merely been

"fixed" at the surface level—by requiring systems to automatically rewrite prompts to create less racist results for searches like "CEO." How models draw inferences from data, and their deep indebtedness to subtle cues, can pose an inherent problem for AI systems if left unacknowledged, untested, and unaddressed.

Such challenges post specific problems in the context of government programs. For example, facial recognition systems are notoriously bad at recognizing people from communities of color, as UnidosUS staff learned first-hand through our efforts to assist the federal government to enroll taxpayers in Puerto Rico who had become newly eligible for federal Child Tax Credits. Because enrollees needed identity verification through digital systems that often failed to recognize their faces on standard, Puerto Rican government ID cards, this frequently delayed or complicated their receipt of benefits.

Sometimes failure modes are more obscure. Models have been caught, after the fact, gathering clues on factors related to race or gender in hiring decisions, for example, from word choice or specific activities listed on a resume, even when information on race and gender has been omitted, leading companies to discard the models as too inherently biased to be used. Relatively simple automated decision-making models have also been shown to be deeply biased and to lack predictive value in areas such as mortgage lending, given the number of factors that function as proxies for race, even when protected class is omitted. Moreover, the black box nature of many models means that, without specific steps, including interrogation of the model for bias, impacts assessments and other forms of actual empirical evaluation, subtle forms of bias may remain undetected.

In areas like health care, these differences can be deadly if unnoticed. For example, as an article on the underdeveloped science of data quality for AI in *Stanford Medicine* notes:

> Scientists at Duke University Hospital, for instance, designed an AI program to identify children at risk of sepsis, a dangerous response to an infection. But the program took longer to flag Latino kids than white kids, possibly delaying the identification and treatment of Latino children with sepsis. The bias, it turned out, existed because doctors themselves took longer to diagnose sepsis in Latino kids. This taught the AI program that these children might develop sepsis more slowly or less often than white children.

As it turned out, three years into the effort, researchers learned that doctors took longer to diagnose the sepsis in Latino children, possibly because, among other reasons, Latino families were awaiting the arrival of hospital translators. Without external reviewers to gut-check and validate the conditions on the ground that produce data, many AI builders, or users, may not know what they do not know. A September 2020 report on uses of AI in sepsis monitoring programs, *Repairing Innovation*, notes that:

> …all too often, potential solutions remain just that—potential solutions, which may work in theory, given pre-set conditions. Rarely are these solutions tested, verified, or even used "in the wild." For this reason, we need fewer studies proposing how AI technologies could be used to address existing problems in the abstract, and more studies exploring how and in what ways could AI technologies be integrated into existing social processes such that they actually address those problems.

Just because something can be done, does not begin to answer the question of whether it should be done—or even whether it can be done consistent with our values, ethics, practical and societal considerations. The task for the agencies should first be to thoroughly inventory uses (as requested), to

create substantial new guardrails around *current* uses of AI tools in light of the new NIST RMF, and to publicly identify the successes, caveats, criticisms from stakeholders, and shortcomings of these uses as described above.

To the extent that uses of AI will *improve* productivity or performance of their goals and mission relative to current systems or possible alternatives, it could be adopted, of course, but—and this is an essential caveat—only if its risks can be adequately mitigated and the limitations of its uses mapped onto the needs of government agencies for fair, explainable, accountable, safe, and transparent systems. In many areas, these considerations remain an open question—and our governance must acknowledge that this is the case.

Exempting AI Uses at the Heart of Constitutional Governance Undermines Democratic Norms and Undermine Incentives to Develop Technologies that Are Rights- and Privacy-Enhancing

Similarly, perceived efficiencies from current and planned uses in criminal justice, immigration enforcement and related uses, and in public benefits, will likely lead agencies to continue to gloss over deeply concerning data security, stewardship, privacy, and civil liberties concerns. Dragnet forms of surveillance by the Department of Homeland Security (DHS) and law enforcement infringe on the rights of millions. Consider that:

- A September 2023 GAO report found that law enforcement at DHS and the Department of Justice (DOJ) lacked basic protocols or training around the use of facial recognition technologies, In response, DHS Sec. Mayorkas published a memo articulating a policy commitment to constitutional principles.
- Immigration and Customs Enforcement (ICE) used facial recognition technology to search the driver's license photographs of around 1 in 3 (32%) of all adults in the U.S. The agency has access to the driver's license data of 3 in 4 (74%) adults and tracks the movements of cars in cities home to nearly 3 in 4 (70%) adults. When 3 in 4 (74%) adults in the U.S. connected the gas, electricity, phone, or internet in a new home, ICE was able to automatically learn their new address.
- DHS has expanded use of facial recognition technology on travelers, including U.S. citizens, at airports and land borders without obtaining consent.
- A report by the Department of Homeland Security Office of the Inspector General (DHS OIG) revealed that Customs and Border Protection (CBP), ICE, and the Secret Service purchased and used commercial geolocation data in violation of their privacy policies and that DHS components have failed to develop policies governing the purchase and use of location data. According to DHS OIG these failures "occurred because the components did not have sufficient internal controls to ensure compliance with DHS privacy policies, and because the DHS Privacy Office did not follow or enforce its own privacy policies." The report recommended that CBP and ICE discontinue the use of commercial geolocation until they have developed and implemented sufficient policies, including conducting a privacy impact assessment. CBP promised Sen. Ron Wyden to stop purchasing location data by the end of Sept 2023.
- In July 2022, the American Civil Liberties Union (ACLU) published thousands of pages of previously unreleased records about how Customs and Border Protection, ICE, and other parts of the Department of Homeland Security are buying access to and using vast volumes of people's cell phone location information extracted from smartphone apps.

As this makes clear, use by governments of AI tools even in cases involving core matters of civil liberties, extremely vulnerable populations and privacy rights for immigrants and U.S. citizens, as well as legal due process and constitutional considerations, does not inspire confidence that the right guardrails are in place to dramatically expand uses of AI consistent with democratic principles, fairness, and other values.

Although the NIST RMF framework calls for AI to be "privacy-enhancing," OMB's approach fails to ensure that this will matter where it is needed most. Instead, the Memo's proposed waivers are likely to allow some of the most problematic and rights-infringing deployments of AI to continue to avoid even basic forms of public accountability. For example, as a law enforcement agency combining criminal and civil responsibilities, DHS or its sub-agencies may claim that law enforcement and national security exemptions apply or that an activity is "mission critical."

Hard cases cannot be the exception to our policies without undermining our fidelity to constitutional principles that rest at the core of our global leadership on personal freedoms and as a beacon of democracy. Instead, we need tools that allow fidelity to longstanding values and permit effective law enforcement. Therefore, OMB should develop a more tailored approach to these highly sensitive use cases and ensure that DHS is not permitted to side-step the implications of its many safety- or rights-impacting uses. Crucially, such waivers erode any incentive to do the hard work of aligning the design of systems with rights—but the failure to use privacy by design principles should not be characterized as a function of the technology when it is, instead, a choice to sanction unaccountable, untransparent and dangerous practices.

At the same time, we must immediately shut down and replace technologies and data that can be marshalled for authoritarian ends in the future. Given the need to future-proof government from the specter of abuse, OMB and the White House must lead a process of taking full account of current practices and fix them in short order. At a minimum, the OMB should create additional clarity regarding when agencies can seek waivers or exceptions from having to meet risk management requirements. We agree with the Leadership Conference comments that the following is needed:

- When a waiver or exception is granted, there should be a mechanism to seek reconsideration of such a decision.
- The Memo should be clear that waivers and exceptions sunset annually and should be reevaluated in light of these documented harms and risks.
- The Memo should require that agencies consider less rights-impacting alternatives before they are eligible for consideration for a waiver or exception.
- The Memo should require that agencies publicly report seeking waivers or exceptions, and the grounds for this request and its resolution and timing be reported.

In lieu of providing waivers, the U.S. should instead follow the lead of European governments in requiring individualized consent to the use of data without a court order. We should also require privacy by design principles that are compatible with effectiveness, such as strict data minimization, access controls, federated learning, and other privacy-enhancing techniques for government AI uses. Collection by agencies of biometric data, including DNA, should also receive specific scrutiny given its power in the hands of future Administrations that may lack any semblance of democratic restraints.

In keeping with the above, we fully support the Memo's provisions on procurement policies that underscore that AI contracts should align with national values and law, including "those addressing privacy, confidentiality, copyright, human and civil rights, and civil liberties." Since waivers for law

enforcement or mission-critical functions could undermine progress in assuring that federal tax dollars are not spent on systems incompatible with this requirement, consistency across federal procurement policy provides another reason to substantially narrow or eliminate waivers.

Second Pillar: Voice—Impacted Communities Deserve a Powerful Seat at the Table

As explained above, to deploy AI responsibly and ethically, address its false confidence about incomplete results, and incorporate nuanced understanding of its risks will require new and inclusive forms of governance. As we suggest above, systems should ensure they are accountable to the people they impact the most—including workers, creators, communities of color and lower-income people, and others who are often left behind and left out—for example, by traditional research methods (such as in health care), the digital divide (rendering communities invisible to AI models), or shifts in the nature and demands of educational preparation or work. For too many, the last decades of technological innovation failed to allow them to reap the benefits of a global economy even while its hyper-powered gains for the few.

Developing best practices could synergistically advance risk management, resilience, accessibility, privacy, ethics, and security. A prominent critique of technological advances has correctly been that the sector all too often celebrates narrow technical advances while ignoring social, political, health, and economic impacts. For AI, given its pace and reach, we simply must do better. Embracing democratic norms for governance can steer transformative technologies to benefit more people, more powerfully, lending them earned legitimacy at a moment of threat and disruption.

Seizing the Opportunity to Democratize AI Governance: An Interactive Model

Mechanisms like a transparent public dashboard, informed by technical metrics as well as real-world evidence gathering and input mechanisms, inclusive red teaming, impact assessments, and investment would be sound additions to the vision. Creating participatory public dashboards and the other key features of participatory methods outlined below would help to enable organic oversight at scale as AI capabilities advance.

The proposed transparency frameworks could be integrated with and would reinforce key goals within NIST's RMF. As the federal agency with expertise in creating technology standards through an open and collaborative process, NIST is well positioned to lead development of transparency benchmarks that cover both the technical dimensions and societal impacts of AI systems and is already doing so under the EO in consultation with relevant agencies and stakeholders.

The inclusive evaluation ecosystem we envision would underpin accountable, transparent, evidence-based AI governance while centering participatory design and democratic process. As explained below, OMB should:

- Work with NIST to develop **transparent public dashboards** on key performance indicators for models, including:
    - **Develop and iterate a real-time public dashboard** with feedback-informed benchmarks covering a wide range of impacts, as well as technical and societal metrics.
        - Both standards and benchmarks should be developed in consultation with impacted communities and build on existing AI alignment toolkits while expanding monitoring into crucial new dimensions of social impacts.

        - The model's scores should reflect both algorithmic real-time evaluations of the metrics and human input (from an interactive user interface that collects data and facilitates ongoing dialogue about experiences—both positive and negative—between model users and deployers and provides a record of efforts to resolve pending issues).
        - A general set of dashboard metrics applicable to most models can be accompanied by tailored metrics designed to be appropriate for specific rights- and safety-impacting use cases.
        - Metrics should be regularly updated and evaluated for fit-to-purpose based on the feedback, with processes to update parameters as technology and measurement science evolve and as public understanding and participation increase.
    - Create an archive of model cards in both technical and accessible formats to accompany the dashboard, adding needed context when major changes occur, including, at a minimum, flagging publicly when, according to the OMB Memo's definition section, there is a "Significant Modification" to a system.
- Embed diverse experts and stakeholder representatives in **AI Ethics and Impact Advisory Committee review boards to coordinate efforts on impacted community participation and drive outcomes consistent with shared values to**:
    - Conduct multidisciplinary and community-based consultation throughout the training, deployment, impact assessment, and iterative design process for both metrics and ongoing model evaluation and require participatory pilots and user testing in real-world community settings;
    - Develop best practices for transparency through public metrics tracking community feedback on specific AI systems, including consumer complaints and outcomes data; and
    - Trigger reassessments of performance and new mitigation steps when harm is indicated.
- **Formulate best practices for inclusive red teaming** that incorporates diverse perspectives to surface risks. Findings can be used to update metrics and inform dashboard content.
- **Conduct and set methodological standards for participatory impact assessments** to develop empirical evidence based on real-world impacts and ground-truth dashboard metrics and red teaming exercises. Proper design is essential.
- **Invest in community participation, education, and the growth of shared expertise.** As a matter of OMB's budgetary function, it should support dedicated funding and roles for public interest groups and participants to enable meaningful engagement in the development and iteration process for standards and benchmarks for above, alongside public education and training, and to distribute public monitoring, benefits, perspectives, and knowledge.

A Public Dashboard Would Facilitate Timely Monitoring of AI Progress, Impacts, and Risks

A centerpiece of our proposals is that NIST should develop, and agencies should apply, a real-time public dashboard providing broad visibility into AI model capabilities, limitations, and impacts based on both technical benchmarks and the incorporation of feedback from users and affected communities. NIST could develop approaches that apply to foundation models, and similar tools could be adapted across agencies and domains for more specific uses, as applicable.

If done well, an open monitoring infrastructure could enable an organic expansion of transparency as new risks emerge in this rapidly evolving field. At the same time, it could help to inform users in real

time of blind spots and gaps in AI models and systems. To develop fair and representative standards for the dashboard, the process of creating new metrics must include perspectives from impacted communities, civil rights, social scientists, ethicists, and public interest technologists alongside industry.

Keeping in mind the nascent state of the science of AI measurement, we approach this project with humility and a collaborative mindset. As described above, policymakers have already proposed key principles and standards to govern AI systems. The AI Executive Order, AI Bill of Rights, and NIST Risk Management Framework and accompanying materials are foundational documents that highlight many of the areas for specific metrics.

Approaches should also seek to leverage AI alignment lessons from the private sector. While voluntary commitments by the tech sector to the Administration acknowledge common goals and values that track the RMF, as competition has increased, public details about AI safety approaches from leading companies have become generally more constrained. At the same time, the largest companies have formed the Frontier Model Forum, and some are contributing to a new AI Safety Fund. While not by any means an exhaustive list, open-source efforts and attempts to create a transparent type of public performance metric or index thus far include, for example, Credo AI's early work on open-source algorithmic tools for a few performance indicators for AI models, Microsoft's safety coding tools in Azure, and sharing of code on AI safety by many companies and coders on GitHub. Notable efforts towards an index for AI include the Stanford Foundation Model Transparency Index, and the OECD's recently announced efforts on an Index for Trustworthy AI.

Static or hard-to-update regulatory models based on traditional approaches are also unlikely to be sufficiently nimble. Fine-tuning of AI models creates significant accountability challenges because it can radically alter model behavior in ways not captured by static documentation or one-time auditing. Most models are constantly being trained to improve performance, and the amount of data and parameters can range in the trillions. Fine-tuned models can inherit or amplify new biases, lose capabilities in previously demonstrated skills, become brittle and overfit to narrow datasets, introduce novel privacy or security issues, and experience performance drift over time. It is also more practical, in general, to look at model outcomes given the level of technical expertise needed to understand the implications of specific alignment and fine-tuning decisions and shifts. For all these reasons, continuous transparency mechanisms would help to keep pace.

The dashboard could provide a regular reevaluation of capabilities, limitations, and harms across an AI model (or a specific use case). If designed to take in user feedback and such feedback is channeled over time into the metric, the result is likely to be even more robust. As the release of large language models has demonstrated, user testing and input is essential. Ongoing participation by domain experts and affected communities would be equally critical to illuminate subtle risks as models change or are applied to new uses.

Governing AI will require a hybrid approach with coordinated oversight spanning foundational models as well as high-impact applications in sectors where decision algorithms already rule. Rapid growth in the capabilities of smaller AI models that use far less compute, and progress in the sophistication of models of every size, means that we will need governance approaches that can reasonably apply to any -safety or rights-impacting AI system. At the same time, foundation models are likely to be a source of countless future technologies and raise specific risks given their flexible and expanding uses.

Notably, if we can establish a scalable and workable system of quality assurances, it may not be too naïve to think that some opt into the use of monitoring dashboards or other inclusive processes to receive public credit for safe and ethical AI practices or benefit from the funnel of user feedback.

Work could be undertaken to develop and invest in measurement science across a range of metrics, and to consult with impacted communities in development of metrics and benchmarks. The dashboard should also clearly note the limitations of both the data and metrics and solicit help to further refine it with public input. Notably, both dashboards and standardized metrics focus on model behaviors and outcomes, rather than requiring disclosure of proprietary model architectures or parameters. Appropriately scoped metrics can quantify risks, harms, and capabilities without sensitive details.

They could also hew closely to the NIST RMF. NIST explains its model in this way:

> In addition to explainability and interpretability, among other AI system characteristics proposed to support system trustworthiness are accuracy, privacy, reliability, robustness, safety, security (resilience), mitigation of harmful bias, transparency, fairness, and accountability.

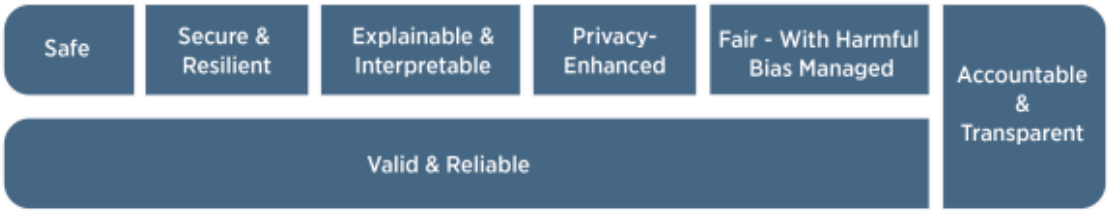


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Current efforts in AI alignment and transparency tend to focus on technical attributes of models. Some key gaps include:

- Minimal transparency around training data sources, curation processes, and labeling practices. This can obscure potential issues like a lack of diversity in training data, built-in biases or racial stereotypes, or misinformation.
- Limited monitoring of real-world performance changes over time as models are updated. Technical benchmarks using static datasets can diverge from the performance of models in deployment.
- Little visibility into how different user populations experience the model, especially marginalized groups that may be vulnerable to specific harms.
- A lack of current requirements to share public information about model limitations, potential risks, or effective mitigation strategies. While the sector has a history of academic openness, competitive pressures are now diminishing such practices.
- Little information for users on accuracy concerns, shortcomings, limitations, language (including cultural or dialect proficiencies and quality) and the quality of mitigations across languages.
- Insufficient information sharing about use or abuse of tools for criminal or illegal uses.
- Scant or few requirements in the U.S. to assure data privacy, data minimization, or privacy by design.

- A lack of agreement on metrics for bias and fairness and underdeveloped tools for communicating when an automated decision is made and how to appeal it.
- A dearth of metrics quantifying impacts such as those for bias, mental health, political polarization, treatment of disadvantaged groups, labor displacement and uses in the workplace and hiring, electoral integrity or democratic values, and other serious concerns.
- A lack of protections for intellectual property, copyright, and the absence of a system for compensation and consent for specific uses in training models, in disclosures to users, and to achieve redress.
- Minimal insight into environmental costs and sustainability practices around energy-intensive model development, iteration, and operation.

Metrics could anticipate and track factors like these, as feasible, and companies could report them on public dashboards. They could also identify levels of performance that are acceptable or problematic according to benchmarks. Each metric could provide a benchmark or range of benchmarks on performance, detail the methodology and data sources it uses, and be frequently updated.

In addition, for each category, companies could provide a means of reporting incidents, violations, or concerns so that public input informs the metrics in real time, and company responses are shared publicly. The ISO 9001 system for quality management (widely adopted by certified technology and other companies) could be an inspiration for an AI system of quality management. This system requires companies to develop mechanisms for corrective action requests and maintain documentation of responses.

As this suggests, there could be a user interface to accompany the dashboard for reporting of specific experiences and concerns by users, and this feedback should be incorporated in some manner into the metrics as appropriate. Models for such a system include consumer and industry incident databases for auto crashes (such as the National Highway Traffic Safety Administration's "early warning" system; the Consumer Financial Protection Bureau's and Consumer Product Safety Commission's consumer complaint databases; or the Federal Trade Commission's database, Consumer Sentinel.) The OECD Policy Observatory has also developed and launched an AI Incident database that collects articles relevant to incidents involving AI systems. A process should be developed so that this critical human feedback is also incorporated into the dashboard as appropriate once reviewed. In addition, the system should empower users to have control over their data, its uses, and to withdraw consent or authorize a third party to do so.

Achieving clarity in standards and best practices is not merely a legal nicety; it is the bedrock upon which businesses can build their strategies, innovate, and compete globally. The current absence of clear guidelines stifles innovation and dampens the economic promise of AI. Over time, the monitoring instruments would be likely to get better at detecting and assessing the measures, even as AI technology evolves rapidly, giving the larger public a voice by design. Users and participants in the reporting system, in turn, will be empowered by the knowledge that by reporting observations, the next round of metrics may reflect their feedback.

By providing continuously updated metrics across a comprehensive set of technical, ethical, and societal dimensions, a public database would offer significant advantages:

- **Timeliness:** Real-time data provides immediate visibility into model issues as they emerge. Rapid identification of problems could enable urgent interventions and address evolving issues with agility.
- **Adaptability:** The dashboard could expand and adapt as models increase in capabilities and reach. Over time, new metrics could be added to address novel risks identified through ongoing monitoring.
- **Transparency:** Metrics that create indicators for accuracy, safety, legal compliance, accessibility, environmental impact, and more could drive innovation in productive ways.
- **Public Knowledge:** Public monitoring could invite a much larger swath of participants into the discussion about the capabilities and limitations of AI models while generating useful data.
- **AI Alignment**: The approach could make use of existing techniques from the AI alignment toolkit.
- **Iterative and Flexible Benchmarks:** The dashboard would provide a framework for transparency without pre-determining rigid standards or thresholds for all metrics. For each metric, reasonable benchmarks could be set and recalibrated over time to measure adequate, superlative, and subpar performance.

An Archive of Model Cards Would Support Understanding of Shifts in Model Performance

In addition to real-time metrics, developing a version-controlled and publicly accessible archive of model cards can further strengthen transparency. Model cards should include details on a model's purpose, performance evaluations, ethical considerations, limitations, and other factors. They could be updated frequently, including, at a minimum, to flag publicly when, according to the OMB Memo's definition section, there is a "Significant Modification" to a system. Requiring simultaneous release of two model card variants could facilitate transparency for public and technical audiences:

- **Public Model Cards** would use non-technical language focused on model impacts on users and society. They outline intended use cases, measures to address risks, and contact points for user complaints.
- **Technical Model Cards** would provide detailed metrics and benchmarks for researchers and developers. They include training data specifics, model architectures, intended applications, technical limitations, and safety considerations.

NIST could develop best practices for model card formats and archives in conjunction with the dashboard standards. At the same time, these approaches could also be designed to protect legitimate commercial interests and the intellectual property (IP) of AI developers.

Creation of AI Ethics and Impacts Advisory Committees Would Inform and Help Coordinate Agency Efforts

While the creation of agency leadership by the Memo is laudable, agencies that are already invested in specific technologies and use cases are unlikely to be as objective as they should be in evaluating potential and actual risks. For this and other reasons, OMB should also require agencies with high-risk use cases to establish participatory oversight boards, such as AI Ethics and Impacts Committees, with representation from impacted groups.

Importantly, the NIST RMF and 2023 Report by the National Artificial Intelligence Advisory Committee, alongside numerous provisions in the EO, encourage "consultation" with affected communities across many stages of AI development and deployment. But such consultations will only matter if they are attuned to the parameters in the EO and OMB memo and include specific requirements for the role of outside stakeholders.

For this reason, OMB should require the creation of foundation model (at NIST) and agency-level, domain-specific **Ethics and Impacts Advisory Committees**, with defined authorities to help agencies assessment of AI systems for accountability, transparency, and other ethical dimensions and with public outreach and engagement. The Committees should be comprised of impacted community members, organizational representatives, subject matter experts and professionals, and relevant legal and professional ethics experts.

Such boards, which could be constituted under the Federal Advisory Committee Act, should play a critical role in reviewing the uses of higher-risk systems prior to deployment, with processes for agency responses on matters of concern, similar to many such arrangements across federal agencies today. In addition to substantive roles in revising use cases and risks, their engagement could include, for example, supporting agency efforts to:

- Host inclusive design workshops to gather diverse perspectives on potential AI applications and their impacts.
- Conduct community data audits to assess biases and gaps in training datasets.
- Organize explainability clinics to convey how systems work to community members.
- Facilitate participatory red teaming (as described below) to surface risks.
- Support deliberative forums for the public to discuss AI's broad societal effects.
- Provide grants for exchanges and co-creation sessions.
- Train community members to audit algorithms and data for biases.
- Fund user experience research focused on inclusive populations.
- Sponsor AI education events and demonstrations tailored for specific communities.

As described below, dedicated funding streams are necessary to enable these critical forms of public interest participation, which well-resourced industry voices can crowd out. At the same time, building capacity with a range of funded and well-structured roles can formalize avenues for impact on standards development and create opportunities for sustained engagement. Importantly, compensating community members respects the significant time and labor that this participation entails.

Agencies should also be expected to support their efforts, and to effectively and publicly communicate learnings and developments that can improve AI systems and use contexts over time, including to:

- Host regular conferences and workshops bringing together AI developers, regulators, civil society groups, and impacted communities to share insights.
- Publish annual reports summarizing key trends, challenges, and best practices identified through oversight processes and widely disseminate them to stakeholders.
- Maintain publicly accessible databases of algorithmic audits, impact assessments, and other empirical evidence that developers can reference.
- Establish public bug bounty programs to incentivize researchers to probe systems for flaws and submit reports.

- Conduct mediated learning exchanges between developers and community leaders.
- Develop open-source libraries of bias testing code, auditing frameworks, and mitigation techniques that developers can integrate.
- Create AI development test environments with synthetic or scrubbed data in which researchers can collaboratively experiment and facilitate sharing of methods.
- Fund fellowships and exchanges between academia, industry, government, and civil society to cross-pollinate ideas.
- Sponsor challenge competitions for developers to create solutions focused on priorities like algorithmic fairness or interpretability.
- Implement public notification systems for spreading urgent warnings about vulnerabilities, flaws, or harmful incidents needing immediate attention.
- Cultivate communities of practice around issues like inclusive data practices, engineering accountability, and human-centered AI.
- Develop AI ethics training programs and make open courseware freely available to all developers to build aligned values and norms.

A community-centered approach would build shared capacity to help govern AI responsibly and holistically. It would also help to unlock the full potential of America's diversity as a competitive advantage in AI development. Sufficient investment in inclusive participation to build and support the AI economy is an imperative for shared prosperity and innovation.

Inclusive Red Teaming Can Surface Assumptions and Overlooked Risks and Harms

Red teaming and similar exercises are essential to stress test AI systems and uncover potential vulnerabilities and failure modes before deployment. Integrating the lived expertise of underrepresented groups into hands-on red team exercises can provide a constructive way to catch vital risks and biases that teams can overlook, including exploring potential for misuse and unintended consequences in a community-based or cultural context.

Multidisciplinary red teams, including domain experts in social sciences, civil rights advocates, ethics, directly affected community members, and other non-technical stakeholders can surface overlooked risks, but are often not a part of reviews. For just one example, with commendable candor, OpenAI's system card from May 2023 acknowledges a lack of inclusivity in process design:

> Participants in this red team process were chosen based on prior research or experience in these risk areas, and therefore, reflect a bias towards groups with specific educational and professional backgrounds (e.g., people with significant higher education or industry experience). Participants also typically have ties to English-speaking, Western countries (such as the US, Canada, and the UK). Our selection of red teamers introduces some biases, and likely influenced both how red teamers interpreted particular risks as well as how they probed politics, values, and the default behavior of the model. It is also likely that our approach to sourcing researchers privileges the kinds of risks that are top of mind in academic communities and at AI firms.

As part of the assignment from the EO to NIST on protocols for red teaming, OMB should request that NIST or another appropriate entity lead development of best practices for more inclusive red teaming methods that intentionally incorporate diverse voices. OMB efforts can then ensure that agencies make use of these best practices.

More intentional and structural involvement of diverse communities would strengthen red teaming of AI systems to:

- Provide perspectives to catch biased assumptions in data or model design that could lead to discriminatory performance, based on real-world experiences of exclusion.
- Identify potential harms or negative social impacts on minority communities that may be invisible to homogeneous teams.
- Evaluate language and speech recognition for robustness across diverse linguistic contexts and dialects.
- Assess visual classification robustness for non-white faces and fairness across skin tones.
- Gauge impacts on economic opportunity and access to services for disadvantaged demographics.
- Model adversarial cases of systems being used for surveillance, over-policing, or exploitation of vulnerable groups.
- Enhance scenarios representing real-world usage by underserved populations and non-mainstream cultural contexts.
- Improve simulations of misuse by bad actors motivated by racism, hate, or unethical profit.

Moreover, for particular domains, intersectional and diverse experts in psychology, political science, labor, electoral and voting integrity, and other relevant experts can offer insights into risks like propaganda amplification, radicalization, divisive polarization, and disparate impacts on vulnerable communities. Experts in ethical norms informed by power differences in a range of settings such as health care, psychology, criminal justice, education, and other settings with decision-making roles over human possibilities and lives may also add depth to multidisciplinary processes. As an intermediate step, OMB should request that relevant agencies work with NIST to help inform best practices for inclusive red teaming of AI systems by:

- Leading by example in red teaming AI projects with diverse external partners.
- Publishing guidelines emphasizing the importance of diverse, multidisciplinary red teams and participatory design processes.
- Providing a framework for identifying stakeholders from impacted groups and compensating them for engagement.
- Developing exercises focused on surfacing risks to marginalized populations.
- Creating rubrics for red teams to self-assess inclusiveness across factors like race, gender, sexual orientation and gender identity, age, and disability.
- Maintaining a public repository of red teaming anti-patterns that led to preventable failures.
- Hosting workshops demonstrating effective inclusive red teaming in collaboration with civil rights and other issue area experts.
- Encouraging the use of techniques like adversarial debiasing, participatory simulations, and sensitivity audits.
- Integrating evaluation of inclusiveness into procurement standards for companies providing AI to government.
- Funding external red teaming by civil society groups on AI systems that impact public interest.

The intentional integration of diverse perspectives into the red teaming process will enable a more rigorous exploration of how AI systems can fail when deployed in complex real-world contexts. In addition, standard steps for data gathering to inform inclusive red teaming exercises could include:

- Soliciting and conducting reviews of impact assessments and outcomes data for specific domains and applications.
- Performing user studies by surveying and interviewing diverse community members about their experiences interacting with an AI system. These could gather qualitative insights on usability, trust, and value.
- Funding community-based organizations to monitor AI systems through methods like user surveys, observational studies, and complaint tracking and synthesize findings in public reports.
- Implementing A/B testing methodologies to try variations of an AI system and measure comparative impacts. Testing could assess dimensions like fairness, interpretability, and user autonomy.
- Conducting pre-deployment impact assessments (as described below) using criteria on dimensions like privacy, accessibility, and bias.
- Integrating feedback surfaces like user rating systems or open comment and complaint forums to receive direct public input.

Leveraging Impact Assessments for Empirical Evidence to Inform Metrics

In addition to inclusive red teaming, as the above list suggests, facilitating and supporting regular impact assessments focused on real-world effects can provide another vital feedback channel to strengthen AI governance. Crucially, these should be conceptualized as "third party audits," and not as internal to government agencies. As AI Now describes:

> Third-party audits stand apart: they have been conducted by journalists, independent researchers, or entities with no contractual relationship to the audit target. From Gender Shades to the audit of London's LFR system to ProPublica's audit of predictive policing tech, these audits have been pivotal in galvanizing advocacy around AI-related harms.

Assessments also must be well-designed to produce tangible, specific, and usable results that inform standards. Formal evaluations conducted in partnership with public interest and other stakeholders can surface overlooked issues, generate empirical insights on how systems perform in actual usage, and center data on the impacts to, and experiences of, affected groups and individuals. Findings could directly inform iterative improvements to policies, model training, dashboard benchmarks, and other governance mechanisms.

As part of its collaborative standard-setting process, OMB should ask NIST and agencies to solicit input from key stakeholders to develop methodologies and priority metrics for AI impact reviews. Public reports would detail findings, recommended actions, and document whether previously identified problems have been suitably addressed. Assessments could be required more frequently for safety- and rights-impacting uses in areas like finance, hiring, housing allocation, healthcare, and education.

Areas for participatory analysis could include, for example:

Fairness and Bias:

- Disparate impact analysis of system decisions and outcomes by demographic
- Assessment of training data biases and gaps for underrepresented populations
- Surveys gauging user trust and sentiment across community groups

Safety and Accountability:

- Documenting issues reported through harm redressal processes and analyzing response adequacy
- Auditing security vulnerabilities disclosed through whistleblower and bug bounty programs
- Tracking rates of problematic or illegal content promotion

Accessibility:

- User studies on accessibility barriers and compatibility gaps with assistive technologies
- Feedback from disability or other advocates on areas for interface improvements
- Language quality for non-English languages and dialects for model performance and user support

Broader Societal Impacts:

- Labor economists that can estimate workforce impacts and job replacement projections
- Mental health experts to gauge psychological risks like addictive system designs
- Public health scientists and others to review scientific accuracy and misinformation risks
- Electoral and democracy scholars to evaluate institutional and integrity exposures and risks

Regular inclusive impact assessments would help provide external validation for oversight processes. Centering community voices and empirical insights within governance cycles fosters accountability and demonstrates that a priority is placed on improving system impacts and real-world outcomes rather than on narrow or technical measures alone.

Specific Comments on OMB Text Highlight a Need for Additional Direction in Some Areas

The OMB Memo helpfully references the EO's definitions of algorithmic discrimination (Section 10(f) of the EO), equity (10(a)), underserved communities (10(b) of EO 14091), and defines automation bias. Perhaps most consequentially, it defines a broad range of rights-impacting AI uses and risks. These represent clear progress on federal policy on AI. We further support the requirement for agencies to "consult and incorporate feedback from affected groups... including underserved communities" and applaud that the Memo specifies that "[i]n the event of negative feedback, agencies must consider not deploying the AI or removing the AI from use."

We would urge more specificity about the kinds of "negative feedback" and consideration of options and mitigations should be required. For example, if an agency were to find that a rights-impacting AI technology was biased or discriminatory, or that it lacked sufficient transparency or accountability, OMB should clarify whether this policy would bar its use, or merely require that an agency "consider" whether not to use it, and on what basis such considerations should be made.

As described above, while frameworks like NIST's highlight the need for community consultation, specific processes and decision-making power must be defined to make this engagement meaningful rather than symbolic. We need new ways of collaborating across sectors and to go beyond check-the-box forms of "consultation" with impacted communities to full partnership.

We also note that the stakeholders named by the draft Memo in this section include "affected groups, such as customers, Federal employee groups, and employees' union representatives." This is laudable, but absent specific, intentional, and well-resourced efforts to solicit and integrate a diverse set of perspectives from all impacted communities, this remains an underdeveloped list. OMB should identify specific steps for agencies to take as part of a program for robust public outreach.

In addition, OMB should require agencies to provide information on AI uses and models that facilitate a full conversation about risks and values, including by requiring agencies to publish information that is available about data used to train models, impact assessments, experiential data and criticisms of model uses. This can be achieved through model, system or data cards providing both technical and plain language information as well as ample other sources of information that are available to the agency. As described above, agencies should also be asked to identify gaps in knowledge or resources about data or impacts used in models, alongside any other factors that inform the agency's uses and views of the model's limitations.

The Memo provides several helpful examples of how feedback may be collected, including direct user testing, general solicitations of comments from the public, public hearings or meetings, and other process to seek public input, comments, or feedback from the affected groups in "a meaningful, equitable, accessible, and effective manner." These positive steps are important for agencies to take, and the specificity is appreciated.

In addition, we welcome the Memo's requirement that agencies must "monitor rights-impacting AI to assess and mitigate AI-enabled discrimination against protected classes that might arise from unforeseen circumstances, changes to the system after deployment, or changes to the context of use or associated data." We note that such monitoring is essential given that models are frequently fine-tuned in ways that can alter performance and change outputs. We agree that "where sufficient mitigation is not possible, agencies must safely discontinue use of the affected AI functionality" and would urge that there be efforts to further define, in specific cases, what kinds of mitigation would be deemed "sufficient" or insufficient in this context, as that is what will matter in the execution of this sound principle.

We support the provision that agencies should "notify negatively affected individuals" and would urge an expansive definition of when it would be "practicable and consistent with applicable law and governmentwide guidance" to do so. Substantive due process will often mean that such notice is or should be required, but there are strong internal disincentives within government to resist such notification. Overall, this section, which includes plain language and multi-language requirements, is strong, but will require active oversight to ensure its appropriate execution. We would urge that OMB develop robust means of monitoring and level-setting across agencies around the implementation of these requirements, including the appeal processes they should engender.

We would also advise, given due process, that agencies not be merely "strongly encouraged" to provide an explanation, but be required to do so. Those negatively impacted by a model are certainly due an

explanation. If the explainability of a model is difficult to achieve, the requirement for an explanation is more (not less) warranted. In contrast, a lack of requirement for one could allow bias to remain undetected. The Memo helpfully indicates that:

> Explanations might include, for example, how and why the AI-driven decision or action was taken. While exact explanations of AI decisions are often not technically feasible, agencies should characterize the general nature of such AI decisions through context, such as the data that the decision relied upon, the design of the AI, and the broader decision-making context in which the system operates. Such explanations should be technologically valid, meaningful, useful, and as simply stated as possible, and higher-risk decisions should be accompanied by more comprehensive explanations.

We also recommend that OMB more fully incorporate NIST's approach to explainability, which provides that explainable systems should adhere to four principles:

- Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
- Meaningful: A system provides explanations that are understandable to the intended consumer(s).
- Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.
- Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

These would provide additional clarity about the components of explainability, including providing a better benchmark for when systems may fail to be adequately explainable to be continued in that use. The OMB Memo further requires that "[w]hen law or governmentwide guidance precludes disclosure of the use of AI or an opportunity for an individual appeal, agencies must create appropriate mechanisms for human oversight of rights-impacting AI." OMB should more specifically define what is an "appropriate" or inappropriate mechanism, as it poses the prospect of potential abuse and a lack of transparency in government decisions or processes.

In addition, the Memo includes a requirement for agencies to "prominently provide and maintain a mechanism to conveniently opt out from AI functionality," noting that "an opt-out mechanism must exist where the affected people have a reasonable expectation of an alternative or where lack of an alternative would meaningfully limit accessibility or create unwarranted harmful impacts." We support this provision but note that it will require significant additional effort to set out what it means in practice and that it is essential that OMB further define its thinking on what it means to "meaningfully limit accessibility or create unwarranted harm" given that these may apply to many AI systems in operation today. The availability of a means to consent (or not) to the use of an AI must be articulated as a fundamental principle, and so it will be well worth additional effort for OMB to specifically solicit additional and focused public comments on this provision.

To further acknowledge accessibility issues, we recommend adding a new part to Section 5.b.ii that specifically calls out the needs of the disability community, who are uniquely vulnerable to AI in any number of existing categories but deserve their own consideration. In addition, OMB should propose a process by which new Safety-Impacting or Rights-Impacting purposes can be added to the AI guidance, including through public input, as technology and its usage evolves.

We join other commenters in noting that previous AI EO and OMB guidance directing agencies to inventory AI regulatory authorities should be retained and strengthened. Previous EO and OMB guidance required agencies to submit to OMB an agency plan that "must identify any statutory authorities specifically governing agency regulation of AI applications, as well as collections of AI-related information from regulated entities." As the Department of Health and Human Services was the only one to publicly release their memo on AI authorities, we join The Leadership Conference in making the point that this inventory of agency regulatory authorities for AI is a vigorous exercise for agencies and is of interest to Congress, civil society, and the public. The OMB AI guidance should require that agency authorities be submitted to OMB and the new White House AI Council and make the list public.

Third Pillar: Investment—Growing an Ecosystem for Equitable Participation, Public Trust, and Innovation Insights

Developing transparent standards for AI systems through a truly participatory process requires proactive efforts to include public interest groups and historically marginalized voices. Many such stakeholders could make the difference in public understanding of and engagement with how technologies evolve, yet currently lack resources to assist or face barriers that may prevent meaningful participation. Layers of overlapping new practices and methods are required.

To democratize the benefits, there is also a need for human infrastructure capable of leveraging AI, including building capacity in and among communities that are too often invisible in developing and deploying technologies. We must close skills and talent gaps—and a robust program to do so will address the needs of a burgeoning industry while mitigating persistent sources of economic inequality.

Substantial funding for creating a more diverse pool of digital economy workers through partnerships with culturally competent community-based organizations, as well as skills-based hiring that connects workers of all skills and backgrounds to an AI economy, would help workers who are currently excluded from upwardly mobile, future-focused employment find career pathways into the AI economy. Such investments would allow the benefits of AI to be more fully shared and could help develop expertise and spark interest. Building technical fluency will allow impacted communities to both help steer these powerful technologies toward sustainable and equitable progress and improve the relevance of AI tools.

Working with the Department of Commerce and other stakeholders, and through its budgetary function with input into the President's 2025 budget, OMB should develop and inform a process for grants directly to impacted communities and nonprofit public interest organizations focused on issues implicated in AI and ethics. This funding could support participation in the standards and governance process and independent research. Potential functions could include:

- As described above, participation and training for inclusive red teaming, impact assessments, and dashboard review.
- Community representatives embedded within an agency and surface issues and gaps.
- Community auditors can train community members to audit or conduct impact assessments.
- Focus groups should be compensated and include diverse individuals who can share experiences and evaluate AI through interviews, surveys, and workshops.
- Participatory designers can create spaces for the public to directly steer technology innovations toward collective priorities, informing benchmarks.

- Inclusive policy fellowships could embed experts from communities within oversight bodies.
- Accessibility consultants could inform fully inclusive design in both standards and products.
- Small business assistance programs could enable AI adoption by enterprises in disadvantaged communities.
- Test participants could include users from underrepresented groups to assess disparate experiences.
- Funding can be set aside for research investigating historical harms from relevant technologies and practices.
- Investigative researchers can include a diverse range of experts, including social scientists funded to study datasets, labeling practices, training processes and model architectures for transparency, fairness, or other gaps.
- Policy consultants can include diverse panels of experts to inform initiatives.

The President and OMB could also recommend that Congress establish a dedicated fund modeled on the CDC Foundation that could support digital skills education and training, community-based AI auditors, participatory technology workshops, and other capacity building to close knowledge and equity gaps. Grants to local organizations would enable national assessments of AI's impacts on disadvantaged groups and workforce needs and could build expertise and understanding of technological tools within impacted groups. It should include funding for community organizations to build AI-specific expertise and support participation in the types of community engagement outlined above.

The CDC Foundation is an independent nonprofit and the "sole entity created by Congress" to mobilize private sector and philanthropic resources to support the CDC's work on public health. It could be funded by a mix of donations from the technology sector, alongside a federal endowment that grows through licensing fees or other federal supports, to provide grants for community organizations initiatives such as:

- Equipping organizations and their members to use the tools and provide feedback (helping to ensure the digital divide does not render communities or individuals invisible);
- Working to solve the talent gap and develop the evidence base for upskilling and skills-based employment through partnerships and innovative approaches to recruitment of an economically and racially diverse workforce;
- Increasing capacity for diverse voices to engage as peers in regulatory and oversight processes around AI tools, including funding nonprofits to train community auditors on AI systems and conduct audits of algorithms, data collection and uses.

The fund could be overseen by an independent governing board with representatives from government, industry, academia, civil rights groups, and community organizations and be housed within an existing agency or as a nonprofit organization to manage the grantmaking process. A national yearly assessment of AI skills gaps, workforce needs, and barriers to digital inclusion could inform its priorities and processes. Nonprofit community organizations could apply to receive grants for purposes such as:

- Providing AI and technology job training programs and internships in impacted areas;
- Developing curricula and certifications for community members to become AI auditors;
- Funding data and policy experts to enable meaningful technical input into AI systems;
- Supporting participatory design workshops and exchanges between developers and community residents;

- Working with recipients to document community partnerships, inclusion plans, and to measure and reflect on impact;
- Funding programs to enable collaborations with schools, employers, and industry partners to create pathways to equitable AI workforce participation.

The fund could also maintain a publicly accessible database of grants awarded, results achieved, and best practices for building AI expertise in local communities to build field knowledge and insights.

Conclusion: From Principles to Implementation: Creating an Ethical AI Ecosystem

Technical progress must be paired with social progress. We must connect principles to action and move to create functional guardrails, meaningful participation, and expanded capacity. By linking ethical guidelines, participatory structures, and human infrastructure, we can build an AI future that reflects our ideals.

Most importantly, inclusive and democratic practices should infuse AI governance itself. Impacted communities deserve structured involvement in shaping these powerful technologies. With broad collaboration and human-centered design, the new powers of AI systems can be governed ethically at the outset rather than regulated after the fact and once harms are already entrenched.

American leadership in developing AI that expands opportunity while respecting Constitutional values is essential for innovation and an inclusive future. Progress will demand openness to creative partnerships and evolving best practices. We applaud OMB's draft guidance as a crucial first step for agency AI systems. We also welcome OMB's commitment to exploring effective ways to safeguard innovation in the era of AI. We also welcome additional opportunities for collaboration and input.

For questions or additional dialogue on these issues, please contact Laura MacCleery, Senior Director of Policy, at lmaccleery@unidosus.org, and Claudia Ruiz, Senior Civil Rights Analyst, at cruiz@unidosus.org.