
**Comments of the World Privacy Forum**

**To**

**The Office of Management and Budget regarding *Request for Comments:Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum OMB-2023-0020.***

*Via www.regulations.gov*

Clare Martorana
U.S. Federal Chief Information Officer
Office of the Federal Chief Information Officer,
Office of Management Budget

December 3, 2023

The World Privacy Forum is pleased to submit comments in response to the Office of Management and Budget's "Request for Comments on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Memorandum." OMB published the request in the Federal Register on November 3, 2023, https://www.govinfo.gov/content/pkg/FR-2023-11-03/pdf/2023-24269.pdf.

The World Privacy Forum is a respected NGO and non-partisan public interest research group focused on conducting research and analysis in the area of privacy and complex data ecosystems and their governance, including in the areas of identity, AI, health, and others. WPF works extensively on privacy and governance across multiple jurisdictions, including the US, India, Africa, Asia, the EU, and additional jurisdictions. For more than 20 years WPF has written in-depth, influential studies, including groundbreaking research regarding systemic medical identity theft, India's Aadhaar identity ecosystem —peer-reviewed work which was cited in the landmark Aadhaar Privacy Opinion of the Indian Supreme Court — and The Scoring of America, an early and influential report on machine learning and consumer scores. WPF co-chairs the UN Statistics Data Governance and Legal Frameworks working group, and is co-chair of the WHO Research, Academia, and Technical Constituency. At OECD, WPF researchers participate in the OECD.AI AI Expert Groups, among other activities. WPF participated as part of the first core

group of AI experts that collaborated to write the OECD Recommendation on Artificial Intelligence, now widely viewed as the leading normative principles regarding AI. WPF research on complex data ecosystems governance has been presented at the National Academies of Science and the Royal Academies of Science. World Privacy Forum: https://www.worldprivacyforum.org.[1]

This Administration's focus on artificial intelligence is welcome and timely. In general. WPF is happy with the President's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Executive Order 13960 of December 3, 2020, https://www.govinfo.gov/content/pkg/FR-2020-12-08/pdf/2020-27065.pdf. We also welcome and support OMB's AI Memorandum. Both documents are good starts at tackling the challenges of AI.

We want to offer a few suggestions for the OMB AI Memo. The focus of the first part of these comments is on activities that involve or affect individuals and their personally identifiable information. We recognize, of course, that there are many AI activities that do not involve privacy matters.

**Part I: Use Existing Privacy Requirements to Assist with New AI Obligations**

Existing law provides a useful method for informing the public about and for seeking comments on agency activities affecting privacy. We refer, of course, to the Privacy Act of 1974, 5 U.S.C. § 552a. The Privacy Act of 1974 directs agencies to use Systems of Records Notices to describe their activities involving the use of personal information. Agency activities that involve personal information are described through notices in the Federal Register with the solicitation of public comment, and the notices remain available publicly to inform the public of agency activities. In many ways, publications by agencies under the Privacy Act of 1974 are unique. Other counties that have more advanced and more comprehensive privacy laws mostly abandoned requirements for publications describing descriptions of personal data systems. In our view, the Privacy Act of 1974's publication provisions are one of the Act's most successful obligations.

The Privacy Act of 1974 became law long before anyone envisioned the need for a privacy impact assessment (PIA). A later law, the E-Government Act of 2002, 35 U.S.C. § 3501 note, imposed a requirement for PIA. The statutory requirement was always inadequate for the purpose, but it was supplemented by useful OMB guidance, the OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002 (September 26, 2003) (M-03-22).

To a significant extent, many of the features of both of these PIA requirements are simply out-of-date and trail current technology and information practices by decades. We do not expect OMB to solve these broader problems in its AI memorandum. However, it should be possible to integrate new agency obligations for AI with existing obligations under both the Privacy Act of 1974 and the E-Government Act.

---

[1] World Privacy Forum's home page includes information about our activities, as well as numerous data governance and privacy research, data visualizations, and resources. https://www.worldprivacyforum.org.

There is a greater opportunity here as well. We note that this Administration is aware of the need to focus more attention on the use by federal agencies of data brokers who provide personal information for the use of federal agencies or of data brokers who receive personal information from federal agencies. WPF applauds that effort with great enthusiasm. Many of these activities are hidden from the American public because existing privacy laws and rules do not expressly direct agencies to disclose their use of data brokers as sources or recipients. Data broker information resources are integral to some AI activities.

It is possible without a major effort to fold the AI and data broker initiative together with a limited reform of existing Privacy Act of 1974 and PIA requirements. Happily, an initiative sponsored by WPF already includes ideas and drafted language that will accomplish all of these tasks. That project is a report titled From the Filing Cabinet to the Cloud: Updating the Privacy Act of 1974 (May 2021), https://www.worldprivacyforum.org/2021/05/from-the-filing-cabinet-to-the-cloud-updating-the-privacy-act-of-1974/.

The goal of that report is a complete revision of the Privacy Act of 1974. That goal will not be accomplished administratively, but some administrative changes can achieve the same ends. For example:

1) **Better publications**. The report proposes on page 85 to expand existing requirements to describe categories of sources of information for a Privacy Act system of records. We note that AI is likely to be a new category of information for many AI activities covered by the OMB memorandum. Language in the report – which can work just as well in an OMB memorandum as in a statute – expands upon the existing requirement to describe the categories of sources of records in the system.

> New language makes it explicit that sources include commercial, governmental, and other sources that the agency routinely reviews, consults, or uses. It is especially important for agencies to inform the public when using commercial sources. For example, if an agency has a contract with a consumer reporting agency ("credit bureau") to use credit records, it must so state. If there is a reasonable prospect that the particular source may change but not the category of sources, the agency may choose to identify the category (e.g., "credit bureau") rather than identifying which specific credit bureau it uses. If an agency routinely uses Internet search engines to find information on individuals, it must so state. If an agency routinely seeks information from social media, the agency should identify at least the major social media used. All the information about sources will help individuals figure out how particular information about them ended up in agency records. This is especially important when the agency uses the information to make decisions about individuals. It is even more important if an agency consults but does not retain a copy of information held by a third party.

This type of disclosure addresses current Administration priorities for AI and for data brokers. OMB could make an adjustment in its Privacy Act of 1974 and PIA administrative requirements to expand public disclosure on data broker activities.

2) **Better PIAs**. Formal impact assessments can be useful to address many different policy concerns. Everyone seems aware of the overlaps between some AI assessments and some privacy assessments. That same report that proposed revising the Privacy Act of 1974 also

suggested revising the existing privacy impact assessment requirement. We will not repeat the details here. We refer you to pages 115-123 of the report.

We summarize here, however, by noting that the report emphasized the need for a PIA ***process*** and just not a flat, one-time, one-size-fits-all assessment. This need will be just as true for an AI assessment as it is for a privacy assessment. We state expressly that: 1) some assessments will require more attention, more consultation, and a longer time than others; 2) some assessments will require regular reviews over time because consequences are not static and because agency programs change over time; and 3) some assessments will require lesser efforts because the risks are smaller. Responsible agency personnel should be allowed to make determinations about which activities need more assessment than others.

Given the significant overlap with some AI assessments and some privacy assessments, we suggest that an agency's Privacy Officer and the agency's AI Officer be directed to identify those overlaps and to work jointly on the assessments that they determine need substantial attention from both perspectives. Other specific ideas on assessments from the Privacy Act revision will also be relevant, and all adjustments to existing guidance can be accomplished administratively by OMB in its AI memorandum.

**Part II. AI Activities**

We have some additional observations specific to AI activities.

In the OMB's definition for risks from the use of AI in its *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum*, the agency rightfully mentions specific factors that create, contribute to, or exacerbate the risks of AI including "AI outputs that are inaccurate or misleading" and "AI outputs that are unreliable, ineffective, or not robust."

We applaud the OMB's recognition that some AI systems simply may not be adjustable in an acceptable manner that adequately mitigates risks, and therefore should not be used. The memo's demand that, "Where the AI's risks to rights or safety exceed an acceptable level and where mitigation is not practicable, agencies must stop using the affected AI as soon as is practicable," is a positive guidepost for AI use by agencies.

Ensuring that AI systems do not produce outputs that create, contribute to, or exacerbate risks is imperative. In addition, there also is a pressing need to ensure that AI governance tools—the tools and techniques used to measure AI system risks and improve their inclusiveness, fairness, explainability, privacy, safety and other trustworthiness factors— *themselves* do not compound those risks.

AI governance tools can improve such issues; however these tools often lack meaningful oversight and quality assessments. Incomplete or ineffective AI governance tools can create a false sense of confidence, cause unintended problems, and generally undermine the promise of AI systems.

U.S. federal agencies have important gatekeeper and quality assurance roles when it comes to ensuring that the tools and techniques employed to measure and improve the AI systems they use are reliable and effective.

We appreciate the opportunity to comment regarding specific questions in the OMB's Request for Comments on the Draft Memorandum. Below are responses to some specific questions.

**Question 2. What types of coordination mechanisms, either in the public or private sector, would be particularly effective for agencies to model in their establishment of an AI Governance Body?**

The OMB memo lists several responsibilities of Chief Artificial Intelligence Officers including "Managing Risks from the Use of AI" through "working with relevant senior agency officials to establish or update processes to measure, monitor, and evaluate the ongoing performance of AI applications and whether they are achieving their intended objectives."

Such evaluation processes should include quality assessment of the AI governance tools and techniques used to evaluate and measure the AI systems. Establishing appropriate processes for this purpose will require evidence gathering and better understanding of realities on the ground before a rush to solutions. Otherwise, these processes may well be unfit for purpose and fail to accomplish the goals which are intended — and in some cases mandated — to attain.

We urge agencies to collaborate with civil society organizations and academia in addition to private sector entities to A) build an evaluative environment for testing AI systems, B) gather relevant evidence and work toward appropriate AI measurement and monitoring processes, and C) work to create appropriate AI benchmarks.

In particular, agencies might look to the National Institute of Standards and Technology (NIST) in convening stakeholders and developing an evaluative environment in which an evidentiary basis for the socio-technical contexts and best practices for AI governance tools could be created.

**Question 5. "Are there use cases for presumed safety-impacting and rights-impacting AI (Section 5 (b)) that should be included, removed, or revised? If so, why?"**

Some use cases involving specific measurement methods for measuring or improving AI fairness or explainability should be assessed thoroughly for appropriate application, and possibly removed from consideration as evaluative AI measures.

In the World Privacy Forum's forthcoming report, *Risky Analysis: Assessing and Improving AI Governance Tools, An international review of AI Governance Tools and suggestions for pathways forward*, we present research showing problems with some commonly-used methods intended to measure or improve AI fairness and explainability. These measurement methods have been shown to be unsuitable including when used in an "off-label" or out-of-context manner if applied to measure many types AI systems.

*A Problematic AI Fairness Measurement Method*

In particular, the research found problems with AI governance tools that incorporate the US Four-Fifths or 80% Rule in an attempt to measure disparate impacts and fairness of AI systems.

The Four-Fifths Rule is well-known in the US labor recruitment field as a measure of adverse impact and fairness in hiring selection practices. Detailed in the Equal Employment Opportunity Commission *Uniform Guidelines on Employee Selection Procedures of 1978*,[2] the rule is based on the concept that a selection rate for any race, sex or ethnic group that is less than four-fifths —or 80%—of the rate reflecting the group with the highest selection rate is evidence of adverse impact on the groups with lower selection rates. The rule has been widely applied by employers,[3] lawyers,[4] and social scientists[5] to determine if hiring practices are lawful and if they result in disparate or adverse impacts against certain groups of people.

While the Uniform Guidelines state that the Four-Fifths rule is "generally" regarded by federal enforcement agencies as evidence of adverse impact, it explains that in some cases, smaller differences in selection rate may constitute adverse impact, and in others, greater differences in selection rate may not constitute adverse impact. In other words, context matters.[6]

Despite its widespread use, legal, employment, and technical experts have cautioned against use of the Four-Fifths Rule as a singular means of assessing disparate impact.[7] Many experts warn against simplistic applications of the rule, both within its historical use in US labor contexts as well as for its use in AI contexts.[8]

In June 2023, the chair of the U.S. Equal Employment Opportunity Commission cautioned against relying solely on meeting the 80% threshold. Calling the Four-Fifths rule "a check" and just one single standard used at the start of federal investigations, rather than the only measure

---

[2] Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607(1978)

[3] *4/5ths Rule - Meaning & Definition*, MBA Skool, (Aug. 16, 2023, 11:01 AM), https://www.mbaskool.com/business-concepts/human-resources-hr-terms/13006-45ths-rule.html.

[4] 1607.4 Information on impact, Legal Information Institute at Cornell Law School, 29 CFR § 1607.4 (Aug. 16, 2023, 11:07 AM), https://www.law.cornell.edu/cfr/text/29/1607.4.

[5] Alexander P. Burgoyne et al., *Reducing adverse impact in high-stakes testing*, 87 Intelligence, art. 101561 (July-Aug. 2021), https://doi.org/10.1016/j.intell.2021.101561.

[6] *Id.*

[7] *E.g.*, M.S.A. Lee & L. Floridi, *Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs,* 31 Minds & Machines 165,191 (June 9, 2020), https://doi.org/10.1007/s11023-020-09529-4 ("Feldman et al. (2015) have formalized the approach to identifying disparate impact, but their methodology for pre-processing the data to remove the bias has shown instability in performance of the technique").

[8] Philip Roth et al., *Modeling the Behavior of the 4/5ths Rule for Determining Adverse Impact: Reasons for Caution,* 91 J. Applied Psych. 507, 522 (May 2006).

used for gauging disparate impact, she said that "smaller differences in selection rates may constitute disparate impact."[9]

Further, according to a U.S. Justice Department legal manual addressing disparate impact, "not every type of disparity lends itself to the use of the Four-Fifths rule, even with respect to employment decisions."[10] Legal scholars also have questioned the limits of the Four-Fifths rule, noting its failure to statistically reflect hiring disparity impact adequately.[11]

Despite those caveats, the Four-Fifths Rule and its 80% benchmark have been repurposed in computer code form and used in a variety of AI fairness metrics and tools.[12] The rule is applied in both employment[13] and non-employment contexts [14] as a means of measuring or "removing" bias or disparate impacts.[15] It is also used outside of the US employment context and is encoded into AI governance tools offered in other jurisdictions.[16]

Again, we suggest that use of AI disparate impact measurement methods based on the Four-Fifths Rule and its 80% benchmark should be assessed thoroughly for appropriate application, and possibly removed from consideration as evaluative AI measures in some or all contexts.

---

[9] Chair Burrows spoke during a keynote speech in June 2023 at the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) attended by World Privacy Forum representatives. She said that it is "worrisome" when employers or vendors suggest that meeting the 80% benchmark is enough to ensure that a hiring approach or system does not create disparate impact.

[10] U.S. Dep't of Just., Just. Manual § 7 (1964).

[11] Jennifer Peresie, *Toward a Coherent Test for Disparate Impact Discrimination.* 84 Ind. L. J. 773, 802 (2009), http://ilj.law.indiana.edu/articles/84/84_3_Peresie.pdf.

[12] Elizabeth Watkins et al., *The Four-Fifths Rule is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness,* Parity Techs. Inc., (March 3, 2022), https://ssrn.com/abstract=4037022.

[13] Hilke Schellmann, *Auditors are testing hiring algorithms for bias, but there's no easy fix,* MIT Technology Review (Feb. 11, 2021), https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/.

[14] *Bias Mitigation with Disparate Impact Remover*, Jupyter nbviewer (Aug. 16, 2023, 11:18AM), https://nbviewer.org/github/srnghn/bias-mitigation-examples/blob/master/Bias Mitigation with Disparate Impact Remover.ipynb.

15 AIF360, GitHub, Trusted AI, Supported Bias Mitigation Algorithms, "Disparate Impact Remover." (November 11, 2023), https://github.com/Trusted-AI/AIF360/tree/master. (Documentation for the Disparate Impact Remover algorithm supported by AI Fairness 360 specifically cites 2015 research introducing a disparate impact measurement based on the Four-Fifths Rule's 80% benchmark.)

[16] Multiple AI governance tools surveyed for World Privacy Forum's *Risky Analysis* report mention or recommend fairness assessments that use encoded versions of the Four-Fifths or 80% Rule to measure disparate impact. See Part I and Appendix C of the report for more detail.

*Problematic AI Explainability Measures*

In the absence of widely-adopted AI explainability standards, two approaches—SHAP and LIME—have grown in popularity, despite attracting an abundance of criticism from scholars who have found them to be unreliable methods of explaining many types of complex AI systems.[17]

Use of both SHAP[18] and LIME[19] has increased in part because they are model agnostic, meaning they can be applied to any type of model that data scientists build. An abundance of accessible and easy-to-use documentation related to the two methods has also fostered interest in them.[20]

However, the applicability and efficacy of both SHAP and LIME are limited, particularly when used in an attempt to explain complex AI systems comprised of non-linear machine or deep learning models. In a typical use case, an AI practitioner might employ SHAP or LIME to explain a single instance of a model input, such as one decision or prediction, rather than the whole model. Because both methods work by approximating more complex, non-linear models (the types that are often called "black-box" models) with more straightforward linear models, they may produce misleading results.[21]

Short for Shapley Additive exPlanations, SHAP is based on a concept known as the the Shapley Value, introduced by Lloyd Shapley in 1951[22] in the context of cooperative game theory. The Shapley Value is a method used to determine the importance or contribution of each player to an overall competition between groups. [23]

---

[17] Dylan Slack et al., *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods, AIES '20 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Ass'n for Computing Machinery (Feb. 7, 2020), https://doi.org/10.1145/3375627.3375830.

[18] *shap*, GitHub, https://github.com/shap.

[19] *lime*, GitHub, marcotcr, https://github.com/marcotcr/lime.

[20] This is based on a description of how SHAP and LIME work and their problems, as intended for a layperson, provided by Tim Miller, professor in artificial intelligence at the School of Electrical Engineering and Computer Science at The University of Queensland, during interviews conducted by World Privacy Forum (WPF) in June and November 2023. Miller was professor in the School of Computing and Information Systems at The University of Melbourne, and co-director of its Centre of AI and Digital Ethics, when WPF spoke with him in June 2023. Miller said that in general, LIME is unstable and inappropriate as an explainability metric for machine learning, while SHAP-based methods are also limited in effectiveness. Professor Tim Miller, Univ. Of Queensland Australia, https://eecs.uq.edu.au/profile/9477/tim-miller.

[21] November 2023 WPF interview with Tim Miller.

[22] Lloyd S. Shapley*, Notes on the N-Person Game — II: The Value of an N-Person Game*, RAND Corp. (1951), https://www.rand.org/pubs/research_memoranda/RM0670.html.

[23] S. Hart*, Shapley Value*, in The New Palgrave Dictionary of Economics 1-6 (1987), https://doi.org/10.1057/978-1-349-95121-5_1369-1.

Today, SHAP is used for another purpose entirely: in an attempt to expose and quantify feature importance, or the importance of factors that contribute to predictions of machine learning models.[24] Oftentimes, SHAP is used in the hopes of revealing how factors affect the outputs of opaque, "black box" AI systems such as deep learning models and neural networks, which are difficult to interpret.

SHAP reflects feature importance numerically. For instance, when using SHAP to determine how certain input features affect a more straightforward linear regression model trained on a California housing dataset, the SHAP value of the median house age in a block group might be expressed as -0.22, and the SHAP value of median income as +0.92. The process would be used to add other features, such as the average number of rooms or average home occupancy, until the current model output is reached.[25]

Although Shapley values have been applied in the context of feature importance for decades,[26] researchers have found several mathematical, practical, contextual, and epistemological problems associated with use of the method for explaining AI systems. For example, when attempting to attribute influence to a large set of features affecting AI model decisions or predictions, the approach relies on the modeler to decide which features count as "players" and which are redundant; these subjective decisions can affect the resulting explanations.[27]

Scholarly research also indicates that some users of SHAP may not understand how to interpret its results. A survey of data scientists using SHAP-based tools showed that many were unable to accurately describe what SHAP values or scores represented.[28] The study also found that the popularity of SHAP-based tools influenced some data scientists to trust the tools even if they did not understand what they did or how to interpret their results.

In addition, research shows that use of SHAP in AI explainability tools may lead users to falsely believe they discovered a precise explanation for why or how a system produced a specific

---

[24] This description is based on an overview of how SHAPley Values work intended for a layperson as provided by Elizabeth Kumar, a Computer Science PhD candidate at Brown University, during interviews conducted by WPF in April and November 2023.
Lizzie Kumar personal website, https://iekumar.com/.

[25] Vinícius Trevisan. Towards Data Science, Medium. Jan 17, 2022. https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

[26] W. Kruskal, *Relative importance by averaging over orderings*, The American Statistician, 41(1):6–10, 1987.

[27] I. Elizabeth Kumar et al., *Problems with Shapley-value-based explanations as feature importance measures.*

[28] Harmanpreet Kaur et al., *Interpreting interpretability: Understanding data scientists use of interpretability tools for machine learning, CHI '20 Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* Ass'n for Computing Machinery, 114 (Apr. 23, 2020), https://doi. org/10.1145/3313831.3376219.

output, such as a decision or prediction. This in turn may lead to misconceptions about what SHAP values represent and the actionable information that can be gleaned from them.[29]

Even scholars who acknowledge benefits of using SHAP to provide insight into certain aspects of models and data suggest they "can lead to wrong conclusions if applied incorrectly," [30] and argue that they can be expensive to compute.[31]

LIME, a similar AI explainability method that has grown in adoption, was first introduced in 2016.[32] Short for Local Interpretable Model-agnostic Explanations, LIME produces explanations by randomly sampling "locally" around the singular instance chosen to be explained. But its randomness is a pitfall: If LIME is used again in an attempt to explain the very same instance, its explanation will be different.[33] The use of LIME for AI explainability has been criticized, and research shows the method can lead to inaccurate results,[34] or be manipulated or "gamed."[35]

Overall, the research indicating that there are vulnerabilities in these popular explainability measures is not reassuring; however, it is not completely unexpected. Trustworthy AI implementation is still nascent, with much work and refinement yet to come.

We suggest that use of measurement methods incorporating SHAP or LIME should be assessed thoroughly for appropriate application, and possibly removed from consideration as evaluative AI measures.

**Question 6. Do the minimum practices identified for safety-impacting and rights-impacting AI set an appropriate baseline that is applicable across all agencies and all such uses of AI? How can the minimum practices be improved, recognizing that agencies will need to apply context-specific risk mitigations in addition to what is listed?**

The OMB draft memo states in section iv., "Minimum Practices for Either Safety-Impacting or Rights-Impacting AI," that agencies must follow specific practices related to AI impact

---

[29] Elizabeth Kumar et al., *Shapley Residuals: Quantifying the limits of the Shapley value for explanations*, Neural Info. Processing Sys. (2021).

[30] Christoph Molnar et al., *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*, Arxiv (2022), https://arxiv.org/pdf/2007.04131.pdf.

[31] Christoph Molnar, *SHAP Is Not All You Need*, Mindful Modeler (Feb. 7, 2023), https://mindfulmodeler.substack.com/p/shap-is-not-all-you-need.

[32] Marco Tulio Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Ass'n for Computing Machinery, 1135–1144 (Aug. 13, 2016), https://doi.org/10.1145/2939672.2939778.

[33] According to a November 2023 interview with Tim Miller.

[34] Romaric Gaudel et al., *s-LIME: Reconciling Locality and Fidelity in Linear Explanations*, Arxiv, (Aug. 2, 2022), https://arxiv.org/abs/2208.01510.

[35] Dylan Slack et al.

assessments; risk mitigation measures; testing, evaluation and monitoring to ensure that AI works in real-world contexts; AI functionality in high-risk decision-making, and more.

It is imperative the tools and techniques used to quantifiably measure AI systems according to safety and rights impacts, and to gauge success according to particular metrics, are effective and reliable in accordance with policy goals. The AI governance tools established for such purposes will help form the foundation of risk scores, consumer scores, ratings or other statistics that agencies rely upon to conduct much-needed impact assessments, risk evaluations and ongoing monitoring. Without meaningful oversight and quality assessments of the tools and measurement methods themselves, new problems or harms could be introduced.

In the World Privacy Forum's forthcoming *Risky Analysis* report, we suggest pathways for creating a healthy AI governance tools environment, and offer suggestions for governments, multilateral organizations, and others creating or publishing AI governance tools.

We found that some AI governance tool providers do not conduct quality assessments, or if they do, they do not always conduct them according to an internationally recognized standard.

The suggestions made in the report include best practices taken from existing AI and other quality assessment standards and practices already in widespread use. Appropriate procedural and administrative controls include: 1) providing AI governance tool documentation and contextualization, review, audit, and other quality assurance procedures to prevent integration of inappropriate or ineffective methods in policy guidance; 2) identifying and preventing conflicts of interest; and 3) ensuring that capabilities and functionality of AI governance tools align with policy goals. If governments, multilateral institutions, and others working with or creating AI governance tools can incorporate lessons learned from other mature fields such as data governance and quality assessment, the result will establish a healthier body of AI governance tools, and over time, healthier and more trustworthy AI ecosystems.

**Question 8. What kind of information should be made public about agencies' use of AI in their annual use case inventory?**

In particular, section iv. of the OMB Draft Memo includes requirements that agencies provide public notice and plain-language documentation in the AI use case inventory. It notes that the documentation should be accessible in contexts where people will interact with or be impacted by the AI, and that even where agencies' use cases are excluded from the public inventory requirements described in the guidance, they may still be required to report relevant information to OMB and must ensure adequate transparency in their use of AI, as appropriate and consistent with applicable law.

All of these minimum practices for documentation are crucial components of AI accountability and transparency. In addition, we urge OMB to include documentation related to AI governance tools used to measure and improve AI systems used by federal agencies. Just as with any tool or product released to the public, agencies should create and make robust documentation available.

As included in World Privacy Forum's *Risky Analysis* report, documentation that can assist in creating more transparency can include information about the developer, date of release, results of any validation or quality assurance testing, and instructions on the contexts in which the

methods should or should not be used. A privacy and data policy is also important and should be included in the documentation of AI governance tools. End users should be made aware of the evaluations in a prominent manner, and the evaluation should be readily understandable by non-expert users.

Ensuring that there is no conflict of interest related to AI governance tools is another quality measure that is readily achievable. For example, core pieces of information to make available to end users should provide details about how development of AI governance tools are resourced and financed, by whom, and who published them. Commercial Interests also should be noted, for instance, if the tool promotes specific commercial products or services. Affiliations, including relationships that potentially impact objectivity, including information about commercial or other entities that donated the tool for open-source use, should also be noted.

Additional items can be provided in the documentation, for example:

- Documentation should provide the suggested context for the use of an AI governance tool. AI systems are about context, which is important when it comes to applicable uses, environment, and user interactions. A concern is that tools originally designed for application in one use case or context may potentially be used in an inappropriate context or use case or "off-label" manner due to lack of guidance for the end user.

- Documentation should give end users an idea of how simple or complex it would be to utilize a given AI governance tool.

- Cost analysis for utilizing the method: How much would it cost to use the tool and validate the results?

- A data policy: A detailed data policy should be posted in conjunction with each AI governance tool. For example, if applicable, this information could include the kind of data used to create the tool, if data is collected or used in the operation of the tool, and if that information is used for further AI model training, analysis, or other purposes.

- Complaint and feedback mechanism: AI governance tools should provide a mechanism to collect feedback from users.

- Cycle of continuous improvement: Developers of AI governance tools should maintain and update the tools at a reasonable pace.

A great deal of existing work has already been done in other areas that could be helpful. For example, significant documentation standards and norms exist around consumer products, software products, and other technology products offered to the public. These norms are

encapsulated in multiple ISO standards,[36] as well as OECD Responsible Business Conduct principles and implementation guidance.[37] [38]

**Conclusion**

In concluding, we add one additional note regarding practical implementations of the work that will be needed in order to contextualize Agency AI activities effectively, safely, and fairly. The opportunities that OMB Circular No. A-119 Revised [39] affords to government agencies to work with stakeholders to create fair, multistakeholder Voluntary Consensus Standards (VCS) with good ground-level rules developed through collaborative work is non-trivial. Given the need for practical multistakeholder work in the area of AI, WPF is particularly interested to see how a VCS process could effectuate positive improvements, particularly regarding the day-to-day governance of AI systems.

WPF stands ready to assist. Thank you again for the opportunity to comment, and we hope to have the opportunity to discuss these issues with you further.

Respectfully submitted,

Pam Dixon,
Executive Director,
World Privacy Forum

---

[36] See for example G. F. Hayhoe, "ISO standards for software user documentation," *2012 IEEE International Professional Communication Conference*, Orlando, FL, USA, 2012, pp. 1-3, doi: 10.1109/IPCC.2012.6408631. See also the work of NIST regarding recommended criteria for cybersecurity labeling of consumer software. While not directly related regarding topic, the procedures and ideas for labeling could be helpful, particularly if tested in the AI governance tools context. See: Recommended criteria for cybersecurity labeling of consumer software, National Institute of Standards and Technology, Feb. 4 2022. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.02042022-1.pdf.

[37] *OECD Due diligence guidance for Responsible Business Conduc*t, OECD. 31v May 2018. Available at: https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm.

[38] Allan Jorgensen, Karine Perset, Rashad Abelson, *Recoding our understanding of RBC in science, tech, and innovation*, OECD. Oct 02 2023. Available at: https://www.oecd-forum.org/posts/recoding-our-understanding-of-rbc-in-science-tech-and-innovation-what-s-new-in-the-oecd-mne-guidelines.

[39] Office of Management and Budget, Circular No. A-119 Revised. The White House, President Barack Obama. Available at: https://obamawhitehouse.archives.gov/omb/circulars_a119.